UNIVERSITE SORBONNE NOUVELLE – PARIS 3

ED 268 (LANGAGE ET LANGUES)

LABORATOIRE DE PHONÉTIQUE ET PHONOLOGIE

Thèse de doctorat en phonétique

Interdependence between Tones, Segments and Phonation types in Shanghai Chinese

acoustics, articulation, perception and evolution

Jiayin Gao

Thèse dirigée par
Pierre Hallé
Soutenue le 21 mai 2015

Jury :

Mme Yiya Chen, Assistant Professor, Leiden University (pré-rapporteur)
M. Pierre Hallé, DR CNRS (directeur)
M. Guillaume Jacques, CR HDR CNRS (pré-rapporteur)
Mme Martine Mazaudon, DR CNRS
M. Alexis Michaud, CR CNRS
M. Laurent Sagart, DR CNRS

Résumé

Cette étude porte sur les corrélats phonétiques des registres tonals *yin* vs. *yang* du shanghaïen parlé dans la région urbaine de Shanghai. Nos investigations acoustique, articulatoire et perceptive ont montré qu'en dehors du F0, des indices multi-dimensionnels comme le voisement (voisé pour *yang* et non-voisé pour *yin*), le pattern de durée (ratio C/V bas pour *yang* et élevé pour *yin*), et le type de phonation (soufflé pour *yang* et modal pour *yin*) participent tous à la définition du registre tonal. Parmi tous ces indices, nous tâchons de distinguer les traits redondants liés aux effets coarticulatoires des survivances de changements diachroniques. En particulier, la voix soufflée qui accompagne les tons *yang* est un trait redondant, issu d'une évolution tonale qui est la transphonologisation de distinction de voisement vers la distinction de registre tonal, ou « bipartition tonale ». Nous proposons que la perte d'un trait redondant issu d'un changement diachronique peut être très lente si ce trait ne contrarie pas les effets coarticulatoires et/ou si le trait a une fonction perceptive.

En nous basant sur les données synchroniques des locuteurs de deux générations (20-30 ans vs. 60-80 ans), nous constatons une tendance vers la disparition de cette phonation soufflée. Nous constatons également une évolution plus avancée chez les femmes que les hommes de leur âge. Dans notre étude, nous essayons d'expliquer ce changement tant par des causes internes que par des causes externes.

Mots clés : Shanghaïen, ton, type de phonation, voix soufflée, voisement, durée, trait redondant, variations, transphonologisation, phonologie panchronique, évolution des sons

Abstract

This study bears on the phonetic correlates of the *yin* vs. *yang* tone registers of Shanghai Chinese as spoken in Shanghai urban area. Our acoustic, articulatory, and perceptual investigations showed that beside F0, multidimensional cues, such as voicing (voiced for *yang* vs. voiceless for *yin*), duration pattern (low C/V ratio for *yang* vs. high C/V ratio for *yin*), and phonation type (breathy for *yang* vs. modal for *yin*) enter in the specification of tone register. Among all these cues, we attempt to distinguish the redundant features related to coarticulatory effects from those that are remnants of diachronic changes. In particular, the breathy voice accompanying *yang* tones, which is a redundant feature, arose from a tonal evolution, namely the transphonologization of a voicing contrast into a tone register contrast, that is, the "tone split." We propose that the loss of a redundant feature arisen from a diachronic change may be very slow if that feature does not conflict with coarticulatory effects and/or if that feature has a perceptual function.

Based on the synchronic data from the speakers of two generations (20-30 years vs. 60-80 years), we find a trend toward the loss of this breathy phonation. We also find that this evolution is more advanced in women than men of the same age. In our study, we try to explain this change by internal factors as well as by external factors.

Keywords : Shanghai Chinese, tone, phonation type, breathy voice, voicing, duration, redundant feature, variations, transphonologization, Panchronic phonology, sound change

In memory of Ren Nianqi 任念麒。

献给我的父亲母亲。

Acknowledgements – Remerciements

Mes remerciements les plus chaleureux vont tout naturellement à mon directeur de thèse, Pierre Hallé, tant pour ses orientations scientifiques que pour ses aides techniques dans les traitements des données, depuis ma deuxième année de Master et tout le long de mes quatre ans de thèse. Je n'aurais pas abouti à cette thèse sans ses corrections minutieuses et constructives. Il sait toujours reformuler mes idées qui parfois partent dans tous les sens.

I am greatly honored to have Yiya Chen, Guillaume Jacques, Martine Mazaudon, Alexis Michaud and Laurent Sagart in the thesis committee. I have been strongly inspired by the seminars and the published work from each of them. I am grateful for their invaluable suggestions on the earlier version of this dissertation. I am especially indebted to Alexis Michaud for his suggestions and criticisms on my work, for his Matlab program for the treatment of the electroglottographic data, for his encouraging song "Tu vas finir ta these !" during the finishing period of my dissertation.

Je me rappelle toujours de mon tout premier cours de phonétique et phonologie que j'ai suivi en tant qu'auditrice libre à l'Université Grenoble 3, avec Anne Vilain. J'ai encore ce souvenir : Anne a posé une question en cours : « comment distinguer la consonne /p/ du /b/ ? », novice en étude phonétique et biaisée par mon intuition sur le shanghaïen, j'ai tenté : « j'ai l'impression que /p/ est plus aïgue que /b/ », auquel Anne a répondu : « c'est vrai, mais ce n'est pas la chose la plus importante. » Un grand merci à Anne pour la découverte de cette discipline !

Je remercie vivement les (anciens) collègues du LPP. Merci notamment à Nicolas Audibert et Thibaut Fux pour leurs nombreux scripts et conseils techniques. Merci également à Laurianne Georgeton et Angélique Amelot pour leur aide dans l'expérience avec le Qualisys Motion Capture System. A ce sujet, je pense toujours à ma première expérience en tant que cobaye et locutrice du shanghaïen avec l'ePGG (external-photo-glotto-graphy) qui m'a initiée à la vraie phonétique expérimentale. Merci à Martine Toda, Shinji Maeda, Kyoshi Honda et Jacqueline Vaissière qui ont rendu cette expérience possible. Merci à Annie, Rachid, Cécile, Didier, Naomi, Cédric, Martine, Claire, Rajesh, Patrick, etc. etc. ainsi que tous les phonéticiens, phonologues et ingénieurs du LPP pour m'avoir donné le goût pour la phonétique et la phonologie.

Merci aux membres du groupe didactique du LPP (Simon, Takeki, Nikola, Laurianne, Altijana, etc.) qui me sortent de temps en temps de mon petit monde du shanghaïen (qui est certes déjà vaste). Merci à Fanny pour nos discussions sur les traitements statistiques. Et merci aux autres petites du LPP (Jane, Nora, Yaru, Charlotte …) pour les bons moments ensemble !

Thanks to different colleagues I met on various occasions. I express my gratitude to John Esling, Scott Moisik for their excellent seminars in Paris and for our discussions on Wu phonation types and on my physiological and acoustic data. By the way, the fiberoptic data, although not reported in this work, were collected on the initiative of Céline Chang and Feng-fan Hsieh from National Tsing Hua University of Taiwan. I am grateful to them for their hospitality. (Thanks to the doctor by the way for his excellent job. "He should have lowered more deeply the tube so that the epiglottis wouldn't be there!" commented Pr. Esling.) I am also deeply grateful to Marc Brunelle and Andrea Levitt for their precious advice on my work. Finally, many thanks to Eric Zee, Carlos Gussenhoven, Phil Rose, Cathi Best and CHEN Zhongmin for helpful discussions, and to Rachel Shen for her advice on statistical analyses.

Merci aux collègues avec lesquels j'ai fait autant de découvertes gastronomiques que scientifiques durant mes années de thèse. Je pense aux "taotonologues", Takeki, Inyoung, Pierre et Ioana, pour le goût que l'on partage pour l'exotisme. Merci également aux copains de la BULAC (Rajesh, Désiré, Aziz, GONG Xun) avec lequels j'ai passé mes derniers moments de rédaction de cette thèse.

Merci à YIN Min pour être locutrice de mes stimuli. Thanks also to my friends in Shanghai and Paris who participated in one or several experiments without complaining. Thanks to Professor DING Hongwei and her students from Tongji University, and to Mme HAO Jia from Shanghai International Studies University for helping me recruiting participants in Shanghai.

A big thanks to my parents for their unconditional support, especially to Mom for helping me recruiting elderly speakers. Thanks to my aunt as well for her kind help.

Last but not least, I thank Professor Arthur Abramson for his quick reply on my request of the dissertation of his late student, REN Nianqi, who conducted the first comprehensive experimental study on phonetics of Shanghai Chinese. This work is in memory of REN Nianqi, who passed away at such a young age.

Table of contents

0 INTRODUCTION

This dissertation attempts at describing in some detail the articulatory and acoustic correlates of Shanghai Chinese tones. One specificity of this dialect is that it retained the Middle Chinese phonological voicing contrast among syllable onset obstruent consonants, like other Wu dialects and contrary to many other Chinese dialects. This specificity is probably related to the prosodic system and the tone sandhi pattern common to northern Wu dialects. The voicing contrast surfaces either as a tone register contrast or as a phonetic voicing contrast, depending on within-word syllable position. We try to document this interesting interplay, and otherwise attempt to identify the most robust and important correlates of Shanghai Chinese tones not only in production but also in perception, at a synchronic level.

We are also interested in the diachronic aspects revealed by the synchronic evidence. A transphonologization process from a voicing contrast to a tone register contrast has been widely observed in East and Southeast Asian languages, with, probably an intermediate stage of the phonation type contrast. Why, then, do the "ancestral" voicing feature and the most recent tone feature coexist in Shanghai Chinese? What is the status of the phonation type? How do all these features evolve in a local "dialect" that is in permanent contact with migrant languages and with the official language, Standard Chinese?

Wu dialects were among the first Sinitic varieties that were investigated phonetically by LIU Fu and CHAO Yuen-Ren (Liu, 1925; Chao, 1928). This strand of research was later taken up by Sherard (1972) and Cao and Maddieson (1992).

Yet, to my knowledge, the only comprehensive experimental study on stops, tones, and phonation type in Shanghai Chinese, including physiological, acoustic and perceptual investigations has been conducted by 任念麒 REN Nianqi in 1992 for his Ph.D. He wrote, in all modesty, that his study "may serve to 抛砖引玉 *pao zhuan yin yu* 'cast a brick to attract jade' (Ren, 1992: 15)". This excellent phonetician and talented singer deceased shortly after completing his Ph.D, which was a great loss.

Twenty-three years later, I submit the present experimental work, probably not as precious as jade, in the memory of Dr. Ren, hoping to add another brick to the temple of linguistic sciences.

This dissertation is organized as follows. Chapter 1 provides a general overview of Chinese dialects and the Wu dialect family before focusing on Shanghai Chinese (or Shanghai Wu), the Shanghai variety of Wu. Geographical and generational variations in Shanghai Chinese are described. Chapter 2 presents the phonetic and phonological system of today's Chinese of Shanghai City (henceforth Shanghai Chinese). It presents the synchronic interdependence between tone, voicing and other glottal settings, followed by a diachronic account of this interdependence. Chapter 3 gives a detailed account of the phonetic correlates of Shanghai tone and voicing contrasts described in the literature. Chapter 4 investigates Shanghai speakers' production of tones and their phonetic correlates, especially the relation between tone, voicing and phonation, based on acoustic and physiological measurements. Chapter 5 examines the perceptual aspect of the most important phonetic correlates found in the previous chapter. Chapter 6 provides a general discussion of the main results of the experiments.

1 SHANGHAI CHINESE AND THE WU DIALECT GROUP

Summary

This chapter begins with a general presentation of Chinese dialect groups, including a discussion on the terminological difference between "language" and "dialect." Political and cultural factors are the main contributors to the adoption of the term "dialect" concerning Chinese dialects (§1.1).

We then take a closer look at the Wu dialect group and its subgroups. According to the traditional view, the phonological feature common to all Wu dialects is its retention of the three-way laryngeal contrast that existed in Middle Chinese (§1.2).

We finally zoom in on Shanghai Chinese, the Shanghai variety of Wu, which can still be divided into several subgroups. A quick description of geographic variation and generational variation is given. This dissertation will focus on the Shanghai City Chinese spoken by both elderly and young speakers (§1.3).

1.1 Demolinguistics of languages in China

Sinitic and non-Sinitic languages are both spoken in the land of China, including languages from Sino-Tibetan, Altaic, Tai-Kadai, Hmong-mien (Miao Yao), Austro-Asiatic, Indo-European, Korean and Austronesian families.

The term "Chinese dialects" is commonly used to designate the "Sinitic languages" spoken in China, although Chinese dialect groups are more like language families. For one thing, they diverged from Old Chinese 2000-2200 years ago. For comparison, French, which derived from Vulgar Latin between the 6th and 9th century, is considered as a separate language from other Romance languages. On the other hand, what we can call "dialectes d'oïl," spoken in the northern part of France including Parisian French, have a much more recent ancestor. For another thing, two speakers from two different Chinese dialect groups without knowledge of the other dialect group are in most cases mutually unintelligible. Unintelligibility is even quite common among speakers of different dialects from the same dialect group.

Mutual intelligibility is often proposed as a probably oversimplified criterion to distinguish "language" from "dialect". "If two varieties of speech are mutually intelligible, they are strictly DIALECTS of the same LANGUAGE; if they are mutually unintelligible, they are different languages." (Crystal, 2008: 319, his emphasis). If we adopted this criterion, there would be hundreds of languages in China, not to mention the fact that tests of mutual intelligibility tend to yield results that are by no means binary, but "of degrees of more or less" (Chambers & Trudgill, 1990: 4). Furthermore, as Crystal (2008: 319) added, political and cultural factors play also important roles in the choice between these two terms. *Putong hua* 普通话 'Common Speech,' or Standard Chinese, the standardized form of Beijing Mandarin Chinese, obtained its official status in the middle of the 20th century and serves as a *lingua franca* for both Han and non-Han people. As a consequence, the other varieties are accepted as "dialects" by their speakers as well as by most linguists.

The word for "dialect" in Chinese is 'fang yan' 方言, literally meaning "local speech, regional speech." In spite of a rather homogeneous development of the literary language, there is strong evidence suggesting that shows vernacular local dialects

already existed in ancient time. The first collection of Chinese dialect words is said to be *Fang Yan* 方言, the full title of which is *Youxuan shizhe juedai yu shi bieguo fangyan* 輶軒使者絕代語釋別國方言 [Local speeches of other countries in times immemorial explained by the Light-Carriage Messenger], edited by Yang Xiong 扬雄 (53 BC – 18 CE). According to Norman (2003), the dialectalization of spoken Chinese began during the great Qin and Han expansion (221 BCE – 220 CE). But Baxter and Sagart (2014: 319) find evidence for dialect diversity even earlier, during the Old Chinese period (i.e., before 221 BCE).

In today's everyday language, names given to Chinese dialects are generally geographical designations, such as Shanghai dialect, Guangdong dialect, etc. Dialects are classified in dialect groups, taking into account typological, historical and geographical factors. The number of Chinese dialect groups varies according to different criteria of classification. The traditional classification is from Yuan Jiahua 袁家骅 (Yuan, 1961) who posits seven dialect groups: 吴 (Wu), 赣 (Gan), 湘 (Xiang), 闽 (Min), 客家 (Kejia or Hakka), 粤 (Yue or Cantonese), and 北方话 (Beifanghua or Mandarin), as shown in Figure 1. This classification is based on the first scientific dialectal classification by Li Fang-Kuei 李方桂 (1937), that further divided the Mandarin group into three groups (Northern Mandarin, Eastern Mandarin and Southwestern Mandarin), but classified 赣 (Gan) and 客家 (Kejia or Hakka) into one group.

Figure 1. Map of Sinitic languages (Chinese dialects) in China.
(http://www.axl.cefan.ulaval.ca/asie/chine-2langues.htm)

In *Language Atlas of China* (Wurm, Li, Baumann & Lee, 1987), it is proposed that the Jin 晋 group, the Hui 徽 group, traditionally treated as subgroups of Mandarin, and the Ping 平 group, traditionally treated as a subgroup of Yue 粵, should be separated as independent dialect groups at the same level as Mandarin and Yue. Although this proposal was supported by the Chinese Academy of Social Science (CASS), the separation of these groups is controversial (cf. Qian, 2010).

A broader classification by Norman (1988) divided Chinese dialects into three large groups based on ten phonological, lexical or syntactical criteria: the *Northern group*, that corresponds to the Mandarin area, the *Southern group*, including the Kejia (Hakka), Yue (Cantonese), and Min, and the *Central group*, which is a transitional area that possesses both northern and southern features. The dialects of the Central group, which originally shared more common features with the dialects of the Southern group, are influenced by Northern features principally brought by migration.

In this work, we employ the term "dialect groups" rather than "language families." Our study bears on Shanghai Chinese (or Shanghai Wu), which belongs to the *Central group* according to Norman's classification.

1.2 The Wu dialect group

The Wu dialect group is the second largest dialect group with 87.1 million speakers (2003), just after the Mandarin group, which represents 68% of the whole population, that is, 867.2 million speakers (2003).[1] Wu dialects are spoken in southern Jiangsu 江苏, in the major part of the Zhejiang 浙江 province, the municipality of Shanghai 上海, as well as in a few counties (Shangrao 上饶, Yushan 玉山 and 广丰 Guangfeng) of northeastern Jiangxi 江西, in Xuancheng 宣城 city (southeast of the Anhui 安徽 province), and in the Pucheng 浦城 county of the Fujian 福建 province.

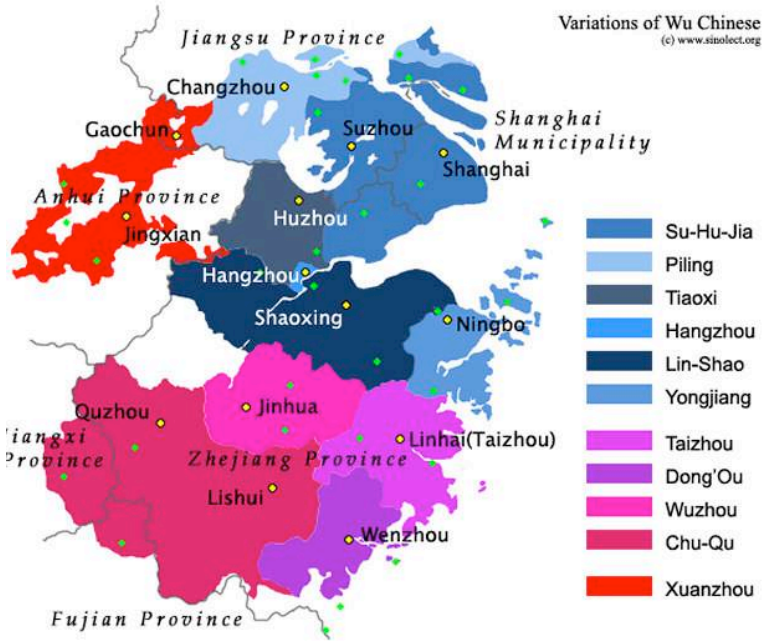


Figure 2. Map of the main Wu dialect subgroups (http://www.sinolect.org).

As proposed by Fu, Cai, Bao, Fang, Fu, and Zhengzhang (1986), the Wu dialect group is further divided into six main subgroups: Taihu 太湖, Taizhou 台州, Oujiang

[1] The data come from *L'aménagement linguistique dans le monde* by Jacques Leclerc: http://www.axl.cefan.ulaval.ca/asie/chine-2langues.htm.

(Dong'Ou) 瓯江, Wuzhou 婺州, Chuqu 处衢, and Xuanzhou 宣州. Figure 2 shows the distribution of the main Wu subgroups (the Taihu subgroup in different hues of blue). Shanghai Chinese belongs to the Taihu subgroup. This proposal is adopted in *Language Atlas of China* (Wurm et al., 1987) and currently widely accepted by Chinese linguists.

According to the traditional view, the major phonological feature that distinguishes the Wu group from the other groups is its uniform retention of the three-way laryngeal contrast that existed in Middle Chinese: Wu dialects possess three stop series, voiceless unaspirated, voiceless aspirated and voiced, labeled *quanqing* 全清 'full clear,' *ciqing* 次清 'secondary clear' and *quanzhuo* 全浊 'full muddy' in traditional terms of Chinese phonology, and two fricative series, voiceless and voiced. The voiced series was lost during the evolution of the other dialect groups, although a few Xiang 湘 and Mandarin dialects still retained the voiced series. For a detailed description of the tonal system of Shanghai Chinese, see §2.1.3.1.

1.3 Variation in Shanghai Chinese

In a broad sense, Shanghai Chinese comprises different varieties spoken in the Shanghai area, including urban and suburban areas, with approximately 14 million speakers. The geographic variation is so important that several dialect subgroups can be classified according to their proper phonological system.

In a narrow sense, Shanghai Chinese designates one of these subgroups, the City subgroup, that undergoes the most rapid evolution among all the varieties, so that important variation can be observed among generations.

1.3.1 Geographic variation

Shanghai is one of the world's largest metropolitan cities with its area of 6340.5 km^2 (suburban areas included). Undoubtfully, there are considerable variations among the subdialects of Shanghai. Chen (2003) subdivided Shanghai Chinese into five dialect subgroups, based solely on their tonal systems: City subgroup, spoken in

the Shanghai urban area, Chongming 崇明 subgroup, Jiading 嘉定 subgroup, Songjiang 松江 subgroup, and Liantang 练塘 subgroup, as shown in the map in Figure 3. Chen (2003) called them dialect groups: some of them can be further divided in several subgroups. The formation of the Chongming, Jiading, Songjiang and Liantang subgroups was principally due to administrative divisions. As for the City subgroup, it was very similar to the adjacent Songjiang subgroup until the middle of the nineteenth century, when the city of Shanghai became a treaty port and the population started to expand, as a consequence of immigration from outside Shanghai and especially from nearby provinces, such as Jiangsu and Zhejiang.

Since then, the City subgroup diverged from the Songjiang subgroup, since it has been largely influenced by migrant dialects, as well as by Standard Chinese, which is used as a *lingua franca* among speakers of different dialects. Meanwhile, the other varieties of the suburban areas have undergone much less rapid changes. This is a well described sociolinguistic observation: the urban variety is more evolutionary and plays an important role in the spreading of linguistic innovations (e.g., Chambers & Trudgill, 1990: 189ff).

For example, the tonal system is the simplest in the City subgroup, in that, first, the number of tones is only five, compared to the more conservative suburban groups that have from six to nine tones; second, the tone sandhi patterns are simpler than the suburban groups, as will be explained in §2.1.3.2.

In this study, we focus on the City subgroup variety. Shanghai Chinese, without specification, will designate the City subgroup Shanghai Chinese. This is also how Shanghai speakers define "Shanghai Chinese": they use otherwise 本地闲话 'local speeches, patois' [pəŋ˧.di˥.ɦɛ˧.ɦu˩] when referring to the suburban varieties.

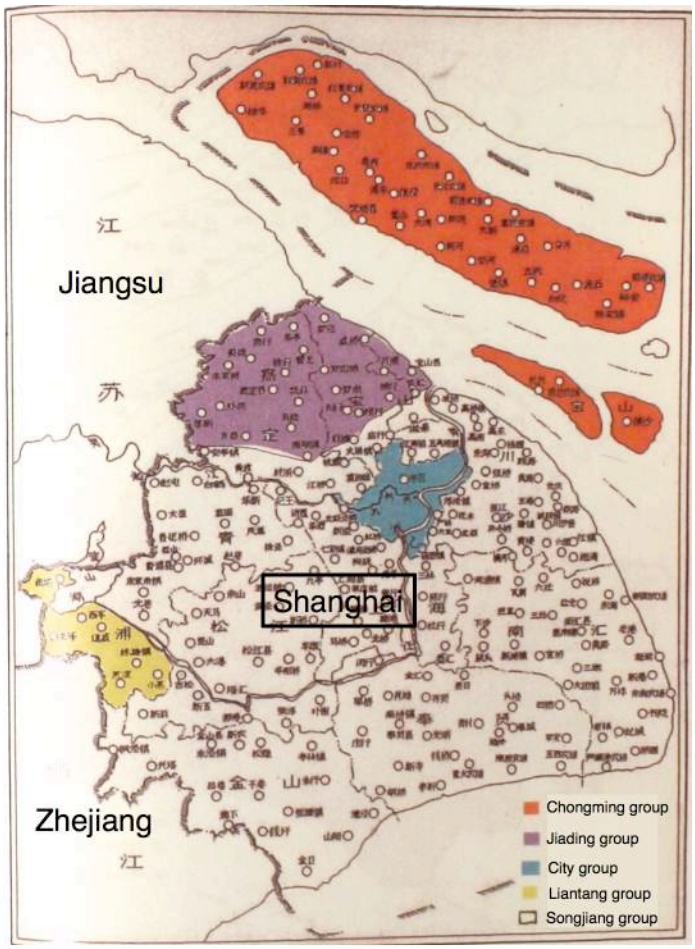


Figure 3. Map of dialect subgroups in Shanghai according to Chen (2003), the City group in blue.

1.3.2 Generational variation

For the City subgroup Shanghai Chinese, Xu and Tang (1988) defined three present-day generational varieties: the Old variety, the Middle variety and the New variety. This division is mainly based on variations in phonetics and phonology. According to Qian (2003), the Old variety is used by speakers born before the 1930s, the Middle variety by speakers born between the1940s and the 1960s, and the New variety by speakers born between the 1970s and the 1990s.

Three major characteristics, among others, distinguish the Old variety from the other two. Firstly, the Old variety maintains the /s/–/ɕ/ and /z/–/ʑ/ dental–alveopalatal contrasts before the vowel /i/, whereas the other varieties have lost these contrasts;

secondly, the Old variety distinguishes the two tone categories *yin shang* 阴上 (44)[2] and *yin qu* 阴去 (35), whereas these two tone categories merged and are both realized as 35 in the other varieties, thus reducing the number of tones to five against six in the Old variety; thirdly, the sandhi pattern is more complex in the Old variety than in the other two varieties (for sandhi pattern of the Middle and New varieties, see §2.1.3.2).

The characteristics that distinguish the Middle variety from the New variety are the following: in the New variety but not the Middle variety, the rimes /ɛ/ and /œ/ start to merge; /aN/ [ã] and /ɑN/ [ɒ̃] have merged, as well as /əʔ/ and /aʔ/.

The evolution from the Old to the Middle variety could be attributed to language contact with other migrant dialects such as the dialects of the Zhejiang (especially Ningbo) and Jiangsu provinces (especially Suzhou), whereas the evolution from the Middle to the New variety could be a consequence of the promotion of Standard Chinese in Shanghai area.

This division dating back from several years, speakers born after the 1990s are not taken into account. Based on our own observations, the youngest Shanghai speakers are even more influenced by Standard Chinese. For example, the loss of the /ŋ/ onset, which is common in the Old to New varieties of Shanghai Chinese but is not permissible in Standard Chinese, is quite systematic in the productions of these young speakers.

[2] In CHAO Yuen-Ren's (1930) tone notation.

2 TODAY'S SHANGHAI CHINESE: PHONETIC AND PHONOLOGICAL SYSTEM

Summary

This chapter describes today's Shanghai Chinese of the City subgroup. We first present the syllable structure of Shanghai Chinese, with its onset inventory (§2.1.1), its rime inventory (§2.1.2), its tone inventory and its tone sandhi rules (§2.1.3).

We then propose a first discussion of the phonological interdependence between tone and other glottal settings described in the literature. The phonological relationship between tone and voicing is discussed in §2.1.4.1: voiced onsets co-occur with the low tone register and voiceless onsets with the high tone register. The phonological relationship between tone and laryngealization is discussed in §2.1.4.2: checked tones co-occur with the final glottal stop (i.e., with laryngealization). The less described co-occurrence between rising tone and final glottal stop may also be phonologized and arise from historical evolution. But for the moment, we cannot exclude the possibility that this factor is due to a phonetic mechanism.

We attribute the above-mentioned phonological interdependence principally to the diachronic evolution of the tone. We compare non-Wu dialects that have undergone tone split and voicing loss with Wu dialects that have undergone the same tone evolution but have preserved the voicing contrast (§2.2.1). Two types of tone development from Old Chinese to today's Chinese are presented (§2.2.2). We especially focus on the tone split theory, according to which the separation into high and low tone registers originates from the voicing of syllable onset. Two points of view are exposed to explain tone split: consonant-based (related to consonants' voicing) and phonation-based (§2.2.3).

2.1 Syllable structure

Syllable structure is quite similar in Shanghai Chinese and in Standard Chinese. It is composed of (1) an optional onset that can only be a simple consonant, (2) a mandatory nucleus, which can be a monophthong, an opening diphthong ([j, w, ɥ] glide followed by non-high vowel), or a syllabic nasal [m̩] or [ŋ̩], (3) an optional coda, which can be a glottal stop [ʔ] or a nasal [N] unspecified for place of articulation, and (4) a lexical tone carried by the rime. Syllables ending with a glottal stop are called checked syllables, and they can only carry Tone 4 or Tone 5, whereas unchecked syllables can only carry Tone 1, 2 or 3. For descriptions of the tone system, please refer to §2.1.3.

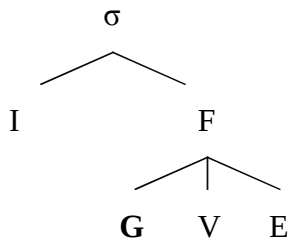
The syllable structures allowed in Shanghai Chinese are (C)V, (C)VC, (C)GV, (C)GVC, and N, where C stands for Consonant, G for Glide, V for Vowel, and N for syllabic nasal. As in almost all Chinese dialects, the combinations of these four elements undergo many restrictions.

In the case of rimes with an opening diphthong, there has been much debate about the status of the on-glide [j, w, ɥ], or the traditionally called "medial vowel" in many Chinese dialects. Most of these discussions concern Mandarin Chinese. The several hypotheses that have been proposed can be summarized as follows, as further shown in Figure 4:

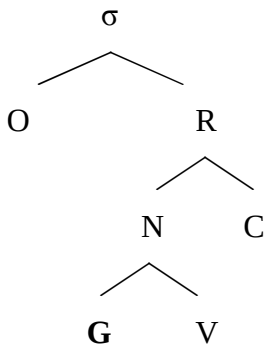
(a) In a traditional approach, a syllable is composed of an optional Initial (I), which can only be a simple consonant (i.e., not a consonant cluster), and a Final, composed of an optional "Medial vowel" (or on-glide), a mandatory Main vowel, and an optional Ending (E), which is an off-glide or a simple consonant (e.g., Chao, 1968). The Medial vowel is thus more closely related to the rime (Main vowel and optional Ending) than to the Initial.

(b) In a hierarchical onset-rime syllable structure, the on-glide precedes the main vowel to form the nucleus, which precedes the coda to form the rime of the syllable (e.g., Wang & Chang, 2001). Again, the on-glide is regarded as part of the rime.

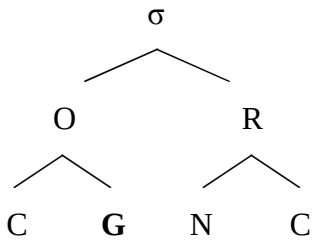
a. Initial-Final: part of Final



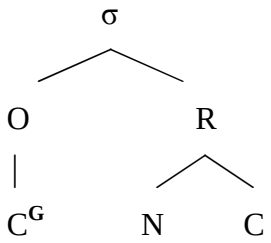
b. Part of Rime



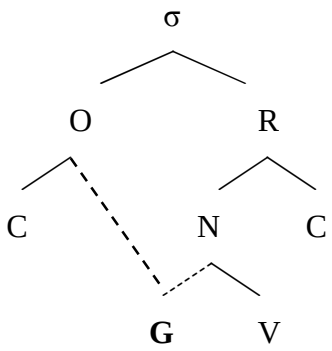
ci. Onset cluster



cii. Onset secondary articulation



di. both Onset and Rime



dii. neither Onset or Rime

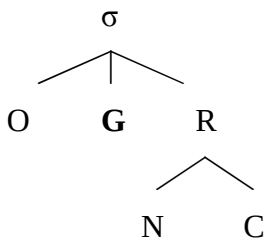


Figure 4. Schemas representating the main proposals for the phonological status of on-glides in Mandarin Chinese. (The first C for the Initial Consonant, the second C for the Coda.)

(c) The on-glide belongs to the onset, with two possible scenarios: (ci) it forms an onset cluster with the preceding consonant (Bao, 1990); (cii) it is a secondary articulation of the Onset (Duanmu, 1990).

(d) The on-glide doesn't belong exclusively to the onset or exclusively to the rime. It depends on the consonant that precedes the glide, and also on the nature of the glide (Wan, 1997; Bao, 1996); alternatively, it belongs to both the onset and the rime as in (di) (Yip, 2003), or to neither of the two, but directly to the syllable node instead,

as in (dii) (Yip, 2003). For Yip (2003) and other authors (eg., Zhang, 2006), the boundary between onset and rime is not consistent, which calls into question the onset-rime structure.

An interesting case is the Cantonese velars with secondary labialization /kʷ/ and /kʷʰ/, considered by most linguists as two phonemes different from non-labialized velars /k/ and /kʰ/ (e.g., Bauer & Benedict, 1997), and adopted by the official Cantonese Romanization system *Jyutping* 粵拼 (http://www.lshk.org/) (as "gw" and "kw", e.g., "gwaa" for 瓜[kʷa] and "kwaa" for 夸[kʷʰa]), although some linguists suggest they should be analyzed as /k/ and /kʰ/ followed by the glide /w/ (Kao, 1971).

Eric Zee, a Shanghai-born Hong Konger, who has carried out many studies on Cantonese, was one of the few linguists who analyzed Shanghai Chinese in the same way as Cantonese. For him, in Shanghai Chinese, /kʷ/ and /kʷʰ/ are two phonemes, not /k/ and /kʰ/ followed by /w/. On the other hand, Zee analyzed [j] as a glide belonging to the rime, as part of a diphthong (Zee, 2003), for example in the syllable 表 [pjɔ] 'watch'.

Eric Zee's analysis is supported by phonological and cross-linguistic arguments. First, phonologically, in Shanghai Chinese, [w] can only be preceded by velar stops, which are thereby labialized, whereas [j] can be preceded by a much greater number of initial consonants. This suggests that [w] has tighter phonotactic restrictions with the initial consonant, so it is economic to attach [w] to a velar stop onset. Note that, since there is a general ban of clusters in onset position, [w] should be analyzed as a secondary articulation of the velar stop. Second, cross-linguistically, many languages have labialized velar series, but much fewer have labialized bilabial or alveolar series. The UPSID database <UCLA Phonological Segmental Inventory Database> [3] (Maddieson, 1984; Maddieson & Precoda, 1990) contains 71 languages with labialized velar stops such as [kʷ], but only 6 languages with labialized bilabial stops such as [pʷ], and only 3 languages with labialized dental/alveolar stops such as [tʷ].

Despite these convincing arguments, we prefer to present the phonological system in a symmetrical and more traditional way, that is, to consider both [j] and [w]

[3] We consulted the simple web interface of the latest available version of the UPSID database by Henning Reetz: http://web.phonetik.uni-frankfurt.de/upsid.html

as glides that belong to the rime. We simply opted for grouping the on-glide and the following vowel together under the nucleus, as in model (b). We will not go into details of phonological and psycholinguistic arguments for each of the hypotheses, for it is not the purpose of this study to clarify this situation. At any rate, our materials only contain syllables without prenuclear glide.

2.1.1 Onset

The onset position, when not empty, is occupied by a simple consonant in Shanghai Chinese, if we consider the on-glide to be part of the rime. Table 1 lists the consonant phonemes and allophones in Shanghai Chinese.

Table 1. Consonants in Shanghai Chinese (shaded cells for allophones).

	Bilabial	**Labio-dental**	**Alveolar**	**Alveopalatal**	**Velar**	**Glottal**
Stop	p p^h b		t t^h d		k k^h g	ʔ
Fricative		f v	s z	ɕ ʑ		h ɦ
Affricate			ts ts^h (dz)	tɕ $tɕ^h$ dʑ		
Nasal	m		n	ɲ	ŋ	
Liquid			l			

As explained earlier, labialized velar stops [k^w k^{wh} g^w] are not analyzed as independent phonemes in our study, contrary to the analysis of Zee (2003).

The alveopalatals obstruents and the palatal nasal are often considered as allophones of the corresponding alveolars. The alveolar obstruents and nasal /s z ts ts^h dz n/ are realized as the alveopalatal obstruents and palatal nasal [ɕ ʑ tɕ $tɕ^h$ dʑ ɲ], respectively, before the high front vowels /i y/ and as alveolars before the other vowels. Note that the Old Shanghai variety distinguishes these two series before front close vowels (e.g., [tsi] 尖 'sharp' vs. [tɕi] 鸡 'chicken').

2.1.2 Rime

The rime position can be occupied by:

(1) V: a single vowel, among the vowels listed in the vowel chart in Table 2;

(2) GV: an opening diphthong (vowel preceded by an on-glide [j w]), namely [ja jɔ jɤ wa wɛ];

(3) (G)VN: a vowel among /i ɛ ɔ a/, or an opening diphthong among [ja jɔ wa wɛ], followed by a nasal coda /N/, unspecified for place of articulation, the phonetic realization of /ɔ a/ being strongly nasalized when followed by the nasal coda;

(4) (G)V?: a vowel among /i a u y/, or an opening diphthong among [ja wa] ([ja?] now being merged with [ɪ?]), followed by a glottal stop /?/ in a checked syllable;

(5) N: a syllabic nasal /N/ unspecified for place of articulation, only when there is no onset.

Table 2. Vowels in Shanghai Chinese (shaded cells for allophones).

	Front	**Central**	**Back**
Close	i y	ʐ̩	u
	ɪ ʏ		ʊ
Upper-mid	ø		o ɤ
Lower-mid	ɛ	ə	ɔ
		ɐ	
Open		a	

Several points are worth mentioning:

(1) The close central vowel [ʐ̩] is produced with the tip of the tongue and thus called “apical” vowel. “Apical” vowels are also found in Mandarin and some other Chinese dialects. Sinologists usually use the symbol [ɿ] for apical vowels preceded by an alveolar affricate or fricative and the symbol [ʅ] for apical vowels preceded by a post-alveolar (sometimes defined as “retroflex”) affricate or fricative. The

International Phonetic Alphabet uses the syllabic fricative and approximant [z̩ ɹ̩], respectively. Since [z̩] only occurs after an alveolar fricative or affricate [s z ts tsʰ dz], some authors proposed not to consider [z̩] as a phoneme but, rather, as an allophone of /i/, based on the observation that [i] cannot be preceded by alveolar fricatives or affricates [s z ts tsʰ dz]. Yet, this account is difficult to reconcile with the fact that /i/ can be preceded by alveopalatal fricatives or affricates [ɕ ʑ tɕ tɕʰ dʑ], if we consider these alveopalatals as allophones of the corresponding alveolars (see §2.1.1), unless [z̩] is considered as an independent phoneme (however, it is impossible to produce this sound in isolation). An alternative account is to consider [z̩] as a phonetically voiced syllabic prolongation of alveolar fricative or affricate onsets.

(2) The front mid vowels are often transcribed as the lower-mid [ɛ] and upper-mid [ø], but their openness is actually between upper-mid and lower-mid.

(3) The distinction between the upper-mid /e/ and the lower-mid /ɛ/ in the first half of the 20th century, the latter evolved from a previous nasal vowel, was neutralized in the 1960s and 1970s (Qian, 2003). Recently, however, young speakers are beginning to make a new vowel split between [e] and [ɛ], probably due to the influence of Mandarin Chinese, in which the rime [ei] corresponds to Shanghai Chinese [e] and the rimes [ai] and [an] correspond to Shanghai Chinese [ɛ] (Qian, 2003). The upper-mid [e] is even slightly diphthongized as [eⁱ] by some Shanghai speakers. This is in line with Zee (2003), who notes a closing diphthong [ej] for the original /e/ vowel, reflecting such diphthongization.

(4) As for the back vowels, both Svantesson (1989) and Chen (2008a) found an overlap in the F1/F2 space between the vowels /u/ and /o/. According to Chen (2008a) and our own observation, /o/ is more protruded and rounded than /u/.

(5) The close vowels /i y u/ are realized as lax vowels [ɪ ʏ ʊ] in closed syllables (with nasal coda or final glottal stop).

(6) /a/ and /ə/ were distinctive when followed by /ʔ/, that is in checked syllables, in the Old Shanghai variety. But the merger between the two was almost accomplished

in the Middle and New varieties, both realized as [ɐ]. /ə/ followed by a liquid sound /l/ is used for literary reading, as an equivalent of the [ɚ] rime in Mandarin.

2.1.3 Tone

Shanghai Chinese has five lexical tones, including three unchecked tones and two checked tones. The checked tones are carried by syllables ending with a glottal stop. Thus, according to some analyses, the checked tones are not phonological but determined by the segmental context.

In Mandarin Chinese, the basic tone contour is carried by the syllable rime, without the prenuclear on-glide, and the F0 contour in the consonantal onset part (including the prenuclear on-glide) may be viewed as merely anticipatory adjustments, as demonstrated by acoustic data (Howie, 1974), as well as by electromyographic (EMG) data (Hallé, 1994), although it is found that the F0 in the consonantal part carries important information for tone perception (Chen & Tucker, 2013). According to our preliminary observation on syllables with zero and nasal onsets, just as in Mandarin Chinese, it is the syllable rime that carries the basic tone contour in Shanghai Chinese.

2.1.3.1 Citation tones

The lexical tones realized on isolated monosyllables are called citation tones, labeled T1 to T5. According to Sherard's (1972) descriptions, T1 is "sharply falling," T2 is a "moderately high level tone" with a "short audible rise in pitch toward the end" in "careful, overprecise speech," T3 is a tone "in low register" with a "clearly audible tone contour with end point higher than onset," T4 has a "moderately high pitch with no discernable change of pitch" … "combined with complete glottal closure," and T5 is "low register [tone] throughout with a short rise in pitch, ending with a glottal stop."

Slightly different tone transcriptions of Shanghai tones can be found among linguists, some broader than others. They are summarized in Table 3 (based on Rose, 1993, and Zhu, 1999: 19). Underlined or single values indicate short tones. Chao

(1928), in his survey of Wu dialects, used the numerical tonic sol-fa system, a pitch scale divided into semitones with the middle point set as *mi flat* (3^{b}) in the scale. Many other linguists use the better-known Chao's (1930) five tone letters system to describe the tones of Shanghai and other Chinese dialects. Rose (1993) used the standard deviation scale who was quite similar to Chao's scale, but with three more points ↑4, ↑3, and ↑2, which are about half a standard deviation above 4, 3 and 2, respectively. Zee and Maddieson's (1979) description seems broader in that it uses the phonological tone symbols H, L, and M, but augmented with arrow diacritic symbols. Their scale consists of "H, M↑ (raised M), M, L↑ (raised L), L" is thus actually similar to Chao's five tone letters scale. All these notations are auditory and/or impressionistic descriptions, not acoustic measurements. Based on the analysis of tone sandhi patterns, Eric Zee proposed both a phonological notation (within slashes) of the five "types of lexical tone melody" and a phonetic notation (within brackets) of the five tones' citation forms (Zee, 2003).

Table 3. Transcriptions of the five citation tones of Shanghai Chinese used by different authors.

	T1	**T2**	**T3**	**T4**	**T5**
Chao (1928)	41	$3^{b}2^{b\,2}$ ~ $^{2}\,2^{b}3$	13^{b}	$4^{\#}$	23
Anonymous[4] (1960)	53	34	13 or 14	5	2
Norman (1988)	42	35	24	55	23
Xu & Tang (1988)	53	34	23	55	12
Zee & Maddieson (1979)	HL	MM↑	LM↑	H	LM↑
Rose (1993)	51	43↑3	22(2)↑3	↑44 or 54	↑22↑3
Zhu (1999)	52 or 42	334	113 or 14	34 or 44 or 54	223 or 24
Zee (2003)	[51] /51/	[34] /351/	[13] /151/	[5ʔ] /451/	[12ʔ] /12/

[4] Editors of *A survey of the dialects of Jiangsu and Shanghai*, whose names remain unknown. Shi Rujie 石汝杰, Professor of Chinese linguistics, announced in his blog (http://shirujie2010.blog.sohu.com/237754175.html), citing Professor Yan Yiming 颜逸明, that the four chief editors were said to be Shi Wentao 施文涛, Bao Mingwei 鲍明炜, Ge Yiqing 葛毅卿, and Ye Xiangling 叶祥苓.

It is the relative values, but not the absolute values, that are the most informative. As Zhu (2006) suggested, there are in fact phonetic variants for each tone category. Inspired by Rose (1988), he proposed three parameters for the Shanghai tone specification: Register (+U(pper) for *yin* tones, and –U(pper) for *yang* tones), Length (long for unchecked tones, and short for checked tones) and Contour (hl (high low) for T1 and lh (low high) for the other tones), as shown in Table 4. In our work, we follow Xu and Tang's (1988) notation for Shanghai tones, and thus note T1 as 53, T2 as 34, T3 as 23, T4 as 55, and T5 as 12.

Table 4. Three parameters specifying the five Shanghai tones, after Zhu (2006: 18).

		Length		
		Long		Short
Register	+U	T1	T2	T4
	-U		T3	T5
		hl	lh	
		Contour		

Chinese tones are traditionally (i.e., in the Chinese linguistic tradition) divided into four categories, called *sheng* 声: *ping* 平, *shang* 上, *qu* 去, *ru* 入. Each category is further separated into two registers, called *diao* 调: *yin* 阴 and *yang* 阳. These eight labels *yin ping*, *yang ping*, *yin shang*, *yang shang*, *yin qu*, *yang qu*, *yin ru*, and *yang ru* are inherited from two waves of tone development, which will be explained in §2.2.2. Modern Cantonese has preserved all of the eight labels and even further divided the *yin ru* category into two sub-categories conditioned by vowel duration. But in the course of tone evolution in the various dialects, the separation of the two registers was not homogeneous across all the dialects, and also, mergers of categories and/or registers occurred in many dialects. All Wu dialects have preserved the two registers *yin* and *yang*. But mergers of categories can be found in the modern City subgroup variety of Shanghai Chinese, in which the *yin shang* and *yin qu* categories are neutralized, as well as the *yang ping*, *yang shang* and *yang qu* categories. Note that *yin shang* and *yin qu* were still distinctive in the Old variety but merged in the Middle and New varieties. The City subgroup underwent more mergers than other dialect

subgroups of Shanghai, resulting in a system of five tones, as shown in Table 5: *ying ping* (T1), *yin shang*/*qu* (T2), *yang ping*/*shang*/*qu* (T3), *yin ru* (T4), *yang ru* (T5). Other Shanghai subgroups have more conservative systems with more tones. In the Songjiang subgroup of Shanghai, for example, the eight categories are all maintained.

Table 5. Historical tone labels of Shanghai tones.

	Ping 平	**Shang 上**	**Qu 去**	**Ru 入**
Yin 阴	T1 (53)	T2 (34)		T4 (55)
Yang 阳	T3 (23)			T5 (12)

2.1.3.2 Tone sandhi and stress

Traditionally, two types of tone sandhi are distinguished in Shanghai Chinese: the left-dominant process (in the Chinese terminology, 广用式 'broad style') and the right-dominant one (窄用式 'narrow style') (e.g., Xu & Tang, 1988; Li, 2009). The left-dominant sandhi applies in the domain of the prosodic word (compounds in most cases but not uniquely), whereas the right-domain sandhi applies in the domain of the syntactic phrase.

The left-dominant sandhi in a prosodic word is accomplished by the spreading of the contour of the leftmost syllable to the following one, and the insertion of a default tone, phonetically implemented as a low tone, on the remaining syllables, except when the first syllable carries the lexical tone T5. Table 6 lists the phonetic tone values of polysyllabic words with the application of left-dominant sandhi rule.

Sherard (1972) was one of the first to provide examples of disyllabic words to illustrate Shanghai sandhi. He explained that the lexical tone of the initial syllable determined the tone contour of the entire word. More precisely, the tone contour of the word is essentially the same as that of the initial syllable but longer in duration. He did not consider the cases of words of more than two syllables. More recent studies suggest that the contour tone should be decomposed into level tones H and L, and it is the second level tone that is spread from the initial syllable to the following one and

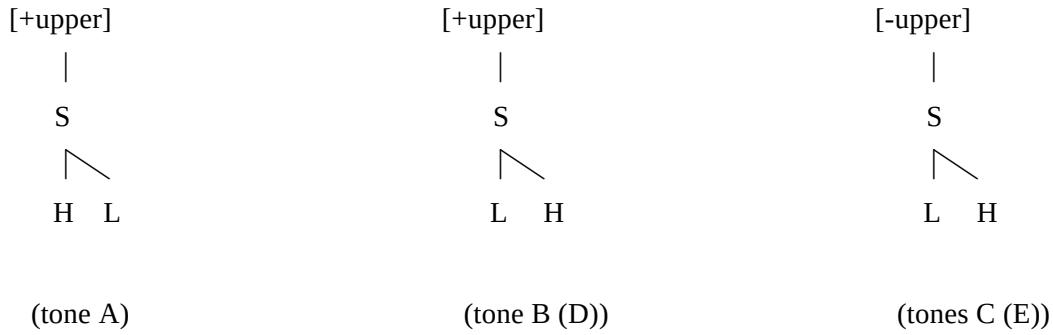
dissociated from the initial syllable (Zee & Maddieson, 1979; Yip, 1980; Wright, 1983; Selkirk & Shen, 1990; Chen, 2008b).

Table 6. Tone values of left-dominant sandi rules for 2-to-4 syllable words, based on Zhu (2006: 38ff). T5 I and T5 II indicate the two possible sandhis after T5 for 4-syllable words. (Underlined values indicate the syllable shortness.)

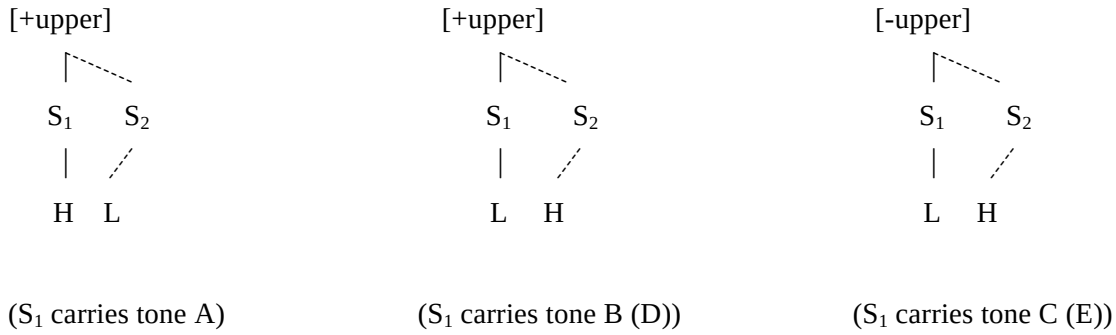
S1 tone	Disyllable	Trisyllable	Quadrisyllable
T1	55 + 22	55 + 44 + 22	55 + 44 + 33 + 22
T2	33 + 44	33 + 44 + 22	33 + 44 + 33 + 22
T3	11 + 44	11 + 44 + 11	11 + 44 + 33 + 11
T4	33 + 44	33 + 44 + 22	33 + 44 + 33 + 22
T5	11 + 24	11 + 11 + 24	11 + 22 + 22 + 24 (T5 I) 22 + 44 + 33 + 11 (T5 II)

Yip (1980: 194ff) proposed a tonal inventory of three tones: A, B, and C. The checked tones D and E are considered as short variants of tones B and C, respectively. These three tones combine the tonal register feature [upper] and the melodic features H and L: tone A [+upper, HL], tone B (or D) [+upper, L], tone C (or E) [-upper, LH]. We propose to change the [L] feature of tone B (D) to [LH], not only because tone B is assumed by many authors to be a rising tone (checked tone D is not, however), but also because the tone sandhi process can be better accounted for with the [LH] feature. Zee and Maddieson (1979) also proposed that tone B is rising (they noted it [LM↑]), but proposed [H] for tone D. Using an autosegmental representation (Goldsmith, 1976), Shanghai tones can thus be represented as shown in Figure 5, with three tiers: the register tier, the segmental tier, and the tone-feature tier. The left-dominant tone sandhi is the output of three ordered rules: (1) Tone deletion on non-first syllables; (2) Right spreading of the second level tone of the decomposed contour tone of the first syllable; and (3) Default tone insertion on the remaining unlinked syllables.

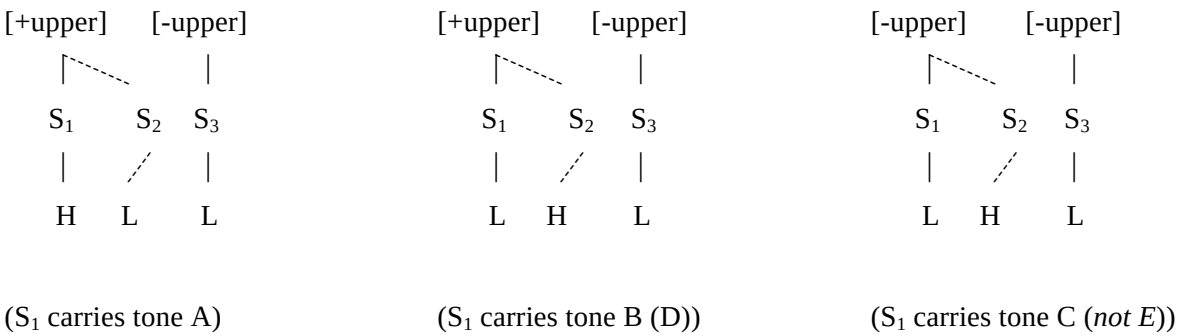
Monosyllables



Disyllabic words



Trisyllabic words



Quadrisyllabic words

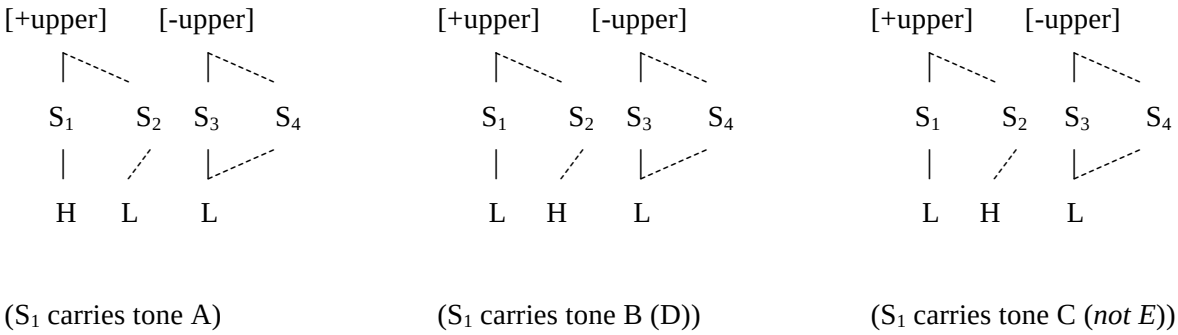


Figure 5. Autosegmental representation of tones of polysyllabic words.

For disyllabic words, when S2 originally carries Tone D, the tone height of S2 is observed by Zee and Maddieson (1979) to be higher than when S2 originally carries another tone than D. (This is observed regardless of S1's tone, except tone A.) Yip (1980: 202-203) claimed that, in this specific case, Tone D is not deleted in S2 and proposed a specific pattern for S2 carrying Tone D. However, we are tempted to interpret the contour of Tone D as phonetically motivated. First, checked syllables may be realized with a higher F0 due to laryngealization. On this account, both Tone D and Tone E syllables tend to be intrinsically higher than Tone A-C syllables. Second, (*yin*) Tone D syllables are realized with a higher F0 than (*yang*) Tone E syllables presumably because they are a voiceless, whereas Tone E syllables are phonetically voiced in word-medial position. Indeed, an F0-lowering effect of voiced onsets in word-medial position is observed in previous studies (e.g., Ren, 1992: 51; Chen, 2011) and in our data (see §4.1.6.1.4).

A particular pattern concerns words whose initial syllable carries Tone E ([-upper], LH). As shown by the phonetic tone values in Table 9, it seems that not only the rightmost H feature but also the L feature spread rightward, resulting in a LH (instead of H) tone on S2 in disyllabic words, and in L and LH tones (instead of H and default L) on S2 and S3, respectively, in three-syllable words. This pattern is described as specific to Shanghai tone sandhi, and is rare in northern Wu (Zhu, 2006: 40). The phonological accounts by Zee and Maddieson (1979) and Yip (1980) of this pattern involve special rules for initial T5 syllables and are not quite satisfying in our view. Perhaps we still need further investigations. Or we need another phonological approach. This is, however, out of the scope of the present work.

In contrast with the left-dominant sandhi, the tone contour of the syntactic phrases to which the right-dominant sandhi rule applies is determined by the rightmost syllable. In a disyllabic phrase, for example, the rightmost syllable keeps its original tone and the first syllable is linked with a level tone, whose height depends on the register of its original lexical tone. According to Zhu (2006: 46), the phonetic tone values of T1 to T5 in S1 position in a right-dominant sandhi phrase are respectively 44, 44, 33, 44, 22. The tone contour of phrases with more than two syllables also depends on the syntactic structure of the phrase, and will not be presented here. Two things are worth to be noted. Firstly, a sequence of morphemes can sometimes be

defined as a compound word and sometimes as a phrase. Their tone values are thus different. For example, 炒饭 /tsʰɔ34.vɛ23/ realized as [tsʰɔ33.vɛ44] when following the left-dominant sandhi rule means "fried rice," while the same morphemes realized as [tsʰɔ44.vɛ23] when following the right-dominant sandhi means "to cook rice." Secondly, that the type of sandhi be solely determined by the syntactic analysis of a sequence (compound word vs. syntactic phrase) is probably an oversimplification. The relationship between syntactic and prosodic structures is too complexe and debatable to be presented in this work.

Based on the fact that the left-dominant sandhi is the dominant sandhi pattern in Shanghai Chinese, even across Northern Wu dialects, the stress pattern in these dialects are often refered to as a "heavy-light" pattern (Ballard, 1989: 105). More specifically, at the word/compound level, the stress is assigned leftmostly (Chen, 2000: 309).

2.1.4 Interdependence between tone and other glottal settings

In the studies of Chinese dialects, a traditional approach consists in correlating the modern tone categories with the historical categories. According to Sherard (1972), the focus on this historical correlation, rather than on synchronic description, is one of the "deficiencies" of the traditional approach of dialectal study. This is quite true. But still, this approach is necessary to understand not only the evolution of tones in Chinese languages, but also what determined the phonetic correlates of tones. Do these phonetic correlates have an articulatory or a perceptual motivation? Or can they be explained by diachronic changes? The *yin-yang* register separation and the *ru* category (entering tones) in Shanghai Chinese are remnants of the diachronic evolution of Chinese languages (see §2.2.2). This led to a synchronic relation between the *yin-yang* register and the voicing distinction, as well as between the *ru* category and the laryngealization.

2.1.4.1 Tone and voicing

In Shanghai Chinese, the voicing distinction and the tone distinction are closely related. Voiceless onsets co-occur with the *yin* register and voiced onsets with the *yang* register. In other words, syllables with voiceless obstruent onsets may only carry tone T1, T2 or T4, whereas syllables with voiced obstruent onsets may only carry tone T3 or T5, as illustrated in Table 7. The historical aspect of the tone-voicing relation will be developed in §2.2.2. With respect to voicing, three issues must be noted. Firstly, we only refer here to the phonological voicing distinction, regardless of its phonetic realization. Chapter 3 will give a detailed phonetic description of voiceless vs. voiced onsets. Secondly, sonorant onsets are phonetically voiced, as is very common cross-linguistically, but can co-occur with either *yin* or *yang* tones, although nasal and liquid onset syllables carrying *yin* tones are relatively rare. As far as we know, no clear explanation has been offered why *yin* tones may co-occur with sonorants onsets. According to Zhengzhang (2000: 65), the glottal stop in coda position of Old Chinese, which later led to the rising tone, might cause sonorants onsets to become glottalized and therefore behaving as voiceless onsets in the tone register evolution (see §2.2.2) whereas the other sonorant onsets behave as voiced onsets tonally. According to Baxter and Sagart (2014: 171-173), the sonorants that were preceded by voiceless pre-initial consonant in Old Chinese occur with *yin* tones in Min, Hakka and Cantonese: for example, 毛, Old Chinese *C.m^{ʔ}aw > Cantonese mou (*yin* tone T1). Thirdly, voicedness and *yang*, that is, low tones seem to be associated perceptually in Shanghai speakers' mind. This is hard to show from the reports of naïve Shanghai speakers/listeners who have little or no idea of what is voicing, but phoneticians/phonologists who are speakers of Shanghai Chinese associate voiced onsets with low tones perceptually. In a perception test conducted by Cao (1987), phonologists and trained phoneticians familiar with Wu dialects perceived low tone syllables in Shanghai and Mandarin Chinese as produced with pre-voiced onsets. Note that the Mandarin Chinese has *no* phonological voicing distinction. REN Nianqi, a trained phonetician (and a talented singer), reported the following anecdote in his PhD dissertation (Ren, 1992: iii): When asked about the stop consonants in his native language, during the phonetics class taught by Professor Arthur Abramson, he

answered with confidence that one Shanghai stop series was voiced unaspirated; after listening to these sounds, Prof. Abramson judged that they were not voiced phonetically. This anecdote motivated Ren's excellent fiberoptic and acoustic study on Shanghai stops under the supervision of Prof. Abramson.

Table 7. Co-occurrence between onset voicing and tone in Shanghai Chinese.

Onsets	**Tones**
p, p^h, t, t^h, ts, ts^h, tɕ, $tɕ^h$, k, k^h, f, s, ɕ, h, l, m, n, Ø	T1 T2 T4 (Yin tones)
b, d, (dz), dʑ, g, v, z, ʑ, l, m, n, Ø	T3 T5 (Yang tones)

2.1.4.2 Tone and laryngealization

The term "laryngealization", as we use it in this work, covers all voice qualities involving laryngeal constriction, including creaky voice, glottalization, and glottal stop.

In addition to the final glottalization of the falling T1 syllables, probably due to a phonetic mechanism accompanying very low F0, as found in the low 3rd tone and the falling 4th tone of Mandarin Chinese (Belotel-Grenié & Grenié, 1994, 1995), there are two other types of laryngealization correlated with tones in Shanghai Chinese.

The first type concerns the *ru sheng* (入声). The *ru sheng* (entering tone, also called "checked tone") is better defined by its segmental than tonal characteristics, at least originally. In Middle Chinese, an entering tone is carried by a syllable ending with a voiceless stop /p, t, k/. A syllable carrying a checked tone is called a "checked syllable." The *ru sheng* is traditionally considered as a separate tone category on the basis of its phonetic difference with the other tone categories. Phonologically, the syllable stop codas [p, t, k] can be analyzed as allophones of the nasal codas /m, n, ŋ/ which surface in [m, n, ŋ] in syllables carrying non-*ru* tones. From a different viewpoint, a complementary distribution could also be established between the *ru* category and the other three categories on the basis that *ru* tone syllables end with /p, t, k/ whereas syllables carrying the other tones never end with /p, t, k/.

In any case, the stop codas of Middle Chinese *ru sheng* syllables evolved in different ways in different dialects. They are lost in Standard Mandarin and most Mandarin dialects, but well preserved in Cantonese, Min and Hakka. In Wu dialects and Jianghuai Mandarin 江淮官话 dialects, including Nanjing Mandarin 南京官话, the stop codas lost their original place of articulation and neutralized into a glottal stop (/p, t, k/ > /ʔ/). In the case of Shanghai Chinese, for example, T4 and T5 are *ru* (checked) tones carried by syllables ending with a glottal stop /ʔ/. This type of laryngealization thus refers to a phonologized feature of tones T4 and T5. Non-prepausal checked syllables, however, generally lose the glottal stop coda. Besides, checked syllables have another acoustic correlate: their duration is noticeably shorter than that of unchecked syllables.

For sake of coherence with the presentation of the Shanghai tone system, we will consider the *ru sheng* as a distinctive tone category. In this case, the laryngealization is phonologized, that is to say, systematically realized with the checked tones in a prepausal context.

The second type of laryngealization is not often described in the literature of Shanghai tones. If [ʔ] is generally described as occurring in coda position in checked tone syllables, we rarely find such description for the unchecked rising tones T2 and T3. However, as demonstrated by the acoustic data from Rose (1993) and my own unpublished fiberoptic data (with myself as the speaker), T2 and T3 syllables are produced with a clear final glottal stop. Phonetically, raised pitch involves tensed vocal folds, which may either lead to or be caused by laryngeal constriction, but not necessarily. Such mechanism is never reported on the rising tone of Mandarin Chinese, for example. John Esling (personal communication) suggested that the final glottal stop might also be the consequence of an utterance-like elicitation of the syllable. Under this interpretation, the glottal stop is utterance-final in nature rather than a characteristic of a lexical tone. We will learn in §2.2.2 that the Middle Chinese rising tone category *shang sheng* (上声) originated from the Old Chinese [ʔ] coda (see also Mei, 1970). Although today's Shanghai rising tones originate from two or three categories which have been neutralized, we are tempted to think that the final glottal stop could be a remnant of the original rising tone. Evidence comes from other modern

Chinese dialects largely distributed geographically, in which *shang sheng* (上声) ends with a glottal stop, such as Wenzhou 温州 dialects, the dialects of Huangyan 黄岩, Tiantai 天台 and Sanmen 三门 spoken in Taizhou 台州, quite a few Huizhou 徽州 dialects of south Anhui 安徽 and Jiangxi 江西, etc.; similar patterns are also found in related languages such as Vietnamese (Zhengzhang, 2000: 65).

2.2 Comparison with non-Wu dialects

2.2.1 Main difference: Phonological voicing distinction

The first scientific dialect study conducted by LI Fang-Kuei 李方桂 (1937) proposed two main criteria for Chinese dialect classification: (1) the evolution, from Middle Chinese to present-day dialects, of voiced initials, and (2) the evolution of final consonants and entering tone. Li (1937) thus based the classification of the Wu dialect group mainly on two criteria (1) its retention of the ancient voiced stops as *aspirated voiced consonants*, and (2) its retention of the "entering tone" as a short tone with a final glottal stop.

The triple series of stops as well as the double series of fricatives, both including a voiced series, have often been cited later as the major phonological characteristics of the Wu dialect group. Using this criterion to distinguish Wu from Non-Wu dialects is, however, problematic. From a general point of view, classification cannot be solely based on retention. Some specialists of historical phonology warn that "only innovations are useful in linguistic classification" (Sagart, 1998). More specifically, one counterexample for the voiced series retention criterion is that some Xiang dialects, which can hardly be related to the Wu family, also retained the voiced series of Middle Chinese.

If the phonological voicing distinction is a common feature of the Wu dialects, the phonetic realizations of the voiced series are not homogeneous among all the dialects. Norman (1988: 199) claimed that in southern Wu dialects, phonologically voiced stops are phonetically fully voiced, whereas in northern Wu dialects, voiced stops are phonetically voiceless in word- or phrase- initial position. CAO Zhiyun (2002), based

on a survey of Southern Wu dialects, suggested that voiced obstruents are realized as voiced in most southern dialects, including Wenzhou Chinese, but as voiceless in several other dialects.

Contrary to these claims, based on impressionistic descriptions, the acoustic data from CAO Jianfen and Maddieson (1992) showed that in Wenzhou Chinese, the VOT (voice onset time) of phonologically voiced stops is about zero in monosyllabic words. Rose (2002) also found that Wenzhou voiced stops are predominantly voiceless although there are more free variations than in Northern Wu dialects. While more and more acoustic data seem to confirm the realization of voiced stops as voiceless in Northern Wu dialects, among which Shanghai data are abundant, larger-scale acoustic data and analyses are still needed to investigate the phonetic realizations of the voiced series in Southern Wu dialects.

2.2.2 Other dialects: Tone split and loss of voicing contrast

Whereas the tone contrast and the voicing contrast coexist in Wu dialects, in a great majority of other dialect groups, the voicing contrast has been lost. This is one of the consequences of a tonal development called "tone split."

It is widely accepted that tone contrasts in languages of Asia originated from segmental contrasts. From the toneless Old Chinese to the present-day tonal Chinese, at least two types of tone development took place: the development of contour tones, and that of register tones.

The first development transphonologized segmental contrasts in syllable-coda position into tonal contrasts between four contour tones. It probably took place before the sixth century CE and characterizes the transition from Old Chinese to Middle Chinese. "Transphonologization" is a term used for diachronic sound changes, referring to a process through which a phonological contrast tends to be neutralized and replaced with another phonological contrast, compensating for the loss of the disappearing contrast (Hagège & Haudricourt, 1978: 75, see also Hyman, 2008). This process was already described by Roman Jakobson, using the term "retransphonologization" (Jakobson, 1931). As summarized in Table 8, in the course of

this first development, open syllables and syllables ending with a sonorant gave rise to the level tone; syllables ending with a glottal stop gave rise to the rising tone; syllables ending with /h/, which evolved from /s/, gave rise to the departing tone; finally, syllables ending with a /p, t, k/ stop gave rise to the entering tone. Laurent Sagart (1986, 1989), on the other hand, reconstructed the departing tone as arising from an ancient creaky phonation realized in a short syllable, based on the observation that such phonation is still preserved in a great number of Chinese dialects and Southeast Asian languages. Whereas the terms "level" and "rising" more or less indicate the phonetic realization of the level and rising tones, the other two terms are more difficult to interpret, although we are almost sure that the "entering" tone was short and ended abruptly. As mentioned in §2.1.4.2, the entering tone should be considered separately, because the /p, t, k/ endings did not disappear and still contrasted with sonorant endings or with open endings; the other coda consonants (presumably /ʔ/[5] and /h/) disappeared and gave rise to contour tones. Haudricourt (1954) provided a detailed account of a parallel evolution in Vietnamese and suggested that a similar evolution took place in Thai and Miao-yao (Hmong-Mien), probably under the influence of the prestigious Chinese culture. For alternative Chinese tonogenetic theories, see Sagart (1999) for an overview.

Table 8. Correspondence between Old Chinese syllable structures and Middle Chinese contour tones, based on Haudricourt's reconstruction.

Old Chinese syllable	Contour tone of Middle Chinese
* -V(N)#	level 平
* -ʔ#	rising 上
* -s# > -h#	departing 去
* -p,t,k#	entering 入

NB: * notes hypothetical reconstructions.

[5] In §2.1.4.2 we mentioned that the /-ʔ/ ending did not totally disappear in some Chinese dialects.

The loss of the voicing contrast in obstruents is related to the second development, which transphonologized the voicing contrast in syllable-initial position into a tonal contrast between two tonal registers,[6] increasing the number of tones from four to eight. It took place around the tenth century during the transition from Early Middle Chinese to Late Middle Chinese.

In Middle Chinese, consonants were divided into four series: *quan-qing* 全清 'full clear', *ci-qing* 次清 'secondary clear', *quan-zhuo* 全浊 'full muddy', and *ci-zhuo* 次浊 'secondary muddy'. In the *Yunjing* 韵镜, (literally "rime mirror," a rime table in which characters are sorted by syllable structure and tone, and whose date of composition remains unknown), these series were called *qing* 清 'clear', *ci-qing* 次清 'secondary clear', *zhuo* 浊 'muddy', *qing-zhuo* 清浊 'clear-muddy', respectively. The first three series qualify stops and affricates, whereas the fourth one covers all the sonorants. Linguists interpret the distinction between "clear" and "muddy" as a voicing contrast, "clear" standing for "voiceless" and "muddy" for "voiced." These two adjectives are still frequently used in today's linguistic descriptions. The difference between "full" and "secondary" clear is supposed to describe the aspiration contrast. Therefore, "full clear," "secondary clear," and "full muddy" defined the triple series of Middle Chinese stops and affricates: voiceless unaspirated, voiceless aspirated and voiced, respectively.

There was a time when the fundamental frequency (F0, or pitch) at the beginning of syllable's rime was affected by the syllable onset: voiceless onsets entailed higher-pitch and voiced onsets lower-pitch rime beginnings. During a certain period, this onset F0 perturbation was not perceptible. Probably, this is one of the reasons why Chinese scholars attached greater importance to the four contours (四声) than to the two registers with respect to the tone system in Middle Chinese. In the

---

[6] The term "register" is used for several different meanings: tonal register (high vs. low), phonation register (e.g., modal vs. breathy), vowel quality difference (e.g., tense vs. lax vowel), and intonation register. For the dependence and the difference between them in East Asian languages, see Yip (1993). For the relationships between them in Mon-Khmer languages, see §2.2.3.2. In this chapter, the "register" distinction refers essentially to the tonal height difference.

Qieyun 切韵 (literally "cut rimes," the first official rime dictionary edited chiefly by LU Fayan 陆法言 in 601 CE), characters were sorted according to the four contour tone (one or two volumes per tone), but the two registers associated with the onsets were not distinguished at all. In the *Yunjing* 韵镜, characters were dispatched into four large rows, each corresponding to one contour tone, and the separation between "clear" and "muddy" onsets was a sub-categorization under each onset's place of articulation but clearly not a tone distinction. Maspero (1912) pointed out that, in Middle Chinese, there existed a difference between the upper register, which was associated with voiceless onsets, and the lower register, which was associated with voiced onsets. He attributed Chinese scholars' neglecting the register difference to their theoretical background, narrowly focused on the four-tone system from the very beginning.

We still do not know the chronological relationship between the composition of the above rime dictionaries and the phonologization of the pitch difference on rime's beginning. However, we believe that this difference was purely phonetic and coarticulatory before phonologization occurred. Only later did this difference begin to be perceived and reproduced by listeners/speakers. The perception of this difference as a property of the sole rime, instead of a property common to both onset and rime, was the origin of the transphonologization. This scenario would be consistent with Ohala's view of historical sound changes: "A misparsing of the speech signal is a potential sound change," as Ohala (1993a, 1993b) showed for various cases of language evolution in several languages. The *ping* tone 平声 was probably the first tone that underwent the register separation. The labels of the two registers, *yin* 阴 for the upper register, and *yang* 阳 for the lower register, as mentioned in §2.1.3.1, were first employed by ZHOU Deqing 周德清 in his *Zhongyuan yinyun* 中原音韵, literally "phonology of the Central Plains," a rime table compiled in 1324 CE which arranged characters according to rimes and tones. The *ping* tone 平声 was divided into *yin* 阴 and *yang* 阳 in the rime table, but not the other three tones. The register separation should have been well established —at least for the *ping* tone 平声— before this date. When register separation was extended to the other three tones, the total number of

tones doubled. But it is only when the voicing contrast was lost by the devoicing of the voiced onsets that the transphonologization can be considered as completed.

This tonal development, called "bipartition tonale" ('two-level tone split') by Haudricourt (1961) is widespread among East Asian and Southeast Asian languages. Haudricourt gave an overview of languages from different families, such as Cantonese, S'gaw-Karen from Tibeto-Burman, Vietnamese, some Thai languages, some Miao-Yao dialects, etc. He also illustrated some cases of "tripartition tonale" ('three-level tone split') in languages that originally possessed three series of nasal onsets (aspirated, modal, and glottalized). In some languages, after the three nasal series neutralized, three register levels were created and the number of tones tripled.

Many of present-day Chinese dialects have undergone the tone split, and some have preserved two tonal registers. We will just take a look at Haudricourt (1961)'s illustration of the Cantonese tone system, presented in Table 9. In Wu dialects, the tone split took place as well. But the specific thing about the Wu dialects is that the voicing contrast was also preserved. Chapter 3 will account for how the voicing and the register contrasts coexist in Shanghai Chinese.

Finally, Table 10 offers a summary of the two tonal developments from Old Chinese to Late Middle Chinese explained in this section.

Table 9. Tone split in Cantonese conditioned by syllable onset, from Haudricourt (1961). The checked tones (*ru sheng*) are not illustrated.

Initials (onsets)	**level**	**rising**	**departing**
p, p^h, t, t^h, k, k^h	53	35	33
b > p ~ p^h, d > t ~ t^h, g > k ~ k^h, m, n, l	31	24	22

Table 10. Correspondence between Old Chinese syllable structures and late Middle Chinese tone contour and register. Initial "P" or "B" represents respectively any voiceless or voiced stop onset.

OC syllable	Contour tone of LMC	Contour × register of LMC
* P_V(N)#	平 level	阴平 yin-level
* B_V(N)#		阳平 yang-level
* P_ʔ#	上 rising	阴上 yin-rising
* B_ʔ#		阳上 yang-rising
* P_s# > P_h#	去 departing	阴去 yin-departing
* B_s# > B_h#		阳去 yang-departing
* P_p|t|k#	入 entering	阴入 yin-entering
* B_p|t|k#		阳入 yang-entering

NB: "V" stands for vowels, "N" for sonorants, and "#" for syllable boundary.

2.2.3 The causal factors of tone split

2.2.3.1 Voicing and F0

It is universally attested that voiceless and voiced consonants have different effects on the fundamental frequency (F0) of adjacent segments, in particular, the following vowel. Voiced consonants lower F0, whereas voiceless consonants raise F0 at the beginning of the following vowel. These effects have been observed in both tonal and non-tonal languages, and may result either in predictable phonetic variation or in phonological alternation. In the previous section, we already mentioned several tonal languages from East and Southeast Asia analyzed by Haudricourt (1961) which illustrate this phenomenon. In several tonal languages from Africa, consonants' voicing similarly interferes with tones, but also with tonal processes such as tone

spreading (Hyman & Schuh, 1974). In non-tonal or marginally tonal languages such as English and Swedish, acoustic analyses showed that F0 on a vowel is higher after a voiceless than voiced stop, and that this F0 difference is significant during the first 100 ms after vowel onset (Hombert, 1975).

The question remains unanswered as to whether voiceless consonants' F0-raising effect and voiced consonants' F0-lowering effect are purely physiological or deliberately controlled, universal or language-specific.

On a physiological and universal basis, two main proposals have been brought forward to account for these consonant-conditioned F0 differences, namely the "aerodynamic" hypothesis and the "laryngeal tension" hypothesis. According to the "aerodynamic" hypothesis (Ladefoged, 1967; Kohler, 1984), the increase of the oral pressure during a voiced stop's closure causes the decrease of the transglottal pressure, and as a consequence, the F0 is lowered at the release of the voiced stop. Voiceless stops, on the contrary, yield high transglottal airflow at the beginning of the following vowel and thus raise the F0 at vowel onset. However, this hypothesis cannot explain why the F0 difference lasts for almost 100 ms after vowel onset, given the fact that the high transglottal airflow is only present for a few milliseconds.

The alternative "laryngeal tension" hypothesis can be subdivided into two theories. The first theory is based on the tension of the vocal folds. It was initially proposed by Halle and Stevens (1971), and further explored by Löfqvist, Baer, McGarr and Story (1989). In Halle and Stevens' words, vocal folds are stiffer during the production of voiceless stops and slacker during the production of voiced stops, and the degree of stiffness perturbs the F0 of the adjacent vowels. The stiffness of the vocal folds is related to the activity of the (intrinsic) cricothyroid muscles. During the production of voiceless stops, the cricothyroid muscles contract to prevent voicing, thus stretching the vocal folds and accordingly raising F0 when they start vibrating. The second theory is based on the larynx height. The larynx is in a lower position for the production of voiced stops (Ewan & Krones, 1974; Ewan, 1976), presumably to enlarge the oral cavity for the purpose of maintaining the transglottal pressure necessary for voicing. F0 is thus lowered because F0 decreases when the larynx moves down, which is a well-established fact (Moeller & Fischer, 1904; Ohala, 1972; Ewan, 1976). For Maddieson (1996), the two mechanisms both take part in the control of F0:

the contraction of the intrinsic cricothyroid muscles mainly contributes to F0 raising, and the lowering of the larynx is the main determinant of F0 lowering.

More recently, Honda, Hirai, Masaki, and Shimada (1999) investigated the F0 control mechanism related to the vertical larynx movement, based on the measurement of magnetic resonance images (MRI). They found that in the high F0 range, the larynx height remains high and constant, so that the horizontal movement of the hyoid bone facilitates the tilt of the thyroid cartilage, which leads to the stretching of the vocal folds; whereas in the low F0 range, the jaw, hyoid bone, and the larynx move downward, so that the cricoid cartilage rotates along the cervical spine, which leads to the shortening and relaxation of the vocal folds (also see Hirai, Honda, Fujimoto, & Shimada, 1994). This finding is coherent with the proposal of Maddieson (1996). In our opinion, this account seems the most plausible, as illustrated in Figure 6.

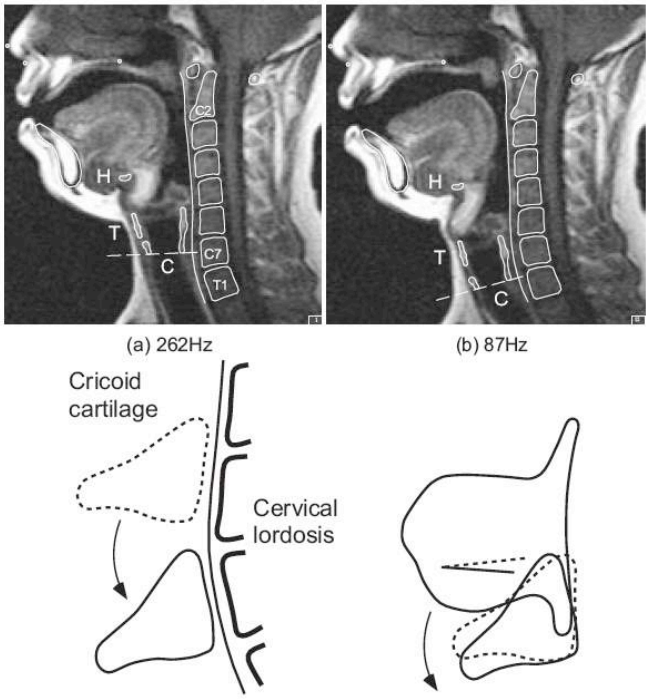


Figure 6. Vertical larynx movement and its effect on F0, from Honda (2004).

On the other hand, Kingston and Diehl (1994) and Kingston (2011) defended the idea that F0 lowering results from a *controlled* articulation rather than from universal, uncontrolled articulatory and aerodynamic mechanisms. For these authors, the above-mentioned hypotheses based on universal, automatic mechanisms suffer from fatal flaws, since counterexamples can be found. As an alternative, they

proposed that the purpose of F0 lowering is to enhance the perceptual distinctiveness of *phonological* (but not *phonetic*) voicing.

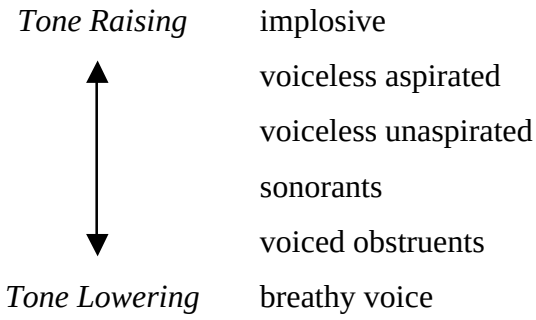
2.2.3.2 Phonation type and F0: phonation-based tone split

In §2.2.3.1, we listed both physiological and perceptual evidence supporting a cause-effect relationship between the stage of the voicing contrast and that of the *yin-yang* tone split in Middle Chinese. However, doubts have been cast on the direct causal factor of tone split. Suggestions have been made that the voiced onsets first have induced a breathy phonation, and that this breathy quality, not the segmental voicing, led to perceptible F0 lowering. Tone split would not be based on a segmental voicing contrast but on this phonation difference. This hypothesis is the "phonation-based" tone split or, using the terminology proposed by Thurgood (2002), the "laryngeally-based" as opposed to the "consonant-based" account mentioned above. (Similarly, Thurgood explained the formation of the rising tone in Vietnamese, and probably in Middle Chinese as well, with the "laryngeally-based" account).

Physiologically, this hypothesis is perfectly plausible. As we saw in the previous section, the production of voiced stops requires some larynx lowering, which slackens the vocal folds, due to the rotation of the cricoid cartilage along the cervical spine. The slackening of the vocal folds causes an increase in glottal aperture, thus producing breathy voice. In turn, because the vocal folds are not completely adducted in this configuration, the Bernouilli effect is weaker, leading to a momentary lower subglottal pressure, resulting in a lower F0. In other words, on this physiological account, the production of voiced stops entails both breathy voice and F0 lowering. Another contributing factor might be laryngeal laxness. Breathy voice involves a lesser activity of the adductor muscles, namely the lateral cricoarytenoid and the interarytenoid (Hirose & Gay, 1972), which participate in F0 regulation (Hirano, Vennard & Ohala, 1970). In fact, larger F0-lowering effects are found with breathy voiced (or aspirated voiced) stops than with plain voiced consonants, for example in non-tonal Hindi (Kagaya & Hirose, 1975) and Gujarati (Pandit, 1957: 169; Khan, 2012). Also, low tone syllables are realized with breathy voice in tonal languages such as Gurung (Glover, 1970) and Green Hmong (Andruski & Ratliff, 2000). Hyman and Schuh (1974)

proposed the phonetic hierarchy shown in (1) for consonant-induced tonal effects, according to which breathy obstruents have a larger F0 lowering effect ("tone lowering" effect in their terminology) than plain voiced obstruents.

(1)



Further investigation still needs to be done in order to factor out all the physiological factors involved in the relationship between breathiness and F0 lowering. Some researchers propose an altogether different account of this relationship. Kingston (2011) claimed, again, that breathy voice is a controlled, deliberate articulation. He proposes that breathy voice enhances the low-frequency energy in the vowel, providing a perceptual cue to the voicing of the preceding voiced stop. In other words, the purpose of both (controlled) breathy voice and (controlled) F0 lowering is to enhance the distinctiveness between voiceless and voiced stops. In any case, this account does not compromise the phonation-based hypothesis.

More compelling evidence for a "phonation-based" tone split comes from historical records of tone descriptions and analogy with evolution in other languages from East and Southeast Asia.

In 880 CE, a Japanese monk, Annen 安然, described the tones of Middle Chinese in *Shittanzō* 悉曇藏, referring to four cultural transmitters, Asian savants who brought Chinese reading to Japan, two of whom dated around early and mid eighth century, and the two others around late ninth century. In the reading of the two transmitters of late ninth century, Annen noted "四声之中，各有轻重" ('Each of the four tones has the *light* and *heavy* [allotones]'). The English translation is proposed by Mei (1970). In the reading of the two other transmitters, the "light" and "heavy" allotones were only observed for one tone (the level tone) or two tones (the level and rising tones). The four tones are of course the well established four contour tones, but the terms "light" and

"heavy" call for attention. Mei (1970) interpreted them as a description of the pitch difference between two "allotones" that would develop later into the *yin-yang* registers. Pulleyblank (1978), however, interpreted these terms differently. He reasoned that because the level tone was already described as "平声直低" ('the level tone is level and low'); the difference between "light" and "heavy" must have referred to something else, something probably related to "voice quality." The other three tones underwent other changes, so that there were other reasons for them to be qualified as "light" or "heavy." We will not go into further detail for these arguments and refer the interested reader to the work of Pulleyblank (1978, 1984). Last but not least, the traditional terms "clear" and "muddy," defined for the voicing contrast by linguists of the modern era, might themselves suggest some phonetic components related to phonation.

The evolution from a voicing contrast to a phonation difference has been first identified in Mon-Khmer languages. These languages are usually considered as non-tonal languages. In Mon, Shorto (1967) noted an ongoing disappearance of the voicing contrast and a distinction between two registers, [7] "head register" and "chest register," the first characterized by a clear voice (for voiceless onsets) and the second by a breathy voice (for voiced onsets). In Khmer, a similar distinction is found between these two registers, but only in isolated utterances, and the voicing contrast is completely neutralized (Henderson, 1952). In addition to the voice quality, the register distinction also includes a vowel quality opposition between tense and lax (tense vowels for the "head register" and lax vowels for the "chest register"), and a slight pitch difference (higher for the "head register" and lower for the "chest register"). The vowel quality difference is often considered as a primary element of the register distinction (see also, Ferlus, 1979; Wayland & Jongman, 2003). For these authors and Pulleyblank (1978) (also see Thurgood, 2002), this is a transitional period between the voicing distinction and the tonal distinction, as Middle Chinese might have experienced in the eighth and ninth centuries. It should be noted that Haudricourt

[7] The register distinction used in this section refers to the distinction of voice and/or vowel quality, possibly accompanied with pitch difference, but not to tone distinction, as used elsewhere in this work.

(1965) has also given a brief overview of the register distinction in Mon-Khmer languages, based on data collected by Michel Ferlus and Henri Maspero, just several years after the publication of his consonant-based tone split theory. His interpretation was somewhat different that of the above-mentioned authors. Instead of considering the distinction as a transitional stage before the tone split, he believed that the voicing contrast evolved differently in tonal and non-tonal languages. In tonal languages such as Middle Chinese, the loss of the voicing contrast gives rise to a pitch contrast and doubles (or triples in some case) the number of tones. In non-tonal languages, the originally voiced onset "relâche" ('relax') the following vowel and result in a lax vowel with a lax voice. Note that in both cases, the evolution is conditioned by the consonant onsets.

Furthermore, register development is frequently found in languages of Southeast Asia (and beyond) (Denning, 1989). The Chamic languages, a group of Austronesian languages, have undergone "register split" from an original voicing contrast, characterized by vowel quality and pitch difference in Western Cham (Edmondson & Gregerson, 1993) and phonation and pitch difference in Eastern Cham (Phu, Edmondson & Gregerson, 1992; Brunelle, 2005).

The evolution from voicing to tone contrast through a phonation difference is somehow illustrated in Tamang-Gurung-Thakali-Manangke, a group of Tibeto-Burman languages of Nepal (Mazaudon, 2012; Mazaudon & Michaud, 2008). In conservative dialects including Risiangku Tamang, Sahugaon Tamang, Tukche Thakali, Syang Thakali and Ghachok Gurung, low pitch syllables are accompanied with occasional voicing of the onset and by breathy voice on the rime. On the other hand, in Manang, a more evolved dialect, pitch is distinctive but breathiness and voicing have both disappeared. The more conservative dialects are believed to reflect an intermediate stage with breathiness on lower pitch syllables as a redundant feature, in between an ancient onset's voicing contrast and a pitch difference evolving toward a tone contrast. In addition, initial voicing can also be present as a trace of the "mother-feature."

The same development is also found in Punjabi, an Indo-Aryan language, in which the reconstructed Indo-Aryan syllable-initial *bɦ gave rise to a low tone and consequently de-voiced and de-aspirated (Bahl, 1957; Chatterji, 1940: 113-114;

Haudricourt, 1972). In the same vein, we are perhaps witnessing a laryngeally-based emergence of tonal contrast in contemporary Korean according to some acoustic evidence (Silva, 2006): lax stops trigger a (default) low tone, and aspirated and tense stops trigger a high tone.

Concerning Chinese dialects, in a recent study (Zhu, 2010), the presence of breathy voice has been shown acoustically (although without statistical analyses) in many Central and Southern Chinese dialects, such as Wu, Xiang, Gan and Yue.

3 MULTIDIMENSIONAL CUES OF TONE/VOICING CONTRAST

Summary

In this chapter, we define a toneme with its multidimensional cues. We further investigate the phonetic cues of the tone register and the voicing contrast.

First of all, instead of interpreting the voiceless vs. voiced realization as the surface form of the tone register contrast in word-medial position, we consider both the voicing and the tone register contrast as phonological. They are conditioned by the context. More precisely, the voicing contrast is complementary to the *yin–yang* register contrast according to the position in the word (§3.1): in word-initial context, the *yin–yang* contrast is tonal between the two registers, and the voicing contrast is neutralized; in word-medial context, the tone register contrast is neutralized by virtue of a tone sandhi rule, and the voicing contrast is maintained.

In §3.2, we review the phonation types related to tones in Shanghai Chinese described in the literature. We try to determine the phonetic elements reflected by the traditional description of voiced stops 清音浊流 'clear sound followed by muddy airflow' (Chao, 1928). "Muddy airflow" is interpreted as a breathy phonation in recent studies. We give terminological precision as well as articulatory and acoustic definitions of this phonation type (§3.2.2), summarize the debate about the time course of breathiness in Shanghai Chinese (§3.2.3), and lastly provide an overview of paralinguistic and linguistic use of different phonation types (§3.2.4). We compare the pitch-independent use with the pitch-dependent use of phonation types.

In §3.3, we present the results from the literature on the duration pattern related to tones in Shanghai Chinese. Phonologically voiced consonants are shorter than their voiceless counterparts, even when they are phonetically voiceless. Rimes following phonologically voiced consonants are longer than following voiceless consonants in Shanghai Chinese, which is rarely found in other languages. Several explanations are proposed.

Voicing, F0, phonation type, and duration are all of crucial importance and will be investigated in our experimental part. In §3.4, we present our research goals and related questions. We mainly aim at (1) giving a detailed description of articulatory and acoustic correlates in production and perception; (2) identifying different types of redundant features (due to coarticulatory effects or remnants of diachronic evolution?); (3) studying how these features evolve in the rapidly changing Shanghai Chinese.

In the previous chapter, we explained some aspects of the interdependence between tones and glottal settings in Shanghai Chinese. Some of them strictly correlate with tones and are traditionally integrated in the descriptions of tones. For example, phonological voicedness is attached to the *yang* register, and glottal stop codas to the checked tones. In addition to these phonologized and redundant features, secondary articulatory and acoustic cues are also important in the definition of each tone. Rose (1982a) argued that, in the study of a tone language, a "polydimensional" approach composed of different phonetic parameters should be used, in order to account for the complex physical reality of the tone system. In the same vein, Mazaudon (2012) proposed to "conceptualize a toneme as a bundle of cues, some of them non-pitch features, rather than as defined by a single distinctive feature accompanied by 'redundant' features'." (Mazaudon, 2012: 140). In particular, languages with large tonal inventories need other cues than pitch, such as phonation and duration, to maximize tonal contrasts (Brunelle, 2009; Kuang, 2013a).

The first question we might ask is about the phonological status of the tone register contrast (*yin* vs. *yang*) and of the voicing contrast. We may consider legitimately that the voicing contrast is not phonological and that obstruents' phonetic voicing is nothing more than the surface form of the low tone in word-medial position. The other way round, that is, a phonological voicing contrast with tones as surface forms, is not legitimate because, for one thing, voiceless onsets would surface as two different *yin* tones, and for the other, nasal onsets can carry either *yin* or *yang* tones. We prefer to give a phonological status to both the tone contrast *and* the voicing contrast. The tone contrast and the segmental contrast both exist on simple minimal pair tests: [tɛ˧] 'gallbladder' ≠ [tɛ˩] 'egg', [pi˩ tɛ˧] 'sheet' ≠ [pi˩ dɛ˧] 'preserved egg'.

The phonetic realizations of the two contrasts are presented in the next section. In the following sections, we review some other correlated phonetic cues of the *yin-yang* tone/voicing contrast, as found in the literature.

3.1 Context-conditioned contrast

In Shanghai Chinese, as mentioned in the preceding chapter, the voicing contrast and the tone contrast are conditioned by the context. More precisely, the voicing contrast is complementary to the tone register contrast according to syllable context. Both contrasts are archiphonemic (or architonemic).

As early as in 1925, LIU Fu completed his PhD at the Sorbonne University, Paris. He studied the tones of three Chinese dialects, among which the Jiangyin 江阴 dialect, a Northern Wu dialect spoken in the south of the Jiangsu province. He demonstrated, using kymographic data, that, in this dialect, voiced obstruents ("sons obscurs" 'obscure sounds', in his words) were phonetically voiceless ("consonnes muettes" 'silent consonants' in his formulation) in isolated or word-initial syllables, and became voiced when preceded by another sound. An example is given in Figure 7.

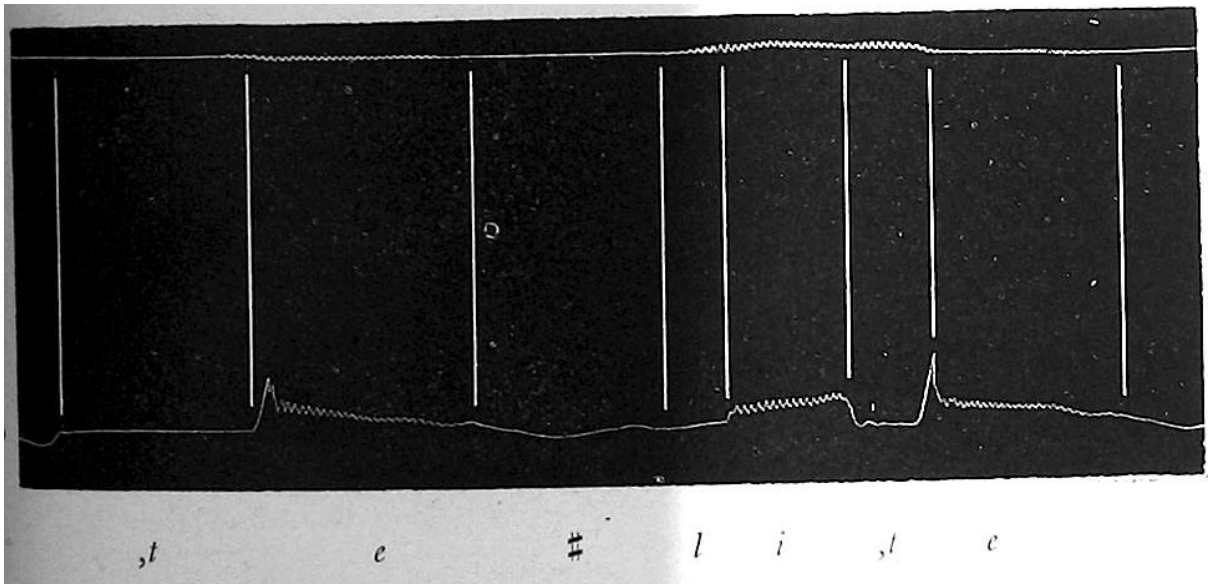


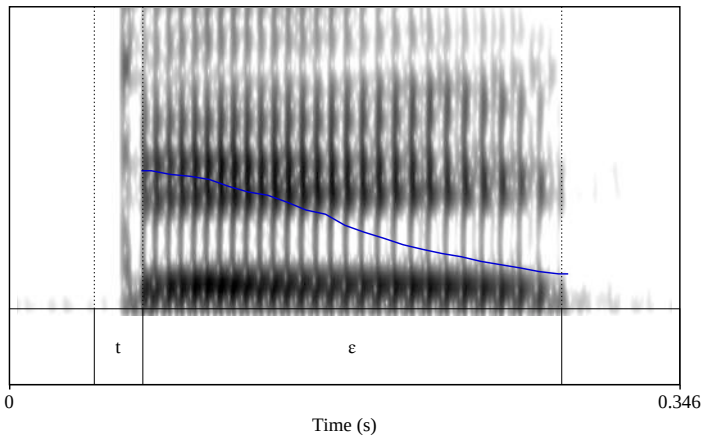
Figure 7. Laryngeal vibrations (top) and oral vibrations (bottom) for the syllable /dɛ/, with a phonetically voiceless onset in isolation (left: no oscillations), vs. a phonetically voiced onset in the word /lidɛ/ (right: oscillations). From Liu (1925: 60ff).

During the last century, the phonetic details of obstruent voicing in Wu dialects have often been discussed. Among other linguists, Chao (1928) agreed with Liu's findings, whereas Karlgren (1915-1926) presumably supported a different opinion, as suggested by his transcription of "voiced" sounds with voiced symbols.

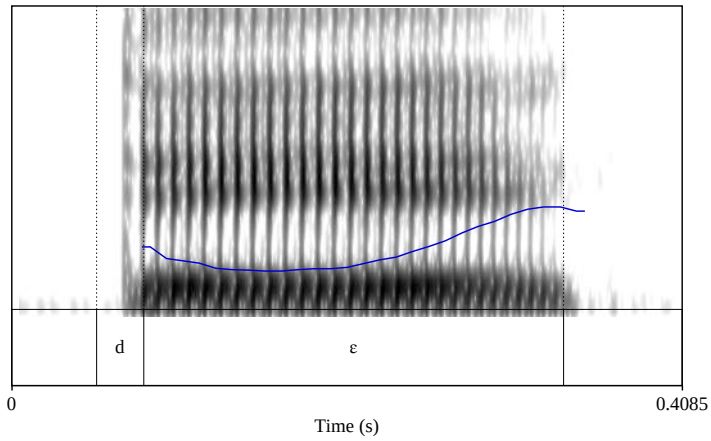
Liu Fu's kymographic method was certainly time-consuming and meticulous. Decades later, using simple VOT (voice onset time) measurements, the phonetic realizations of the voiced series in Wu dialects can be clearly assessed. Shanghai Chinese is found to share the same pattern with Jiangyin Chinese.

In word-initial or stressed position, the voiced stops are phonetically voiceless, as shown by their positive VOTs (Cao & Maddieson, 1992; Ren, 1987; Gao, Hallé, Honda, Maeda, & Toda, 2011, etc.). The original voicing contrast is neutralized but the tone register contrast between *yin* and *yang* is maintained. The phonologically voiced fricatives are also presumed to be voiceless in this position, or at least voiceless during the first half (Chao, 1928). Acoustic measures of voicing in fricatives are quite rare. However, Gao and Hallé (2012), using a voicing ratio measurement, found that phonologically voiced fricatives are often realized as phonetically voiced.

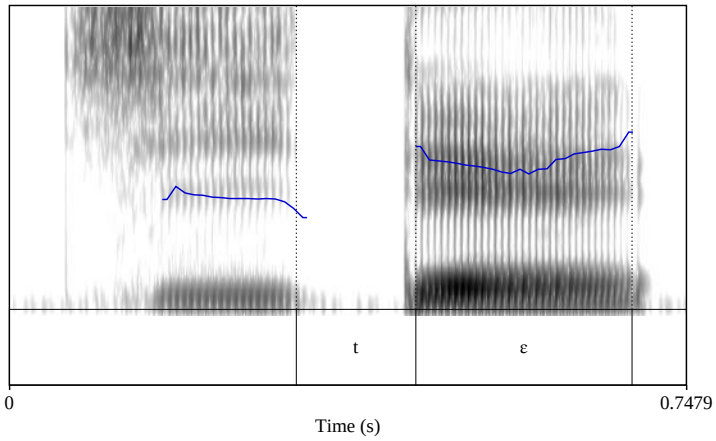
In word-medial unstressed position, the tone contrast is neutralized due to the tone sandhi process (see §2.1.3.2), while the voicing contrast is maintained. The VOT values of voiced stops are negative in this position (Cao & Maddieson, 1992; Ren, 1987); the voicing ratio values for voiced fricatives are close to 100%, indicating complete phonetic voicing (Gao & Hallé, 2012). The spectrograms in Figure 8 illustrate the presence of the tone contrast and neutralization of the voicing contrast in initial, stressed position, and the neutralization of the tone contrast and presence of the voicing contrast in non-initial, unstressed position.



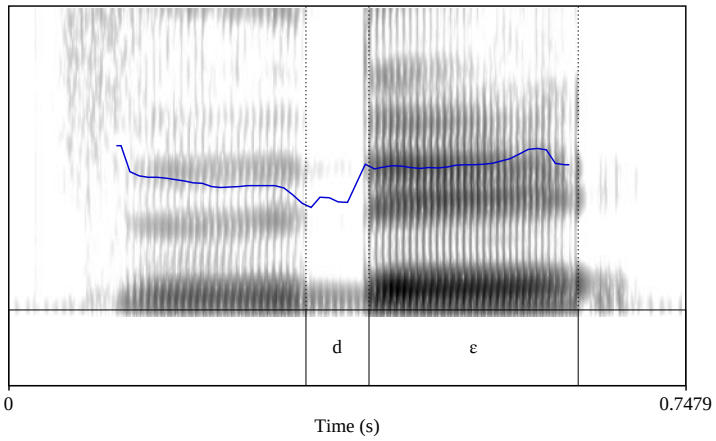
(a)



(b)



(c)



(d)

Figure 8. Spectrograms illustrating (a) voiceless /t/ in word-initial position; (b) phonetically voiceless /d/ in word-initial position; (c) voiceless /t/ in word-medial position and (d) phonetically voiced /d/ in word-medial position. The second syllables in (c) and (d) have similar tone contours (speaker: 25-year-old male). (CD: 3.1_fig9)

3.2 Phonation type

In this work, we use the terms "phonation type" and "voice quality" interchangeably to describe laryngeal settings regardless of supralaryngeal settings.

3.2.1 Impressionistic descriptions: clear sound with muddy airflow

Around 100 years ago, Karlgren (1915-1926: 260) suggested that voiced onsets in Wu dialects were phonetically voiced, but also accompanied by some "voiced aspiration." He described them as follows:

"Les **b** des dialectes Wou, comme les autres occlusives sonores de ces dialectes – explosives tant qu'affriquées – sont accompagnés, à la détente, d'une aspiration sonore. En réalité, celle-ci est tout à fait identique au phonème initial du sanscrit *bharati*. Cependant l'aspiration des dialectes Wou est, à mon avis, trop faible pour mériter d'être désignée."

["The **b** of Wu dialects, like all the other voiced stops of these dialects – plosives as well as affricates – are accompanied by a voiced aspiration at the release. In fact, this aspiration is the same as the initial phoneme of *bharati* in Sanskrit. However, in my opinion, the aspiration in Wu dialects is too weak to worth its name." (my translation).]

This observation is echoed in later descriptions.

As a native speaker of Jiangyin dialect (Northern Wu), Liu (1925), who studied experimentally voiced onsets, as mentioned in §3.1, made the following observation (in 1923):

"我自己所發的濁音，在句首的和在句中的不同。例如'朋友'之'朋'是[pɦ], '一個朋友'之'朋'是[b]。因此圖中有多數濁音，都有兩種音標；如羣為 [kɦ] 與 [g]，定為 [tɦ]與 [d] 之類。究竟一般人所發的濁音是否如此，尚待研究。"

["The production of voiced sounds by myself depends on whether they are in sentence-initial or sentence-medial[8] position. For instance, the syllable '朋' from '朋友 <friend>' is produced with [pɦ], but the same syllable '朋' from '一個朋友 <a friend>' is produced with [b]. Therefore, in the following consonant table, I choose two symbols for these sounds; e.g. [kɦ] and [g] for the syllable '羣', and [tɦ] and [d] for the syllable '定'. However, it still needs to be investigated whether these sounds are really produced this way by naïve speakers." (my translation).]

In the first modern dialectological study on Wu dialects, Chao (1928) agreed with Liu's (1923) transcription of the Wu "voiced" sounds with a voiced *h* [ɦ], and described voiced stops as follows, "they begin with a quite voiceless sound and only finish with a voiced glide, usually quite aspirated, in the form of a voiced *h*." (Chao, 1928: xii) As for voiced fricatives and affricates, he wrote: "the second half may be voiced," and "z" can be transcribed [sɦ] or [sz], for example. He used the term 清音浊流 'clear sound followed by muddy airflow' to describe these sounds, a formulation widely accepted by Chinese linguists. As a linguist highly sensitive to phonetic differences, he judged by ear that most of the Wu-dialect speakers produced voiced consonants this way. In the next section, we will try to define this "muddy airflow."

3.2.2 Breathy voice: definition and terminology

The sounds described by the linguists cited above are indeed quite close to the voiced aspirated stops that exist commonly in Indo-Aryan languages such as Hindi, although they are actually voiceless. Yet, Karlgren (1915-1926: 260) suggested that the aspiration was much weaker in Wu dialects than in Hindi.

The label "voiced aspirated" is commonly used for Hindi, which is described as using four stop series: voiceless aspirated, voiceless unaspirated, voiced unaspirated, and voice aspirated. This label is not appropriate on a phonetic basis since these stops

[8] The following examples illustrate phrase-initial vs. medial rather than sentence-initial vs. medial contexts.

are "neither voiced (in the sense of having regular vibrations of the vocal cords) nor aspirated (in the sense of having a period of voicelessness during and after the release of the closure). … and the release is neither voiced nor voiceless but murmured," according to Ladefoged (1975: 126-127). For him, the terms "breathy voice" and "murmur" are preferable. And the breathy voice can be a consonant's as well as a vowel's property. Many authors including Ramsey (1989: 91), Sherard (1972), Cao & Maddieson (1992) and Zhu (1999) all used one of these two terms to describe Shanghai "voiced" obstruents.

Ladefoged (1971) and Gordon and Ladefoged (2001) proposed a simplified continuum of phonation type based on the glottal aperture, shown in Figure 9. The breathy voice, used in a broader sense, ranges from the left end of the continuum, where the glottis is most open as in the production of a voiceless sound, to the midpoint, where the glottis is moderately closed as in "normal," modal voice.

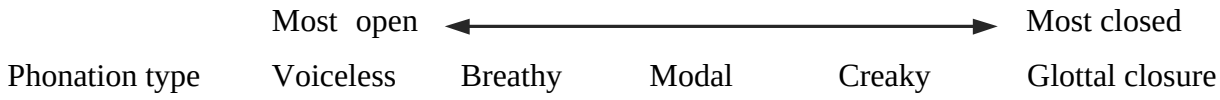


Figure 9. Continuum of phonation types after Ladefoged (1971), from Gordon & Ladefoged (2001).

We may summarize the articulatory settings for breathy voice as follows: (1) slackened vocal folds with minimal muscular tension (adductive tension, medial compression and longitudinal tension) as the vocal folds vibrate (Laver, 1980: 133); (2) larger glottal aperture, and therefore (3) higher rate of airflow through the glottis than in modal voice; (4) open quotient above 50%, that is, longer open than closed glottis phase during a glottal cycle. In addition, the production of voiced aspirated or breathy stops, as found in particular in Hindi, also requires a special timing relationship of the glottal aperture relative to suprasegmental articulation (Kagaya & Hirose, 1975; Benguerel & Bhatia, 1980). For voiced aspirated stops, the maximal glottal aperture is located a few centiseconds after the release of the closure; for voiceless aspirated stops, it is located around the release; for voiceless unaspirated stops, it occurs well before the release; finally, for voiced unaspirated stops, the glottis remains closed during the entire stop closure.

In a narrower sense, breathy voice is only one of the subcategories in the range between voiceless and modal. The division into subcategories and labels is, however, not entirely consensual.

If most authors use "breathy voice" and "murmur" interchangeably, Catford (1977: 96-101) made a distinction between "breathy voice" and "whispery voice (or "murmur")", the latter often "incorrectly" called "breathy voice" in the literature according to him. He described the whispery voice as produced with a narrowing of the glottis, and generating "richly turbulent escape of air". The glottis is not narrowed in the production of the breathy voice. Laver (1975: 134) also distinguished "breathy voice" (compound phonation of breathiness and voice) from "whispery voice" (compound phonation of whisper and voice) on the auditory basis, according to which whispery voice involves a more audible friction component compared to modal component than breathy voice. In other words, whispery voice is noisier than breathy voice. Besides, whispery voice is produced with a higher laryngeal effort and a stronger medial compression.

Whispery voice is described as a phonation type in a northern Wu dialect, Zhenhai 镇海 dialect, spoken in the Zhenhai county of the Northeastern Zhejiang province. Rose (1989) found that Zhenhai *yang* vowels are produced with whispery voice, and *yang* obstruents and devoiced vowels with whisper (that is, without glottal pulsing).

Ladefoged and Maddieson (1996: 57) distinguished "breathy voice" from "slack voice" on a physiological basis. The degree of glottal aperture, the airflow rate and of the slackness of the vocal folds are all higher in slack voice than in modal voice, but are even higher in breathy voice than in slack voice. Thus the breathy voice is located to the left of the slack voice in the continuum of phonation types. Slack voice is reported in Javanese (Fagan, 1988) and Xhosa (Jessen & Roux, 2002). Ladefoged and Maddieson (1996: 64-65) preferred the term "slack voice" than "breathy voice" to describe Shanghai "voiced" stops.

The phonation continuum, useful as it may be for phonation classification, "fails, however, to take explicit account of location differences." (Catford, 1977: 105). Multiple articulatory activities are indeed involved in the production of different phonation types, as demonstrated by Esling and Harris (2003) and Edmondson and

Esling (2006) with larynscopic observations. Instead of the classic view single valve system, on which the one-dimension phonation continuum is based, they propose a more complex system consisting of six vocal tract valves: Valve 1: glottal vocal fold abduction/adduction; Valve 2: ventricular folds incursion; Valve 3: forward and upward sphincteric compression of the arythenoids and of the aryepiglottic folds; Valve 4: epiglottal-pharyngeal constriction; Valve 5: laryngeal raising/lowering by the suprahyoid muscle group; and Valve 6: pharynx narrowing. The authors described canonical breathy voice as a phonation type with a partial closure of Valve 1 (vocal folds), with oscillation at the anterior part of the vocal folds and aperture at the posterior part of the vocal folds.

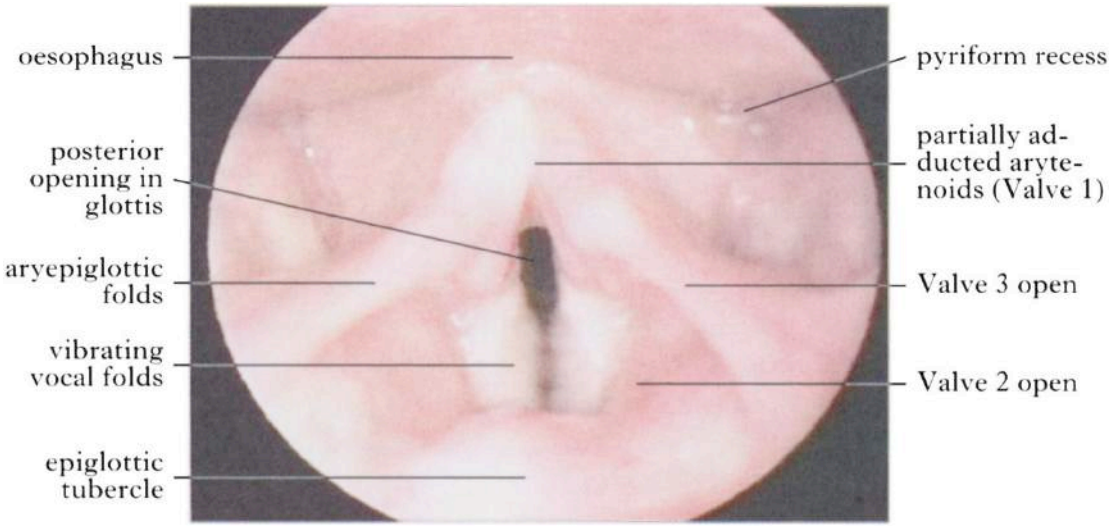


Figure 10. Laryngoscopic image of canonical breathy voice, from Edmondson & Esling (2006).

Acoustically, breathy voice is mainly characterized by (1) the dominance of the lower harmonics, notably the fundamental (i.e., H1), thus leading to a steep negative spectral tilt (Fischer-Jørgensen, 1967; Bickley, 1982; Klatt & Klatt, 1990; Hillenbrand, Cleveland & Erickson, 1994), (2) a decrease in the amplitude of harmonics in higher frequencies (e.g., Fischer-Jørgensen, 1967), (3) the presence of random aspiration noise in the spectrum, particularly at high frequencies (Ladefoged & Maddieson, 1996: 65; Ladefoged & Antañanzas-Barroso, 1985) inducing less regular periodicity (Klatt & Klatt, 1990; Hillenbrand et al., 1994), and (4) perhaps not much investigated and less systematically, the lowering of the first formant (Thongkum, 1988, but see Maddieson & Ladefoged, 1985, for crosslinguistic variations). In Hindi,

aspiration noise is clearly visible in spectrograms, at vowel onset after stop closure release. In Shanghai Chinese, however, there is little or no noise at vowel onset. From the spectrograms in Figure 11, we can notice clear aspiration noise at the release of the Hindi breathy stop, but the Shanghai "breathy" stop is only characterized by a decrease in energy at vowel onset. This weaker noise corresponds to the auditory impression reported by Karlgren (1915-1926: 260) and somewhat justifies the term "slack voice" for Shanghai Chinese. However, the spectral tilt (indicated by H1–A2, not H1–H2 in this case) is found to be larger for Shanghai "voiced" than voiceless stops, as shown in Figure 12.

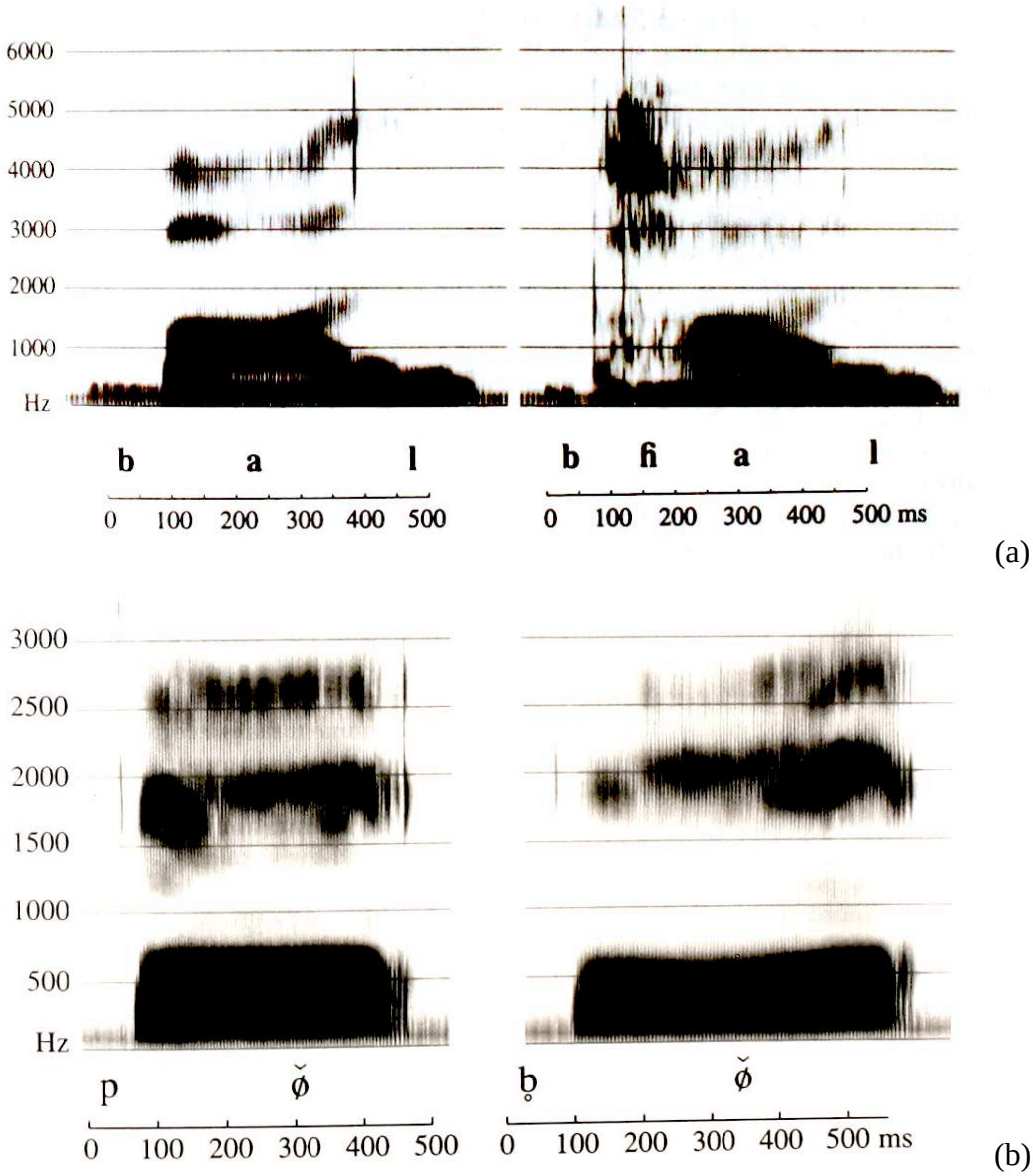


Figure 11. Spectrograms illustrating modal (left) and breathy (right) stops in (a) Hindi and (b) Shanghai modal-breathy minimal pairs. From Ladefoged & Maddieson (1996: 59-65).

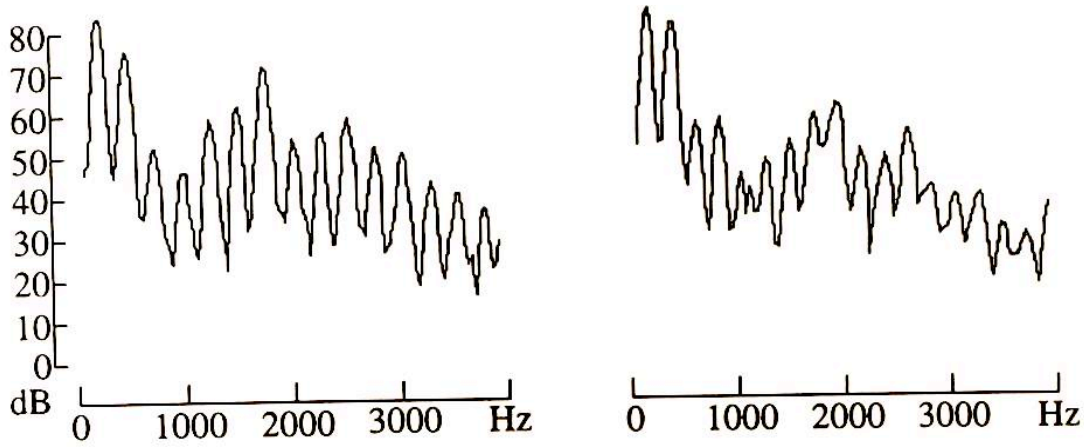


Figure 12. Short term spectra illustrating modal (left) and breathy (right) stops in the Shanghai minimal pair /p-b/, from Ladefoged & Maddieson (1996: 65).

It is true that the Shanghai "voiced" obstruents are produced with breathy voice to a lesser degree than Hindi voiced aspirated stops. But we still employ the term "breathy voice" rather than "slack voice" in this work to describe them. First, we tend to use "breathy voice" in a broader sense in order to make generalizations on this feature, which is common to (Northern) Wu dialects. Second, one purpose of this work is to give an estimation of the degree of breathy voice in Shanghai Chinese.

3.2.3 The domain of breathy voice

Whereas the breathy voice is clearly a property of the consonant in some languages including Hindi, Marathi, Telugu, Newari, Mundari, etc. (Ladefoged & Maddieson, 1996: 57), in other languages, such as Gujarati, Jalapa Mazatec, breathy vs. modal phonation is instead contrastive in the vowel rather than in the consonant: that is, breathiness is maintained throughout the entire vowel (Gordon & Ladefoged, 2001).

What about Shanghai Chinese? There are three main proposals for the domain on which breathiness distinguishes yang syllables: the syllable onset, the entire syllable, and the phonological word. On a fourth proposal, breathy voice is treated as a property of the syllable nucleus. In our view, this latter view can hardly be defended.

The impressionistic descriptions mentioned in §3.2.1 and the label "clear sound with muddy airflow" suggest that breathy voice should be interpreted as a property of

the onset. From a diachronic point of view, breathiness, as explained in §2.2.3.2, is a property which developed from the ancient voicing of the obstruent onset. Phonetic evidence seems to support this proposal. Cao and Maddieson (1992), as well as Ren (1992), found that only the consonant's release and the following vowel onset are affected by breathiness, and interpreted the breathy property as part of the consonant that extended to the vowel onset due to coarticulation.

On the other hand, however, Ramsey (1989: 91) stated that "the breathiness (in Shanghai Chinese) is acoustically quite prominent; it pervades the entire syllable, beginning in the initial consonant and lasting throughout the syllable vowel." But he did not provide any phonetic data. Chao (1934) and Sherard (1972: 87) reported the same observation. Phonological arguments for the syllable as the domain of breathiness are perhaps more relevant than phonetic ones. Rose (1989, 2002) and Yip (1993) pointed out that zero onset and sonorant onset syllables are also concerned with the phonation difference, suggesting that phonation type is not determined by the [+voice] feature of the syllable onset. (But Duanmu [1988, cited in Yip, 1993] has suggested that zero-onset is the surface form of an underlying obligatory onset slot. Accordingly, phonation type is the property of this onset slot.) Yip (1993) further argued that phonation type is a property of the entire syllable because, as explained in §2.1.3.2, in polysyllabic words, both tone and breathy voice are deleted in non-first syllables, whereas voicing is not. According to Yip, breathy voice is a feature (the [+murmur] feature) assigned to the tonal root node in a feature geometry model (Clements, 1985): like the tone feature(s), it modifies the entire syllable in monosyllables, and it is deleted in par with the tonal root node's deletion of non-first syllables in polysyllables.

Gao (2011), Gao & Hallé (2012) also showed the same pattern in monosyllabic words. Contrary to Ren's (1992) results that showed no breathy voice at vowel midpoint, they found H1–H2 difference between *yin* and *yang* syllables at the vowel midpoint, although the difference was less remarkable than at the vowel onset.

In this study, more time points will be used in order to track more precisely the time course, and thus the phonetic domain of the breathy voice.

3.2.4 Different uses of phonation types

3.2.4.1 Paralinguistic use of phonation types

Paralinguistic use of phonation types is widespread in different languages, tonal and non-tonal. For example, in Tzeltal (Mayan language of Mexico), sustained falsetto is used as an honorific feature, and creaky voice as an expression of commiseration and complaint (Brown & Levinson [1978: 272, cited in Laver, 1980]). In many languages, breathy voice is used for intimate communication (Laver, 1980). In British English, breathy voice furthermore indexes feminine and "desirable" qualities (Henton & Bladon, 1985), whereas creaky voice rather is a sociophonetic marker of masculine and authoritative voices (Henton, 1986). Indeed, female voices are on average breathier than male voices (Henton & Bladon, 1985; Klatt & Klatt, 1990; Hanson & Chuang, 1999). This said, today's young female speakers of American English frequently use creaky voice with a completely different social indexation. Their creaky voice is perceived as indexing "educated, urban oriented and upwardly mobile" women (Yuasa, 2010, among others).

3.2.4.2 Linguistic use of phonation types and relation between pitch and phonation

Phonation difference can be contrastive in some languages, as in non-tonal Mon-Khmer languages (see §2.2.3.2).

In some tonal languages, the contrastive use of phonation types is orthogonal to that of tones. Mpi (Tibeto-Burman) contrasts two phonations and six tones independently (Silverman, 1997; Blankenship, 2002). Jalapa Mazatec contrasts three phonation types and three tones independently (Garellek & Keating, 2011). Southern Yi, Bo and Luchun Hani (Tibeto-Burman Yi languages) distinguish two phonation types on mid and low tones (Kuang, 2013b). In these languages, phonation categories constitute another perceptual dimension than tone categories. Kuang (2012, 2013b) called this use of phonation types as a "pitch-independent use." Pitch-independent

phonation types are not related to pitch but add the phonation dimension to the dimension of tonal contrasts.

There exists, of course, a "pitch-dependent use" of phonation types. According to Kuang (2012, 2013b), pitch-dependent phonation types are often produced with super high and super low tones, so as to enhance the perceptual differences for these two types of tones. The interdependence between fundamental frequency (F0) and phonation has been abundantly described for singing voice (e.g., Sundberg, 1987; Henrich, 2001). It is also found in languages world-widely. Non-modal phonation types, both creaky and breathy, are often associated with low F0.

The association between laryngealized voice and low F0 is found in many languages, synchronically or diachronically, for example in Mam, Northern Iroquoian languages (see Gordon and Ladefoged, 2001, for a review), in Mandarin Chinese (Davison, 1991; Belotel-Grenié & Grenié, 1994, 1995), and in Cantonese (Yu & Lam, 2014). Gordon and Ladefoged (2001), however, suggested that this correlation is not universal, since F0 can also be raised by laryngealization, especially in the case of final glottal stop. Kingston (2011) explained that if laryngealization only involves the contraction of the thyroarytenoid muscle, F0 is lowered and adjacent vowels become creaky, while if laryngealization also involves contraction of the cricothyroid muscle, F0 is raised and adjacent vowels become tense. He found evidence for these patterns in Athabaskan languages (Western North America), in which the stem-final glottic consonants have developed opposite tones: high tones in Chipewyan but low tones in Gwich'in (Kingston, 2005). Another good illustration of these two-way laryngealization-induced F0 contrasts is the tonal development of the glottal stop coda, which gave rise to a rising tone from Old Chinese to Middle Chinese (see §2.2.2), and is now developing into a falling tone in Modern Tibetan (Mazaudon, 1977). In sum, there are two kinds of laryngealization: a tenser articulation raises F0, and a laxer articulation (commonly known as "vocal fry") lowers F0.

If breathy voice affects F0 in a language, it tends to lower F0. §2.2.3.2 provides an overview of the synchronic associations between breathy voice and F0 lowering, as well as the diachronic evolutions from breathy voice to low tones that languages have undergone or are undergoing. Languages that exhibit the opposite correlation, that is,

breathy voice associated with F0 raising, are rare but do exist, for example Chanthaburi Khmer (Wayland & Jongman, 2003).

When phonation types are pitch-dependent, they can be phonemic or allophonic, although it is not always easy to determine whether tone conditions phonation type or vice versa. Even in those languages that clearly have contrastive tonal categories, phonation types and other non-tonal properties can function as primary perceptual cues. In Northern Vietnamese, laryngealization is used as a primary cue to identify the rising contour and the falling-rising contour (Brunelle, 2009). In Sgaw Karen (Karenic, Tibeto-Burman), phonation type and duration are crucial in tone perception (Brunelle & Finkeldey, 2011).

Finally, some languages in which neither tone nor phonation type has a linguistic function exhibit a correlation between F0 and breathiness (as measured by H1–H2 or the open quotient (OQ) in voicing periods), but not always in the same direction. Holmberg, Hillman and Perkell (1989) already observed an increase in OQ with the increase in F0 in their correlation between several glottal airflow measurements. Weak positive correlation between OQ and F0 has been found in Dutch (Koreman, [1996, cited in Iseli, Shue, Epstein, Keating, Kreiman & Alwan, 2006]), while negative correlation between corrected H1–H2 has been found in American English (Iseli et al., 2006).

3.3 Duration

In his PhD dissertation, Liu (1925: 62) already observed that "*La consonne initiale d'un son clair est, pour la plus part des cas, plus longue que celle d'un son obscur correspondant*." [Initial clear consonants are in most cases longer than initial obscure ones.] (my translation; "clair" and "obscur" refer to voiceless and voiced, respectively; *qing* 清 "clair" also is the Chinese term used for "voiceless").

The duration aspect has been less investigated in Shanghai Chinese. Shen, Wooters and Wang (1987) found that voiceless stops onsets are longer than their voiced counterparts in word-medial position, where phonological voiced stops also are phonetically voiced, as well as in word-initial position, where all stops are phonetically voiceless. A methodological problem remains, however. In their study, the stop onset

target syllable was elicited in the frame sentence [t̥ʊʔ _ ɪʔ pi] 'Read __ once' and the beginning of word-initial stop closure was defined as the end of the preceding syllable. Whether the closure portion so defined contained a silent pause or not can of course not be determined from the inspection of the acoustic signal.

Gao (2011), and Gao and Hallé (2012) investigated duration differences between voiceless and voiced fricative onsets. They found that phonologically voiceless fricatives are consistently longer than their phonologically voiced counterparts, in word-medial as well as word-initial position, whether phonetically voiced or not.

It is universally attested that voiced obstruents are shorter in duration than their voiceless counterparts (e.g., Umeda, 1977). This is accounted for by what Ohala (1997) called the "aerodynamic voicing constraint." During the closure of a voiced stop, the oral air pressure P_{oral} increases as air accumulates in the oral cavity, gradually becoming close to the subglottal pressure $P_{sub\text{-}glot}$ so that the transglottal pressure differential ($\Delta P_{glot} = P_{sub\text{-}glot} - P_{oral}$) decreases and becomes insufficient to maintain the airflow necessary for glottal pulsing. As a consequence, in the absence of vocal tract expansion, voicing cannot be maintained and closure duration is limited. To produce voiced fricatives, it is necessary to maximize P_{oral} to produce friction which is, again, incompatible with maintaining the transglottal pressure differential ΔP_{glot} for a long time. As a consequence, a compromise has to be reached between glottal pulsing and supralaryngeal friction and voiced friction cannot last long. In Shanghai Chinese, phonological voiced obstruents are phonetically voiceless in word-initial position but still, they are shorter than phonological voiceless obstruents. (Note, however, that voiced realizations also exist, see §4.1.6.2.) Gao (2011), and Gao and Hallé (2012) interpreted this difference as remnants of a presumed phonetic voicedness, which preceded the diachronic change from a voicing contrast to a tone register contrast.

Vowel duration varies with the voicing of the following consonant: vowels followed by a voiced consonant are generally longer than those followed by a voiceless consonant and this trend is widely found cross-linguistically (for English: Lisker, 1957; Peterson & Lehiste, 1960; for French: O'Shaugnessy, 1981; Abdelli-Beruh, 2004). In particular, this pattern is also found in Shanghai Chinese (Zhu, 1999: 191). In contrast, vowels do not differ in duration according to the voicing of the preceding consonant in English (Peterson & Lehiste, 1960). In Shanghai Chinese, however,

voiceless consonants are followed by short vowels and voiced consonants by long vowels (Gao & Hallé, 2012). Gao and Hallé (2012) attributed the vowel's duration difference to duration compensation at the syllable level. A similar trend is found in Korean: fortis stops, which are longer than lenis ones, are followed by shorter vowels (Chung, 2002).

An alternative explanation is that vowel (or tone) duration partly depends on tone height, range, and contour. Based on a study of Thai vowel duration, Gandour (1977) proposed a universal phonetic tendency for vowel duration to depend, at least partly, on tonal characteristics: vowels are longer when produced with a rising tone than with a falling tone, as also found in Mandarin Chinese (Howie, 1976); they are also longer when produced with a low tone than with a high tone. The F0 range covered in the tone contour also seems to influence vowel duration. Based on an acoustic study with F0 patterns produced deliberately by five American speakers, Hombert (1977) found that for rising F0 contours, vowel duration is positively correlated with the covered F0 range. He also found that low F0 contours are shorter than high F0 contours. Yet, in Shanghai Chinese, among checked tones, the lower tone T5 was found longer than the upper tone T4 (Zee & Maddieson, 1979; Rose, 1993). Among the Shanghai Chinese unchecked tones, the falling tone T1 (52) was found to be the shortest (Zee & Maddieson, 1979; Rose, 1993). This might partly explain why Gao and Hallé (2012) found that the same vowel [ɛ] is shorter with *yin* tones T1 and T2 than with *yang* tone T3. Rose (1993), on the other hand, did not find systematic differences between the two rising tones T2 and T3.

A third explanation is that vowel duration partly depends on phonation type. In many languages, breathy vowels are longer than modal vowels (Fischer-Jørgensen, 1967, for Gujarati; Samely, 1991, for Kedang; Kirk, Ladefoged & Ladefoged, 1993, and also Silverman, 1995, for Jalapa Mazatec). But this pattern is far from being universally attested. The opposite pattern has indeed been found in San Lucas Quiaviní Zapotec, for example (Gordon & Ladefoged, 2001).

3.4 Research goals

In this dissertation, we attempt to give a detailed description of the articulatory and acoustic correlates of Shanghai tones. We try to identify the most robust and important correlates in production as well as in perception.

In our attempt at describing Shanghai tones, we follow Mazaudon (2012), who conceptualizes a toneme as a bundle of pitch and non-pitch cues. She also distinguishes the redundant features that are due to coarticulatory effects from those who were distinctive in a previous state of the language and survived a diachronic transphonologization process. The former redundant features are universal and predictable, and the latter ones are presumably language-specific and may explain inter- and intra- speaker variations. However, universal laws in sound change may also be revealed from the second type of features.

In addition to the main goal of this dissertation —the description of the articulatory and acoustic correlates of Shanghai tones and their weights in production and in perception— we asked more general, more ambitious questions.

The general questions asked in this dissertation are to what extent the phonetic cues correlated to the different tones of Shanghai Chinese (1) are universal and predictable, (2) are related to diachronic evolution, and (3) are changing over time in the recent past or in the present period.

The latter question is related to Mazaudon's (2012) conception of phonological and phonetic changes over time. In line with her conception, we believe that the bundle of cues that characterizes a toneme is *dynamic*, and that a phonological system is not static. Some cues are being replaced by others. The synchronic variation in the coexistence of multiple cues to a given distinction may represent a zoomed view of a diachronic evolution in progress. We are therefore more particularly interested in the evolution of the bundle of cues that characterize Shanghai Chinese tones across the two generations of speakers we focus on. This focus is further motivated by the specificity of Shanghai Chinese, which has always been evolving at a fast rate in the recent past (as we have already described) and is now confronted to perhaps even faster changes, due the overwhelming influence of Standard Chinese.

More specifically, the following questions are addressed in this dissertation:

(1a) What are the articulatory and acoustic correlates of Shanghai tones?

(1b) Which phonetic measures are the most efficient to distinguish Shanghai Chinese breathy voice from modal voice?

(1c) To what extent are there intra- and/or inter- speaker variations in the realization of these phonetic correlates?

(1d) How do Shanghai listeners perceive all these cues?

(2) Are redundant features due to coarticulatory effects or are they remnants of diachronic changes?

(3) How do features, distinctive or redundant, evolve in modern Shanghai Chinese?

In the next chapters, we report the data from our physiological, acoustic, and perceptual experiments that address these issues.

4 EXPERIMENTAL INVESTIGATIONS OF THE PRODUCTION OF "YIN" VS "YANG" SYLLABLES

Summary

In this chapter, we report two production experiments in order to provide a phonetic description of the *yin* vs. *yang* tone register distinction in Shanghai Chinese.

In Experiment 1 (§4.1 and §4.2), we examined unchecked and checked target syllables, in all the possible tones and with most of the possible onsets, occurring in monosyllables, or in the first or second syllable of disyllables. Both the speech and EGG signal were recorded. We conducted phonetic-acoustic analyses of these target syllables, including: F0 contour on the target syllable rime, voicing-related measures (VOT and v-ratio), voice quality (on obstruent onsets: HNR; on vowels: various measures of spectral tilt, cepstral peak prominence, and F1 for monosyllables, only H1–H2 for disyllables), durations of all the onsets except phonetically voiceless stops, stop release durations, and rime durations. The EGG data was analyzed, where possible, to measure the open quotient (OQ), which is believed to be a reliable index of breathy voice.

All these analyses roughly confirmed the results found in prior studies, with some exceptions. They were conducted for four groups of speakers, crossing age (young: ~25 y vs. elderly: ~69 y) and gender (male vs. female), with the hypothesis that young and elderly speakers' productions differ, reflecting a change in progress in Shanghai Chinese. We indeed found that young speakers tended to lose breathy voice in *yang* syllables, compared to elderly male speakers who produced *yang* syllables with clearly breathy voice on all measurements. Elderly women also had begun to lose breathy phonation, thus preceding men in this change. Another surprising result was the robust differences in C/V duration patterns with not only shorter durations for *yang* than *yin* obstruent onsets but also longer durations for *yang* than *yin* rimes. Stop durations cannot be measured from the acoustic signal in word-initial position because they are phonetically voiceless in that position.

We ran a pilot Experiment 2 (§4.3) using a motion capture system to measure the time course of lip aperture in labial stops and estimate voiceless closure durations. The preliminary results suggest shorter durations for *yang* than *yin* labial stops in word-initial position.

Using acoustic and physiological methods, we aim at measuring phonetic correlates of Shanghai tones, including essentially F0, voicing, phonation type and duration pattern.

In this chapter, we report the results of Experiment 1, in which acoustic and electroglottographic data were collected simultaneously, and the results of Experiment 2, which is a pilot articulatory study to investigate closure duration in word-initial position. Compared with previous studies, as summarized in Table 11, our study is based on more speakers, richer materials and more comprehensive measurements for phonation types. This table only includes experimental studies that included phonation types among acoustic/articulatory measures.

Table 11. Summary of methods used in previous studies compared to this study.

Study	**Speakers**	**Measurements (acoustic / physiological)**	**Material**
Cao & Maddieson (1992)	2M + 2F (age not reported)	H1–H2, H1–A1 (vowel onset: 30 ms) AF/AP (ratio of the air flow to the air pressure)	9 monosyllabic word pairs (unaspirated vs. voiced labial, dental and velar stops T1-T3), 3 disyllabic word pairs (S2 carrying T1-T5), 3 short sentences
Ren (1992)	3M + 1F (aged late 20s – mid 30s) 2 speakers (gender not reported, who also participated in the acoustic study)	H1–H2, H1–A1 (3 time points), F0 Rothenberg Mask, fiberoptic transillumination	5 reduplicated syllables (including 4 hypothetical names) (aspirated, unaspirated and voiced labial stops, S1 carrying T1, T2 or T3)
Chen (2011)	2M + 4F (born between 1935-50)	VOT, F0, H1-H2 (3 time points)	18 hypothetical disyllabic names (aspirated, unaspirated and voiced dental stops, S1 carrying T1, T2 or T3), in focus vs. no focus condition
Gao (2011)	2 age groups: - young: 6M + 6F - elderly: 2M + 3F	F0, VOT, v-ratio, H1-H2, HNR	60 target syllables (in monosyllables and disyllables), labial and dental stops and fricatives, zero and nasal
Gao et al. (2011)	1F (aged 23)	VOT, H1-H2, HNR, ePGG	26 monosyllables
This work	2 age groups: - young: 6M + 6F - elderly: 4M + 6F	F0, H1–H2, H1–A1, H1–A2, H1–A3, CPP (5 time points), F1, VOT, v-ratio, duration	32 monosyllabic words, 20 disyllabic words with target syllable in S1, 40 disyllabic words with target syllable in S2 onset: labial and dental stops and fricatives, zero and nasal
	2 age groups: - young: 3M + 3F - elderly: 3M + 1F	EGG	32 monosyllabic words onset: labial and dental stops and fricatives, zero and nasal

4.1 Experiment 1 (acoustic data): production of phonetic correlates of Shanghai tones

4.1.1 Recording procedures

Speakers were recorded individually in a quiet room, and simultaneous audio and electroglottographic (EGG) data were collected. For EGG recording details, please see §4.2.1. The audio recordings were made with an AKG C520L headband microphone through an EDIROL external sound board connected to a laptop in stereo mode, using the Sound Studio software: one channel for the audio signal, and the other for the EGG signal. Both signals were coded in WAV format, sampled at 44.1 kHz, with 16 bit resolution.

4.1.2 Participants

Twenty-two native speakers of Shanghai Chinese participated in the recordings. They were divided into two age groups. The young group included 12 speakers (6 male and 6 female) from 21 to 29 years of age (mean 24.9); the elderly group included 10 speakers (4 male and 6 female) from 61 to 79 years of age (mean 68.7) at the time of the recordings. All of the speakers were born in Shanghai urban or suburban areas, except one elderly male speaker who was born in the Jiangsu province but came to Shanghai before the age of one. All of them had spent most of their lifetime in Shanghai urban area. Two of the elderly speakers, a couple, had lived in the Chongming 崇明 county for nearly 30 years, where the Chongming subgroup of Shanghai Chinese is spoken.

All of the young speakers had learned Standard Chinese before the age of eight; all were learning English; one was speaking German and three were speaking French; one of them spoke some Wu dialects of Jiangsu, one spoke Chuansha 川沙 dialect (Songjiang subgroup), and the others did not speak any other dialect than Shanghai Chinese and Standard Chinese. For all young speakers, their self-evaluation of their

competence in Standard Chinese was higher than or equal to that in Shanghai Chinese, but their usage of Shanghai Chinese was generally as frequent as that of Standard Chinese. As for the elderly speakers, all had learned Standard Chinese at primary school or at adult age. None of them spoke any foreign language, but six of them spoke another Wu dialect of Jiangsu or Zhejiang, or Chuansha dialect. For all elderly speakers, both the frequency of usage and the self-evaluation of language competence were higher for Shanghai Chinese than for Standard Chinese.

The result of young speakers' self-evaluation is probably not due to a lesser competence in oral communication for Shanghai than Standard Chinese. Apart from those who communicated with their parents almost always in Standard Chinese from infancy, all the speakers have no difficulties in daily life Shanghai communication, although it is true that code-switching between Shanghai and Standard Chinese is a highly frequent phenomenon observed in the speech of young Shanghai speakers. In fact, one of the reasons of the low self-rating of their competence in Shanghai Chinese is that young speakers realize they often rely on expressions in Standard Chinese instead of local expressions. Another reason might be the young participants' awareness of the difference between their speech and that of their parents' or grandparents'. They judge the older generation variety of Shanghai Chinese more "correct." For these naive speakers, the evolved Shanghai Chinese spoken by the young generation is considered a deviation from the "standard" form. Meanwhile, their own judgment is further supported by that of their parents and grandparents, as well as that of some professors and linguists who are devoted to the promotion of "genuine" Shanghai Chinese. ZHU Xiaonong, who studied experimental phonetics in Australia, took a purist stance, "Quite a lot of young Shanghai people even cannot speak their native language properly." (Zhu, 2006: 2).

4.1.3 Speech materials and design

We used onsets of all manners of articulation except glides and affricates (zero onset, stops, fricatives and nasals) and in all the five lexical tones T1–5 of Shanghai Chinese. We used the monosyllabic context in order to examine the phonetic

correlates of Shanghai tones in their citation forms, and the disyllabic context for the purpose of examining situations of sandhi-modified tone contours. In the disyllabic context, the target syllables were either the first syllable, which should partly maintain its tonal identity, or the second syllable, which should lose its tonal identity by virtue of the tone sandhi whereby the tone contour of the non-initial syllables of a polysyllabic word is determined by the sole first syllable. We used only two rimes, /ɛ/ in unchecked syllables and /aʔ/ in checked syllables, in order to avoid the influence of heterogeneous vocalic context on our different measures of phonation types. Low vowels should be privileged for measures of phonation types to avoid influence of low-frequency formants on the first and second harmonics. These rimes were also chosen because they both occur after almost all the onset consonants and for almost all the tones we used in this study. We could not use the same vowel for both checked and unchecked syllables, because /a/ does not occur after /f, v/ in unchecked syllables, and /ɛ/ does not occur in checked syllables.

Thirty-two monosyllabic and sixty dissyllabic words were produced in a carrier sentence /_ gə ə zɻ ŋo nin tə ə/ (__ 这个字/词我认得的。'__ this character/word, I know it'). The target syllable was elicited in sentence-initial position in order to avoid potential intervocalic voicing of phonologically voiced obstruents. We used three contexts: target syllable as a monosyllable (M), as the first syllable of a dissyllabic word (S1), and as the second syllable of a disyllabic word (S2). Each target syllable carried one of the five citation tones. The unchecked syllables (T1–3) shared the /ɛ/ rime and the checked syllables (T4–5) shared the /aʔ/ rime. The meaning of each word can be found in Appendix 1.

In the monosyllabic context (Table 12), the onset could be empty, or was a labial or dental stop or fricative or nasal, that is, within the /Ø(zero), p, (b), t, (d), f, (v), s, (z), m, n/ set. (There are very few velar stop/nasal onset syllables and no velar fricatives.) There is no T2 /nɛ/ syllable, nor T4 /maʔ/ or /naʔ/ syllable. This made a total of 32 (= 7 onsets * 5 tones – 3) monosyllables. Each of the 22 speakers repeated the word list twice, except for one young female speaker and one elderly male speaker who read the word list only once due to technical problems.

Table 12. List of speech materials of Exp. 1 – monosyllabic words.

Syllable type	Tone	zero	stop	fricative	nasal
unchecked	T1 (yin)	ɛ 哀	pɛ 杯, tɛ 堆	fɛ 翻, sɛ 三	mɛ 蛮[9], nɛ 拿[10]
	T2 (yin)	ɛ 爱	pɛ 板, tɛ 胆	fɛ 反, sɛ 伞	mɛ 美
	T3 (yang)	ɛ 咸	bɛ 办, dɛ 谈	vɛ 饭, zɛ 才	mɛ 梅, nɛ 难
checked	T4 (yin)	aʔ 鸭	paʔ 八, taʔ 搭	faʔ 发, saʔ 杀	—
	T5 (yang)	aʔ 盒	baʔ 白, daʔ 踏	vaʔ 罚, zaʔ 石	maʔ 麦, naʔ 纳

In the S1 context (Table 13), the onset was a labial or dental stop or fricative, that is, belonged to /p, (b), t, (d), f, (v), s, (z)/ set. As in monosyllables, the unchecked syllables shared the /ɛ/ rime and the checked syllables shared the /aʔ/ rime. This made a total of 20 (= 4 onsets * 5 tones) S1 syllables.

Table 13. List of speech materials of Exp. 1 – disyllabic words with target syllable in S1.

Syllable type	Original target tone	After sandhi	stop	fricative
unchecked	T1 (Yin)	**55**–21	**pɛ**.tsɻ 杯子, **tɛ**.ɕiŋ 担心	**fɛ**.ɪʔ 翻译, **sɛ**.ɕi 三鲜
	T2 (Yin)	**33**–44	**pɛ**.tɕɪʔ 背脊, **tɛ**.tsɻ 胆子	**fɛ**.wɛ 返回, **sɛ**.piŋ 伞柄
	T3 (Yang)	**22**–44	**bɛ**.koŋ 办公, **dɛ**.tsɻ 台子	**vɛ**.wø 饭碗, **zɛ**.nəŋ 才能
checked	T4 (Yin)	**33**–44	**paʔ**.paʔ 八百, **taʔ**.sɛ 搭讪	**faʔ**.lɪʔ 法律[11], **saʔ**.tsʰu 塞车
	T5 (Yang)	**11**–23	**baʔ**.pɛ 白板, **daʔ**.zaʔ 踏实	**vaʔ**.kø 罚款, **zaʔ**.dɤ 石头

[9] This character has another reading with T3, so the lexical context was given to elicit the T1 reading.

[10] There are at least two phonetic variants of this character. The speaker was instructed to produce the desired reading.

[11] This word is produced [faʔ.lɪʔ] by most elderly speakers, but [faʔ.lʏʔ] by all young speakers.

In the S2 context (Table 14), the realized tone value of the target syllable depends on the underlying tone of the first syllable. It is realized with a high pitch (noted 44) when the preceding syllable originally carries tone T2, T3 or T5 and with a low pitch (noted 21 or 23) when the preceding syllable carries originally tones T1 or T4. Thus, we studied two sub-contexts: one with realized high pitch (preceded by T2) and one with realized low pitch (preceded by T1). As in the S1 context, the onset was a labial or dental stop or fricative, that is, within the /p, (b), t, (d), f, (v), s, (z)/ set; the unchecked syllables shared the /ɛ/ rime and the checked syllables shared the /aʔ/ rime. This made a total of 40 (= 4 onsets * 5 tones * 2 preceding tones) S2 syllables.

Table 14. List of speech materials of Exp. 1 – disyllabic words with target syllable in S2.

Syllable type	Original target tone	After sandhi	stop	fricative
unchecked	T1 (Yin)	high: 33–**44**	tsɔ.**pɛ** 早班, tɕi.**tɛ** 简单	tsʰɔ.**fɛ** 吵翻, tʰɔ.**sɛ** 套衫
		low: 55–**21**	ku.**pɛ** 科班, sʐ̩.**tɛ** 私单	sɛ.**fɛ** 三番, i.**sɛ** 衣衫
	T2 (Yin)	high: 33–**44**	ɕi.**pɛ** 死板, tɕiɔ.**tɛ** 校对	tɕʰi.**fɛ** 遣返, tsɤ.**sɛ** 走散
		low: 55–**21**	kø.**pɛ** 干贝, tsʐ̩.**tɛ** 猪胆	sɛ.**fɛ** 三反, ɕiɔ.**sɛ** 消散
	T3 (Yang)	high: 33–**44**	pʰɛ.**bɛ** 配备, tsʐ̩.**dɛ** 子弹	tsɔ.**vɛ** 早饭, tɕiɔ.**zɛ** 教材
		low: 55–**21**	tsʰɔ.**bɛ** 操办, tɕi.**dɛ** 鸡蛋	tsʰʐ̩.**vɛ** 糍饭, tʰi.**zɛ** 天才
checked	T4 (Yin)	high: 33–**44**	sʐ̩.**paʔ** 四百, pɔ.**taʔ** 报答	zu.**faʔ** 做法, pɛ.**saʔ** 板刷
		low: 55–**21**	sɛ.**paʔ** 三百, i.**taʔ** 医德	kɛ.**faʔ** 开发, tsʐ̩.**saʔ** 知识
	T5 (Yang)	high: 33–**44**	ɕiɔ.**baʔ** 小白, tsø.**daʔ** 转达	tʰi.**vaʔ** 体罚, ɕy.**zaʔ** 选择
		low: 55–**21**	kɔ.**baʔ** 茭白, ɔ.**daʔ** 凹凸	i.**vaʔ** 衣物, ɕy.**zaʔ** 虚实

4.1.4 Data segmentation

We segmented manually the target syllable in each carrier sentence, using Praat (Boersma & Weenink, 1992-2015). For vowel segmentation, we determined the first zero-crossing point after the consonant offset on the waveform as the beginning of the vowel (Figure 13a) and the end of the visibility of the second formant on the spectrogram as the end of the vowel, excluding the portion corresponding to the voice decay time (Figure 13b).

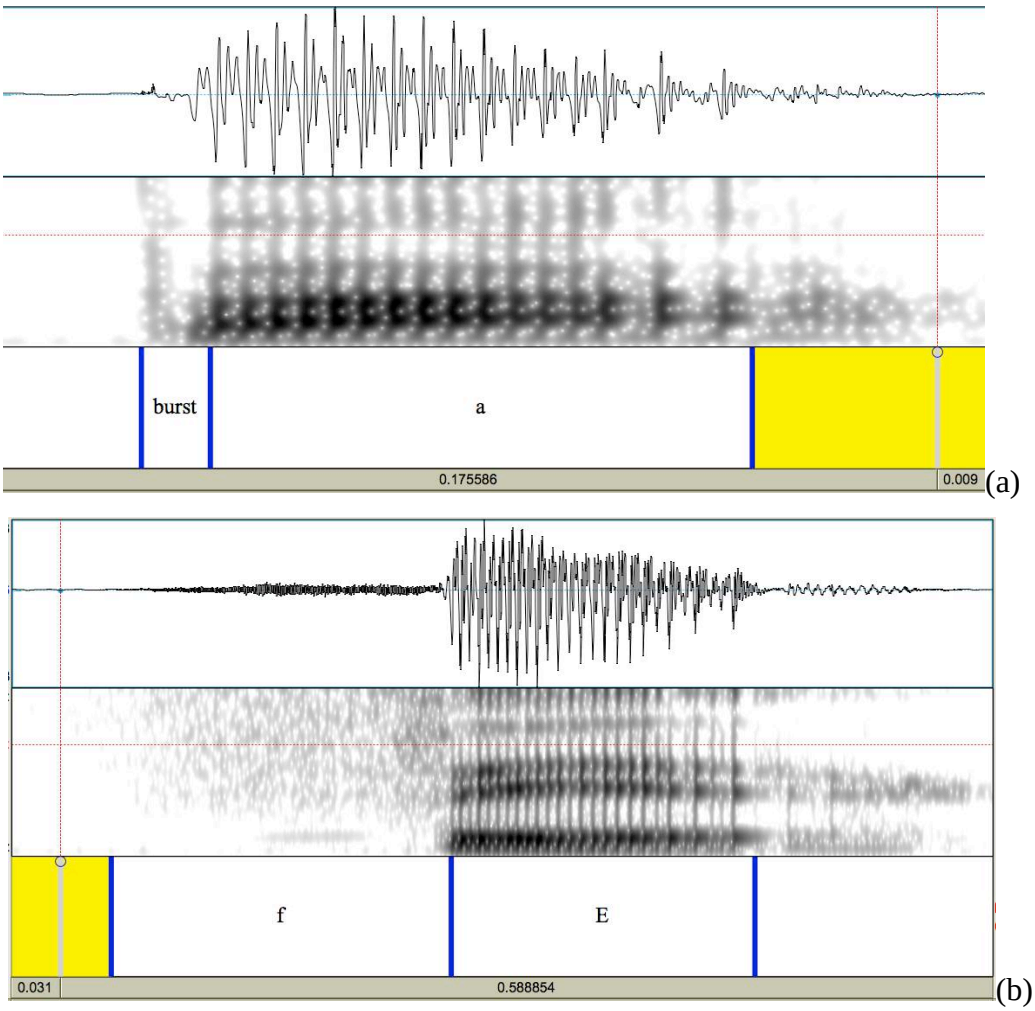


Figure 13. Segmentation of two monosyllables: (a) [paʔ] (T4) and (b) [fɛ] (T1).

For syllables with a stop onset, we determined the beginning of the stop release based on the burst observed on the waveform and on the spectrogram (Figure 13a). For syllables with a fricative or nasal onset, we determined the beginning of the onset

based on the visibility of the signal on the spectrogram and decided that the end of the onset should coincide with the beginning of the vowel (Figure 13b). The dynamic range for spectrogram display in Praat was set at 50 dB.

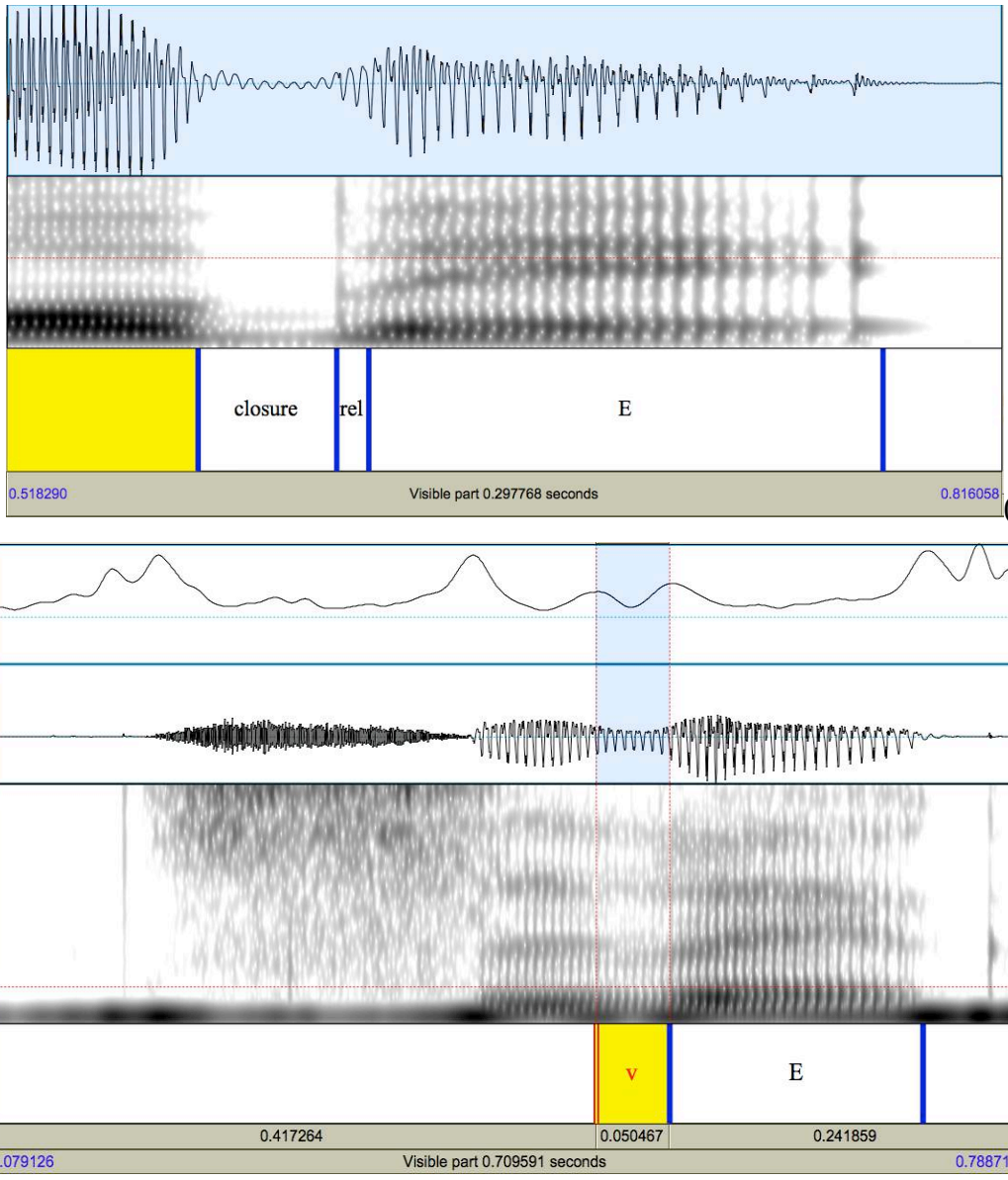


Figure 14. Segmentation of (a) a [bɛ] syllable and (b) a [vɛ] syllable, both second syllables of a disyllabic word. NB: For (b), the function in the top pannel is the spectral derivative.

In intervocalic position, the beginning of the consonant was determined at the end of the preceding vowel, always determined by the end of the visibility of the second formant. Stops in this position were segmented into two parts: closure and release (Figure 14a). In the case of spirantized stops, the measurements were not retained. Voiced fricatives in this position were relatively difficult to segment, due to

their formant structure, which was sometimes in continuity with the adjacent vowels. In order to segment these fricatives, we resorted to the spectral derivative of the speech signal that estimate the spectral variations along time. The troughs of the derivative indicate spectral stability and the peaks of the derivative indicate local maximum velocity of spectral change. We set segments' boundaries at the peaks of the derivative signal (Figure 14b).

4.1.5 Measurements

For each vowel of the target syllable, we measured the fundamental frequency (F0), the duration, and phonation parameters (H1–H2, H1–A1, H1–A2, CPP, F1).

For stop onsets in initial position, we measured the duration of voice onset time (VOT) and harmonics-to-noise ratio (HNR) during the release; for stop onsets in intervocalic position, we measured the release duration, the closure duration and the voicing-ratio.

For fricative and nasal onsets, we measured HNR, consonant duration and voicing-ratio. Details are to be explained in the following.

4.1.5.1 Fundamental frequency (F0)

For the acoustic estimation of F0, we used the cross-correlation method in Praat, setting the default F0 range to [60, 400 Hz] (which generally covered the F0 range of both male and female voices) and the analysis time step at 5 ms. We used default settings for all other parameters (see §4.2.1 for F0 measure from dEGG signal.) For each vowel (i.e., rime) in a target syllable, we calculated mean F0 values over five consecutive equal time intervals covering the entire vowel. That is, we time-normalized the F0 contour data. Because the durations of the rimes ranged from ~70 to ~370 ms (see §4.1.6.4 for more detail), the mean time interval in these F0 contours, ranged from about ~14 to ~74 ms. That is, we computed duration-normalized F0 trajectories consisting of five consecutive, equidistant in time, F0 values.

4.1.5.2 Voicing

Two measures were used to estimate voicing: voice onset time (VOT) and voicing-ratio (or v-ratio). VOT intervals were segmented from visual inspection of spectrograms and a Praat script was used to calculate their duration; v-ratio was the proportion of the voiced part duration out of the total duration of the consonant. The voiced part was determined by the detection of F0, as calculated in Praat using the cross-correlation method. For the v-ratio measurement, the F0 range was set at 60 to 400 Hz, and the time step at 2 ms.

4.1.5.3 Duration

Durations were measured for vowels, fricatives and nasals. As for stop onsets, we could only measure their closure duration in intervocalic position, except when they were spirantized. It was indeed impossible to measure acoustically the durations of phonetically voiceless stop onsets in initial position, This is why we used physiological measurements for the analysis of stop durations (see §4.3).

4.1.5.4 Spectral measures

For the vowel portion of target syllables, we used several measurements as indicators of spectral tilt, including the amplitude difference between the first and second harmonics (H1–H2), between the second and fourth harmonics (H2–H4), between the first harmonic and the first formant (H1–A1), between the first harmonic and the second formant (H1–A2). For all these measurements in the monosyllabic context, we used the VoiceSauce program (Shue, Keating, Vicenik & Yu, 2011), implemented in Matlab. This program also performs corrections of the harmonics' amplitudes, attempting to factor out the influence of vocal tract resonances, as proposed by Iseli & Alwan (2004). The corrected measurements are noted with asterisks (e.g., H1*–H2*). We used a Praat script for the analysis of H1–H2 in the other contexts.

We first intended to use the Voicesauce program for all measurements in all three investigated contexts. Unfortunately, we accidentally discovered an anomaly in the program. At vowel onset, for a period of several milliseconds, the amplitudes computed for the harmonics H1 and H2 were quite often negative, entailing noisy estimations of spectral tilt, for example of of H1–H2. This anomaly is typical of the programming errors occurring at the margins of a computation domain: the few first analysis windows at the beginning of the vowel probably included aperiodic noise from the preceding consonant. Since we averaged spectral tilt values on rather large intervals (width = 20% of vowel duration), the consequences of this anomaly were quite minor in the monosyllabic context, in which the vowel was always quite long. As can be seen in §4.1.6.3.2, similar results obtained using Praat vs. Voicesauce. Because we also wanted to compare the various measures of spectral tilt (using a linear discriminant analysis), which all are available in Voicesauce, we decided to retain the Voicesauce data, averaged on five consecutive 20% time intervals, but only for monosyllables. In the disyllabic contexts (S1 and S2), however, vowels had shorter duration and the spectral tilt values in the first (and perhaps also the last) time intervals thus were less reliable. We therefore only used a custom Praat script, which was limited to H1–H2 estimation, but was free of computation errors at domain margins. H1–H2 values were computed using a 30 ms analysis window, at five time points for unchecked syllables —at 15 ms from vowel onset, at 15 ms before vowel offset, and at 3 intermediate time points (25%, 50%, and 75% in the vowel)— and three time points for checked syllables because of their short durations, with only one intermediate time point (50% in the vowel).

4.1.5.5 Measures of noise

In order to estimate the amount of aperiodic noise, as possibly correlated with breathy phonation, we measured HNR (harmonics-to-noise ratio) for consonant onsets, and CPP (cepstral peak prominence) for the vowel rime. Both measures evaluate the aperiodicity of the speech signal. In a recent study (Gao, 2011), we measured HNR also in the vocalic part, but no significant *yin–yang* difference was found. Since CPP measure is more widely used for vowels, we retained this measure for vowel parts.

A cepstrum is a spectrum of the log power spectrum of a signal. If the spectrum of a signal exhibits a clear harmonic structure, indicating that the signal is strongly periodic, the cepstrum is chararaterized by prominent peaks at multiples of the time corresponding to the fundamental period of the signal (the fundamental "quefrency"). CPP measures the prominence of the first of these peaks relative to the overall cepstrum signal (de Krom, 1993).

For HNR measurement, we used the Harmonicity (cc) function in Praat with the following parameters: 2 ms time step, 60 Hz pitch floor, 0.03 silence threshold, and 4.5 periods per window.

4.1.6 Results

Unless otherwise specified, the error bars represent a 95% confidence level (i.e., standard error multiplied by 1.96), and separate statistic analyses were conducted for unchecked and checked syllables.

4.1.6.1 Fundamental frequency (F0)

4.1.6.1.1 Citation tones in monosyllables

The averaged F0 trajectories of the citation tones are plotted in Figure 15, for checked and unchecked tones, and for each speaker group. The 5-level tone letter notation for the three unchecked tones (T1: 53, T2: 34, T3: 23) by Xu and Tang (1988) reflect globally quite well the phonetic reality, in that T1 is falling, and both T2 and T3 rising but T2 with a higher contour. Looking at the detail, T1 falls rather abruptly and deeply, whereas T2 is almost level, that is, only slightly rising. The F0 difference between T2 and T3 lies in their onsets, but their offsets meet at the end. The checked tones (T4: 55, T5: 12) are usually noted with a single tone-letter or underlined values indicating their short duration: they are perceived as tones with little F0 variation. The results show, however, that T4 is slightly falling and T5 slightly rising. For each speaker group, the F0 onsets of T1 and T4, as well as of T3 and T5, are very close to one another.

The F0 range (i.e., the difference between maximum and minimum F0 in the F0 trajectories) according to group is worth noting. It is larger for female than male speakers when expressed in Hertz units. Yet, when converted to semitones, elderly male speakers exhibit an F0 range quite close to that of female speakers, whereas young male speakers have a much lower range than the others (see Table 15 for details). The small tonal space might have a consequence on the distinction between certain tones, especially those with similar contours, such as T2 and T3. Young male speakers exhibit a difference of only 22 Hz between T2 and T3 onset, which is just marginally sufficient for perceptual contrast in languages (’t Hart, 1981).

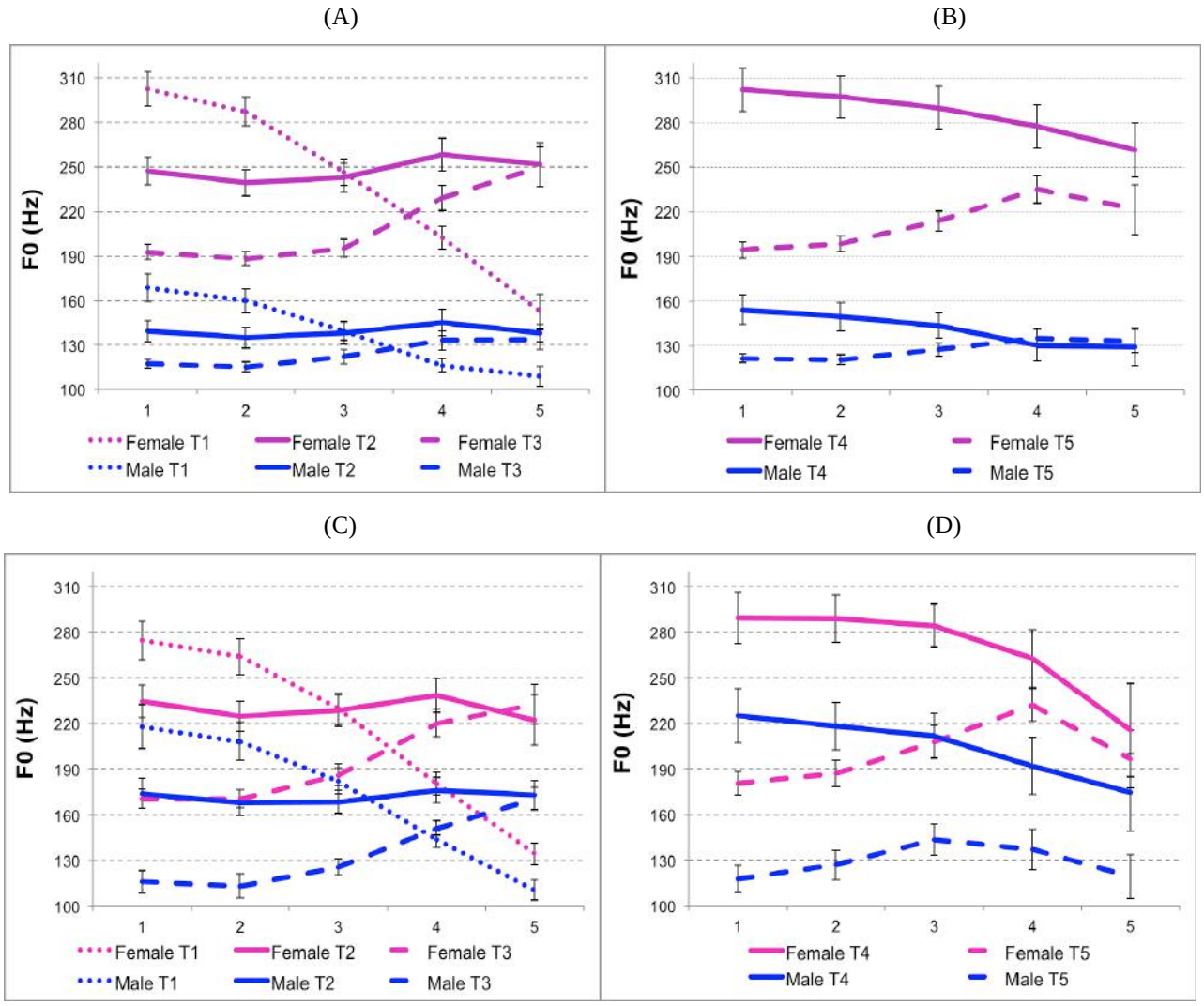


Figure 15. Average F0 trajectories of the five tones in citation (M), time-normalized into 5 intervals: A-B for young speakers, C-D for elderly speakers; pink for female and blue for male speakers. Left pannel: T1-3; right pannel: T4-5.

For young speakers, there is a clear boundary between the F0 range of the male speakers and that of the female speakers, whereas for elderly speakers, there is an overlap between male and female speakers. It seems that male and female F0 ranges are converging with age. The difference in mean F0 between males and females is greater among young speakers than elderly speakers (105 Hz vs. 60 Hz). Taking a closer look at the results, we observe a lowering of F0 range for elderly female speakers compared to young female speakers, and symmetrically a raise of F0 range for elderly male speakers compared to young male speakers. Moreover, elderly male have an overall higher F0 than young male speakers, and elderly female have slightly lower F0 than young female speakers (Table 15).

Table 15. F0 range and mean F0 across the 5 tones for elderly vs. young and male vs. female speakers.

	F0 range (Hz, converted to semitone)		**Mean F0 (Hz)**	
	Young	Elderly	Young	Elderly
Male	60 Hz (7.8 st)	114 Hz (12.2 st)	134	162
Female	150 Hz (11.8 st)	155 Hz (13.2 st)	239	222

It should be noted that among our four elderly male speakers, the one who has the highest mean F0 is an amateur singer of Huju 沪剧 'Shanghai opera'. As most varieties of the Chinese opera, Shanghai opera singers usually sing with very high-pitched voice. An acoustic study on Beijing opera singing, for example, reports the Beijing opera singers' highest scale tone at 611 Hz (pitch about Eb5), which is higher than Western tenors' general pitch (lower than 523 Hz, i.e. C5) (Sundberg, Gu, Huang & Huang, 2012). Concerning F0 in speech, western soprano and tenor singers exhibit a significantly higher F0 than non-singers (Brown, Morris, Hollien & Howell, 1991). Hence, it is not surprising that the amateur opera singer uses a higher pitch level for speech compared to the other male speakers.

However, this speaker has little influence on the general mean F0, nor on the general trend of an overlap between elderly male and female F0 ranges. His data excluded, the mean F0 of the elderly male speakers goes down to 157 Hz from 162 Hz, making little difference. We thus retained his data in all subsequent analyses.

4.1.6.1.2 First syllable in disyllabic words (S1)

The F0 trajectories in S1 rimes are plotted in Figure 16, for checked and unchecked tones, and for each speaker group. They are all realized as level tones, as Xu and Tang (1988) described using Chao's (1930) five tone letters (T1 > 55; T2 > 33; T3 > 22; T4 > 33; T5 > 11). The difference between our data and their tone-letter notations is that the pitch level is lower for T3 than T5, and higher for T4 than T2.

For unchecked syllables (T1-3), Similarly to monosyllables, for young speakers, male and female F0 ranges are separated, but for elderly speakers, they are overlapped.

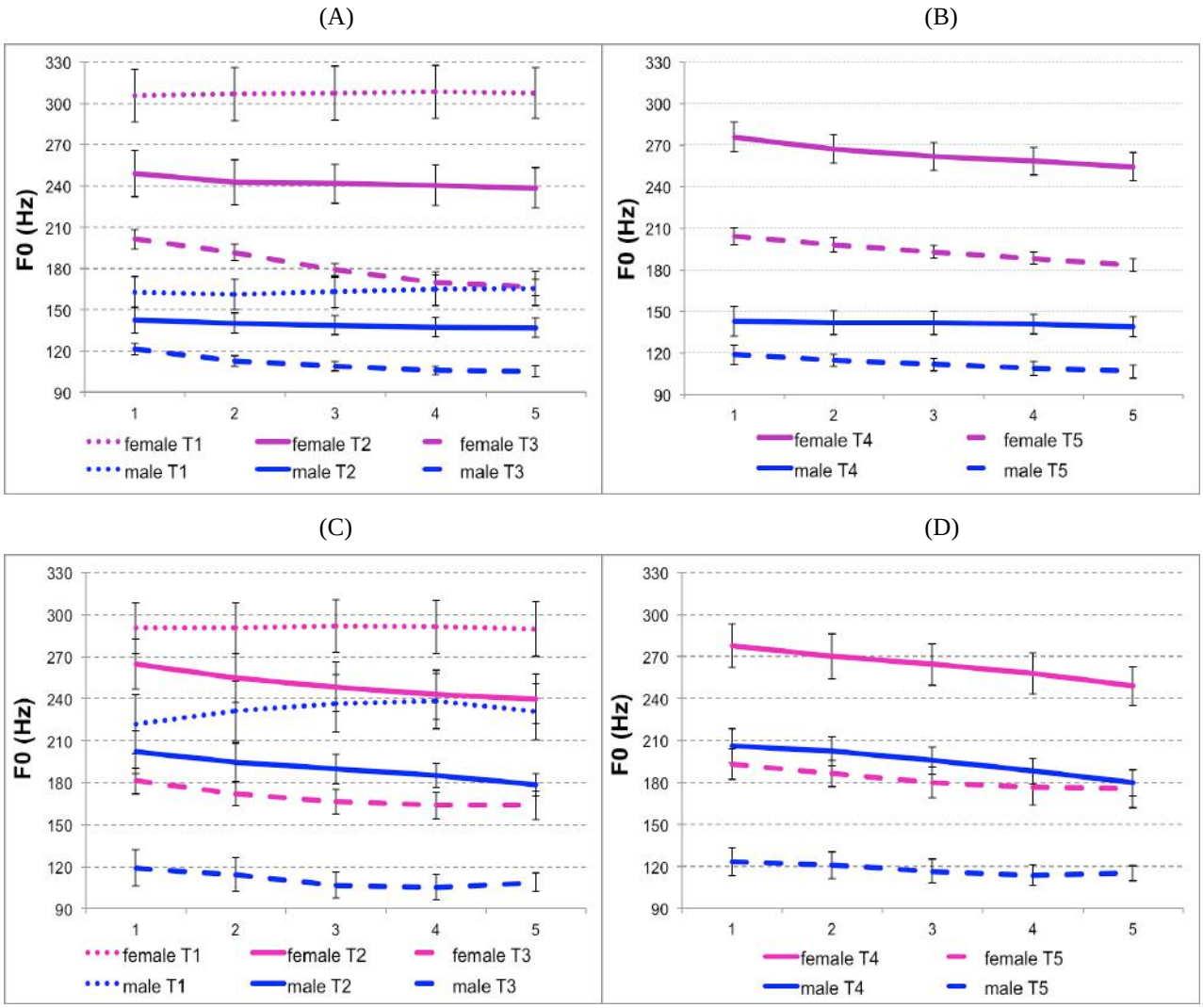


Figure 16. Average F0 trajectories in originally T1-T5 S1 rimes, time-normalized into 5 intervals: A-B for young speakers, C-D for elderly speakers; pink for female and blue for male speakers. Left: T1-3; right: T4-5.

4.1.6.1.3 Second syllables in disyllabic words (S2)

The F0 trajectories of S2 rhymes are plotted in Figure 17 for checked and unchecked tones, and for each speaker group. As explained in §2.1.3.2, in a dissyllabic word, the second syllable loses its underlying tone and the tone contour realization depends on (underlying) tone of the first syllable. Indeed, the figures show that, when

preceded by an originally rising T2 syllable, the second syllable, regardless of its underlying, original tone, is realized with a high-pitch level contour; when preceded by an originally falling T1 syllable, the second syllable is realized with a low-pitch falling contour, regardless, again, of its underlying tone. Xu and Tang (1988) describe post-T1 S2 as 21 and post-T2 S2 as 44 (tone-letter notation). In our data, the post-T1 and post-T2 F0 onsets are very close to each other for unchecked syllables, and their difference is largest at F0 offset.

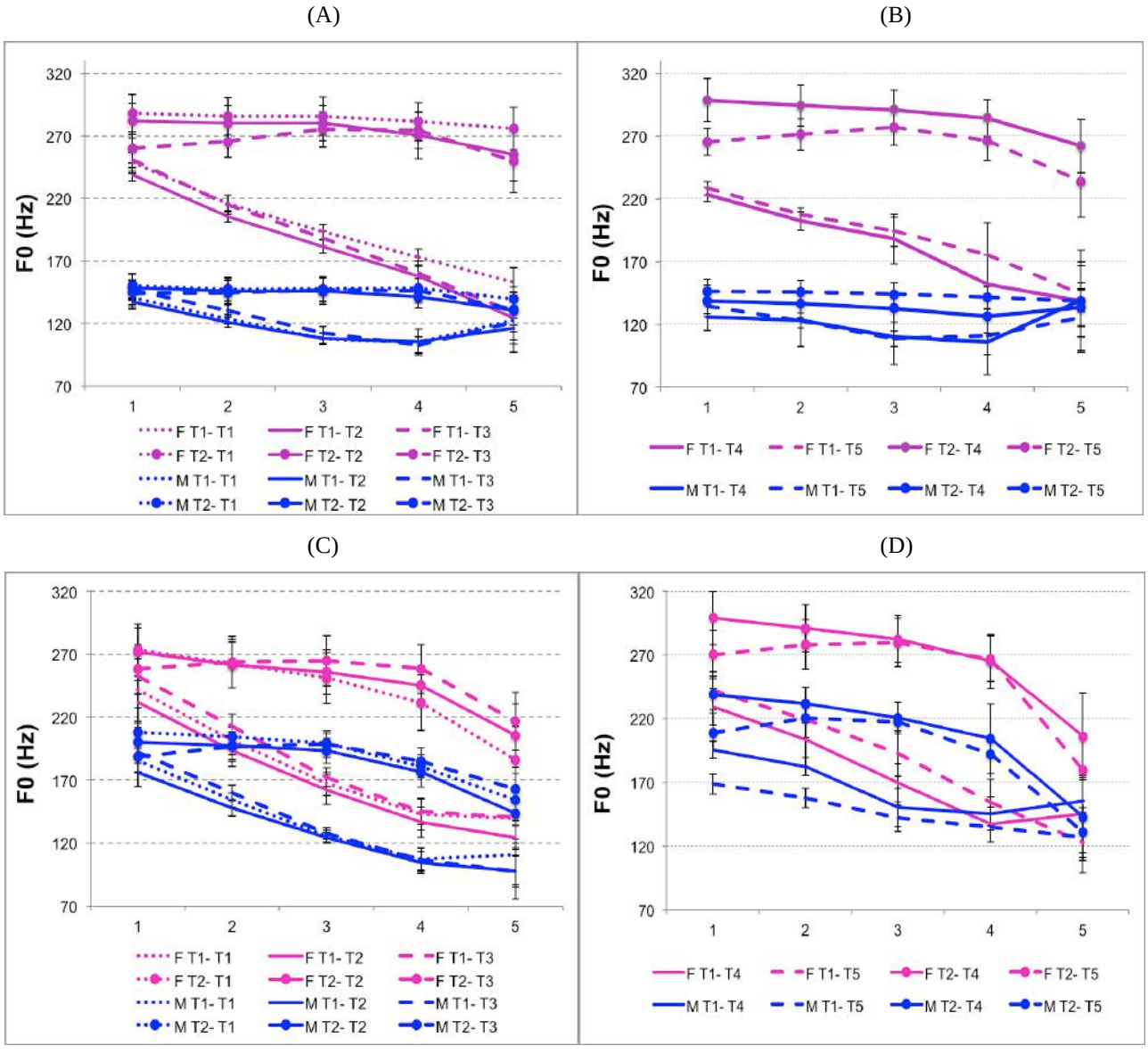


Figure 17. Average F0 trajectories in originally T1-T5 S2 rimes (left: T1-3, right: T4-5), preceded by a T1 (unmarked) or T2 (marked) syllable, time-normalized into 5 intervals: A-B for young speakers, C-D for elderly speakers; pink for female and blue for male speakers.

4.1.6.1.4 Variations of F0 onset in the second syllable

As explained in the previous section, S2 syllables are, in principle, neutralized for tone category. The *yin* vs. *yang* distinction is maintained through phonetic voicing only. However, voiced stop onsets (phonetically and phonologically) are reported to lower F0 at vowel beginning (Ren, 1992: 51; Chen, 2011; Chen & Downing, 2011), and are therefore called "depressors," a term frequently used for describing tonal processes in Bantu languages (e.g., Downing, 2009). The phonetic implementation, however, is different in Bantu languages such as Zulu (or "isiZulu") and in Shanghai Chinese: depressors have F0-lowering effects over the entire syllable in Zulu but only on the vowel beginning in Shanghai Chinese (Chen & Downing, 2011). However, in our data shown in Figure 17, voiced onset S2 syllables did not have systematically lower F0 onset than voiceless onset ones.

Figure 17 shows five time points time-normalized contours, with an approximate time interval between successive data points of ~36 or ~22 ms for unchecked or checked syllables, respectively. Here, however, in order to examine possible fine variations in the F0 onset of S2 syllables, we measured F0 with a smaller time step of 5 ms, within the first 100 ms of the rime. Moreover, we normalized F0 contours, converting the raw F0 values in Hz of each subject into z-score values[12] to minimize intra-speaker variations. These more detailed measurements are shown in Figure 18, averaged across onsets (stops and fricatives) and across all speakers. Figure 18 shows that the pattern is different according to the preceding tone, T1 or T2. Remember that S2 is realized with a low pitch (or rather low falling pitch) when preceded by a T1 syllable, and with a high pitch when preceded by a T2 syllable. When the first syllable's tone is T2 (right column), the "depressor" effect of voiced onsets is clearly observed. At 5 ms from rime beginning, the raw F0 data show a difference of 27 Hz between voiceless and voiced onsets for unchecked syllables, and a difference of 31 Hz for checked syllables. However, when the first syllable's tone is T1 (left column), the voicing of the S2 syllable onset (or, equivalently, its underlying tone) has little effect

---

[12] $z = \frac{X - \overline{X}}{S}$, where X stands for each F0 value, overbarred X for the mean F0 value, and S for the standard deviation of F0 for each speaker.

on F0 onset height. Voiced onsets even seem to yield slightly higher F0 onsets than voiceless onsets and this pattern applies to both unchecked (top pannel) and checked (bottom pannel) syllables, and to all speaker groups, contrary to the data reported by Chen (2011). Note that Ren's (1992: 138ff) data are averaged across all sandhi patterns.

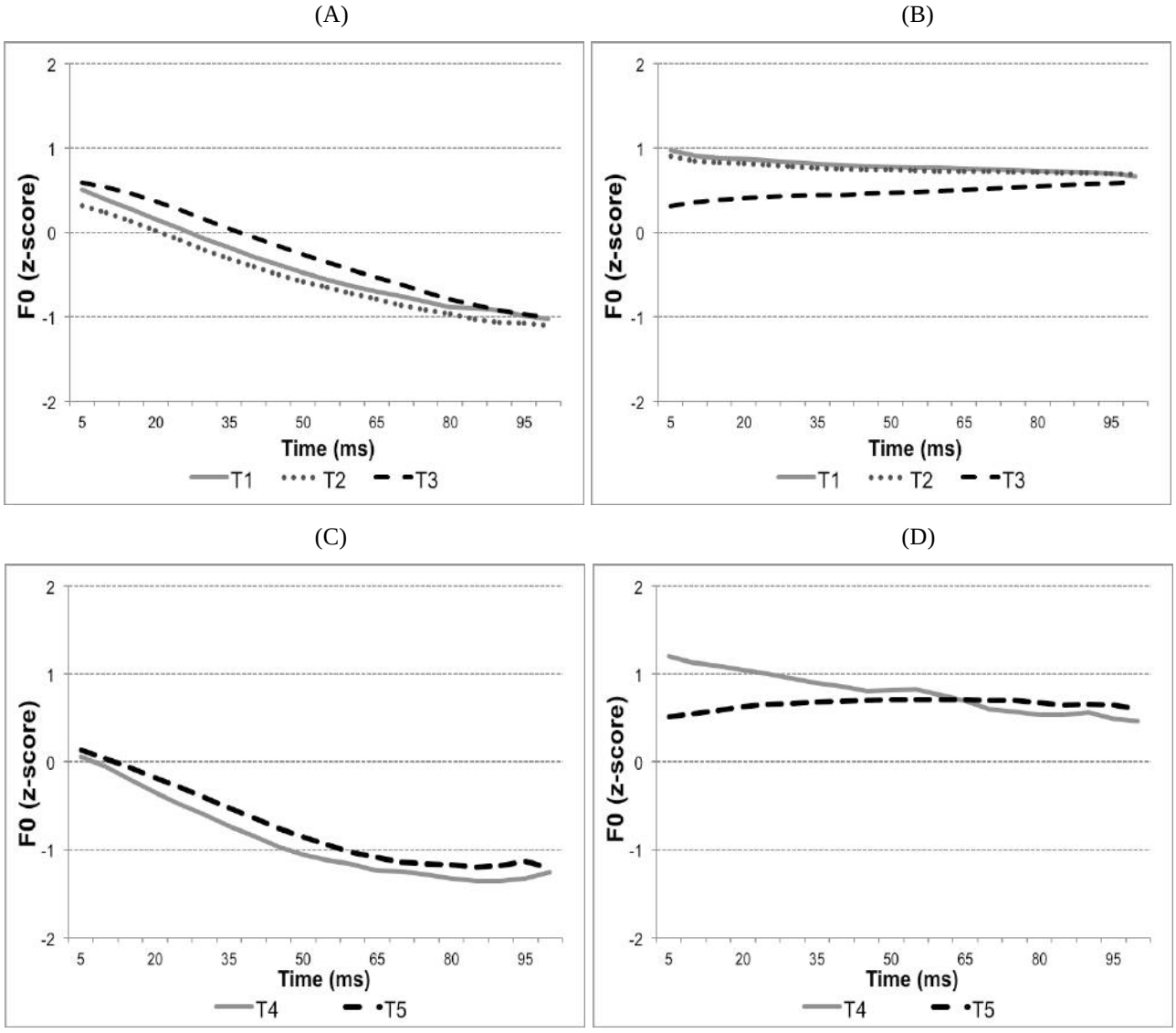


Figure 18. F0 contours (z-score normalized within speakers) from 5 ms to 100 ms of the beginning of the rime of S2 syllable as a function of S2's underlying tone, averaged across onsets and speakers: A-B for unchecked S2, C-D for checked S2; left: after T1, right: after T2.

Looking at Figure 18, one can observe that for post-T2 syllables, the depressor effect of voiced onsets lasts for about 100 ms (almost the entire portion analyzed) in unchecked syllables, but only for about 50 ms in checked syllables. This is presumably

due to the shorter syllable duration of checked than unchecked syllables. We therefore averaged F0 values across the first 50 ms of the rime for all syllables in order give an overall picture of the F0 onset effects. Figure 19 summarizes these effects in showing the averaged (z-score normalized) F0 values at S2 rime onset, according to the first syllable's tone (T1 vs. T2) and the (target) S2 syllable's underlying tone. Originally T3 or T5 yang S2 syllables are shown in brown so that they can be compared visually to T1 and T2 or to T4, respectively, possibly illustrating the F0-depressor effect. This depressor effect (F0 lowering in yang S2 syllables) can indeed be observed when the first syllable's tone is T2 (right side) but not when it is T1 (left side).

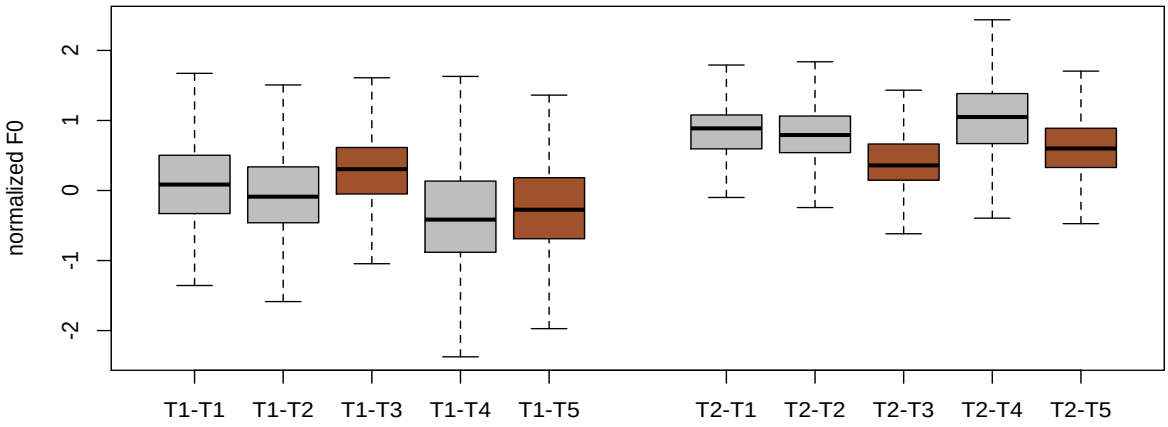


Figure 19. Boxplot of mean F0 in the first 50 ms of S2 rime according to first syllable's tone (left side: T1; right side: T2) and S2 syllable's original tone. Originally T3 and T5 (i.e., yang) S2 syllables are shown in brown (muddy color).

In order to substantiate these observations, we ran by-subject ANOVAs separately on unchecked and checked syllables. In both analyses, z-score normalized *F0* (across the first 50 ms) was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Underlying tone* (unchecked: T1, T2, T3; checked: T4 vs. T5), *Sandhi pattern* (post-T1 vs. post-T2) and *Onset* (/p(b), t(d), f(v), s(z)/) were within-subject factors.

For unchecked syllables, *Underlying tone* was significant overall, $F(2,36)=4.3$, $p<.05$, and interacted with *Sandhi pattern*, $F(2,36)=62.5$, $p<.0001$. In post-T1 sandhi, F0 onset was significantly higher for T3 than T1 syllables (0.18>–0.004), $F(1,18)=13.0$,

p<.005, or than T2 syllables (0.18>–0.14), F(1,18)=37.2, p<.0001. (The T1 vs. T2 difference was also significant, F(1,18)=16.1, p<.001.) In post-T2 sandhi, on the contrary, F0 onset was significantly lower for T3 than T1 syllables (0.41<0.84), F(1,18)=40.8, p<.0001 and T2 syllables (0.41<0.80), F(1,18)=36.8, p<.0001. T1 and T2 onsets did not differ (0.84≈0.80), F(1,18)=1.2, p=.29. The *Underlying tone* × *Onset* interaction was not significant, F(6,108)=1.7, p=.12, but the *Underlying tone* × *Onset* × *Sandhi pattern* interaction was, F(6,108)=9.9, p<.0001. These interactions reflect the fact that, whereas in both sandhi patterns the *Underlying tone* effect was larger in absolute value for fricative than stop onsets, it went in opposite directions for the two sandhi patterns. For post-T1 sandhi, the *positive* T3 vs. T1-2 differential was larger for fricative onsets (Δ=0.42, ps<.0001) than for stop onsets (Δ=0.09, ps<.05). For post-T2 sandhi, the *negative* T3 vs. T1-2 differential was also larger for fricative onsets (Δ= –0.55, ps<.0001) than for stop onsets (Δ= –0.28, ps<.005). Both *Gender* and *Age* were significant (*Gender*: F(1,18)=6.7, p<.05, *Age*: F(1,18)=6.4, p<.05). Overall, normalized F0 was higher for female than male speakers (0.45>0.29) and for elderly than young speakers (0.48>0.25), as already found in the previous sections. Finally, *Underlying tone* did not interact with *Gender* (F(2,36)<1, n.s.) or *Age* (F(2,36)=1.7, p=.19). That is, the same patterns of depressor (or non-depressor) effects were found for all four speaker groups.

For checked syllables, *Underlying tone* was marginal overall, F(1,18)=4.1, p=.059, but the *Underlying tone* × *Sandhi pattern* interaction was significant, F(1,18)=35.3, p<.0001. The pattern was similar to that observed with unchecked syllables: for post-T1 sandhi, F0 onset was significantly higher for T5 than T4 syllables (–0.36>–0.51), F(1,18)=4.8, p<.05; for post-T2 sandhi, it was significantly lower for T5 than T4 syllables (0.64<0.98), F(1,18)=35.4, p<.0001. The *Underlying tone* × *Onset* × *Sandhi pattern* interaction was significant, F(3,54)=4.0, p<.05: the *Underlying tone* × *Onset* interaction was significant only for post-T2 sandhi, F(3,54)=9.9, p<.0001, but not for post-T1 sandhi, F(3,54)=1.3, p=.29. For post-T2 sandhi, *Underlying tone* was significant for all onsets at least at the p<.0001 level, except for dental stop, F(1,18)<1. This is presumably due to an unexpected pronunciation of the /taʔ/ syllable in the word 报答 /pɔ.taʔ/ (T2-T4) 'pay back.' The normally voiceless /t/ onset for 答 was produced as voiced by many speakers, especially the young speakers. The frequencies

of voiced instead of voiceless realization of all speaker groups are summarized in Table 16. The reason for this voiced realization remains unknown, but it explains why no F0 onset difference was found between /da?/ and /ta?/ syllables: the latter were frequently realized as [da?]. *Gender* was marginal, $F(1,18)=3.9$, $p=.065$, and *Age* significant, $F(1,18)=25.5$, $p<.0005$: normalized F0 was higher for elderly than young speakers (0.44>0.004). Finally, *Underlying tone* did not interact with *Gender* or *Age*, Fs(1,18)<1, showing, again, that the same patterns of depressor (or non-depressor) effects were found for all four speaker groups.

Table 16. Number of occurrences of /t/ realized [d] in /pɔ.ta?/ syllable.

	Female	**Male**	**Total**
Young	11/11	7/12	**18/23**
Elderly	3/10	0/7	**3/17**
Total	**14/21**	**7/19**	**21/40**

Another interesting point to note is that the time course of the F0 onset depressor effect of voiced stop onsets for T2 sandhi pattern differed between young and elderly speakers. As shown in Figure 20, for unchecked syllables, the F0 depression induced by the /b, d/ stops lasted for about 170 ms for young speakers, but only about 50 ms for elderly speakers. For unchecked syllables, the duration of the F0 depression induced by /d/ onset was about the same for both age groups (between 60 and 80 ms). (The data for dental stops were not shown for checked syllables because /ta?/ was frequently realized as [da?] as explained earlier.) For young speakers, the F0 depressor effect seems to depend on the rime duration: the longer the syllable, the longer the F0 was depressed by voiced stops. But for elderly speakers, the F0 was depressed locally and rather independently from the rime duration. Note that rime duration did not differ between young and elderly speakers (see §4.1.6.4.3.2). No noticeable difference was observed for fricative onsets: the depressor effect lasted for quite a long period of time for both young and elderly speakers.

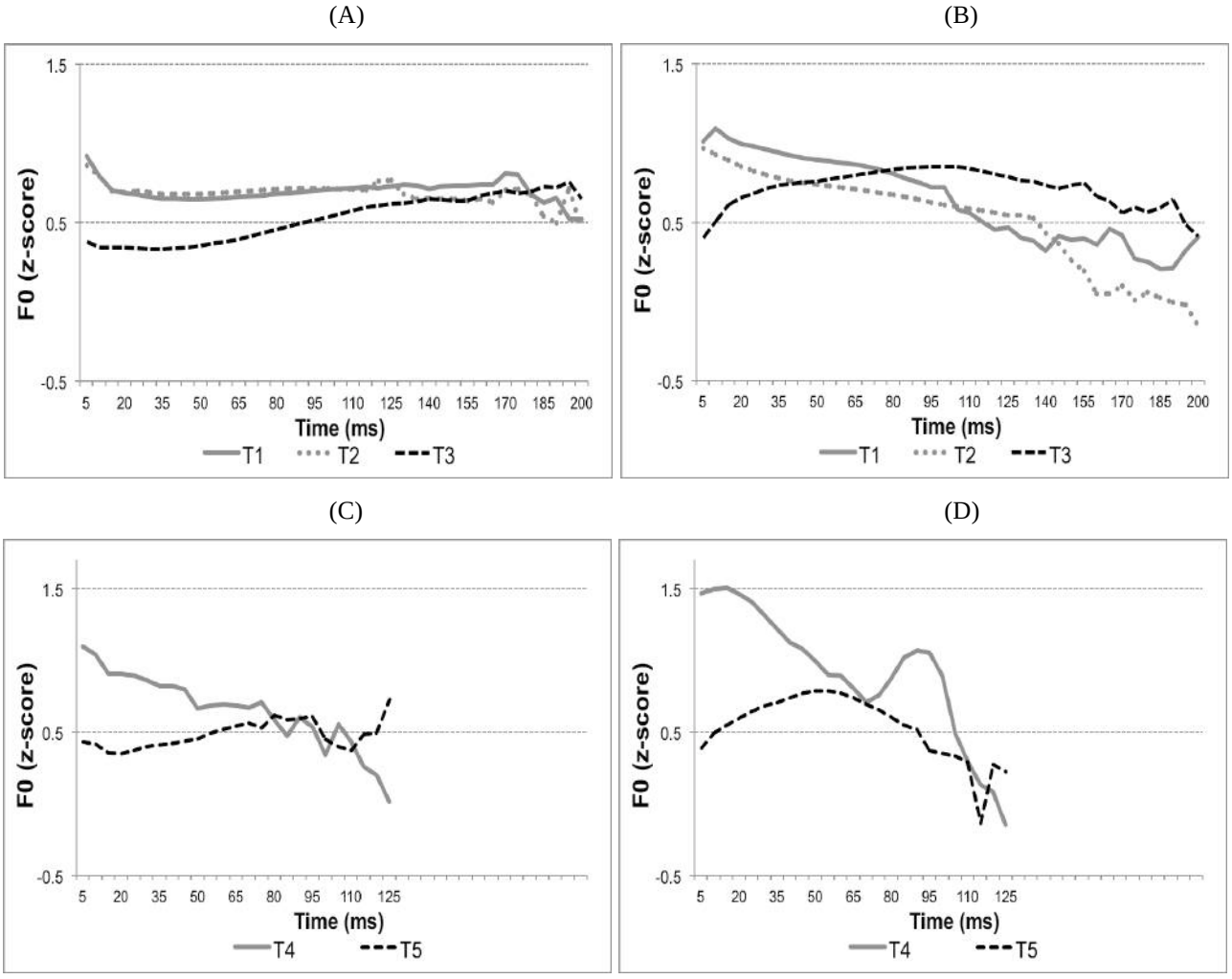


Figure 20. Time course of the F0-onset depressor effect of voiced stop onsets: young speakers (left), elderly speakers (right); unchecked syllables (top), checked syllables (bottom).

4.1.6.1.5 Discussion

- **Tone description**

Globally, the F0 data of Shanghai syllables in monosyllabic and disyllabic contexts presented in this section are in conformity with Shanghai tone descriptions in the literature, in terms of their relative value within a tone or between tones. Some minor deviations, however, can be observed: both rising T2 and T3 are often noted with the same rising step (34 and 12 or 23, respectively), but in our data, T2's range is narrower than T3's; also, T4 is noted with a single tone letter (4 or 5) indicating a short, flat tone, but in our data, T4 is slightly falling.

- **Cross-gender and cross-age F0 difference**

Our data show a clear separation between the F0 ranges of Shanghai young male and female speakers. Furthermore, Shanghai male speakers tend to raise their F0 with age, and female speakers tend to lower their F0 with age. This could be explained by generally agreed vocal aging processes, which predict that, as far as non-singers are concerned, F0 for speech increases with age for males but decreases with age for females (e.g., Mysak, 1959; Brown et al., 1991). However, in the absence of a longitudinal study, we cannot exclude other linguistic or extralinguistic factors than age-related voice change to explain this generational variation.

- **Voiced onsets' F0 depressor effect**

We examined the depressor effect of voiced onsets in the second syllable of a disyllabic word. Previous studies showed a depressor effect of voiced stops. We also included fricative onsets. Previous studies did not explore or did not find differences in depressor effects according to the first syllable's tone. We examined separately the post-T1 and post-T2 contexts. When the first syllable carries T1 (post-T1 context), the second syllable is realized with a low pitch; when the fist syllable carries T2 (post-T2 context), the second syllable is realized with a high pitch. We found a depressor effect of voiced onsets only in the post-T2 context, and for both voiced stops and fricatives.

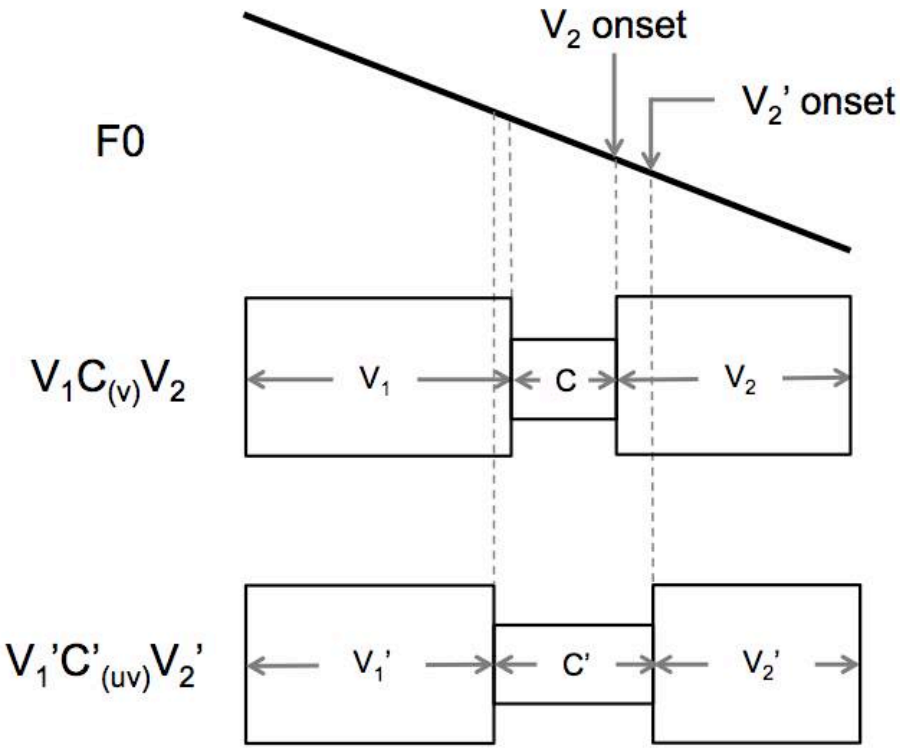


Figure 21. Schema of independent timing tiers for disyllables with S1 carrying T1. Above: F0 movement; middle: VCV segments with intervocalic voiced obstruents (shorter); below: VCV segments with intervocalic voiceless obstruents (longer).

Our explanation for this difference is that, in addition to the depressor effect, there is another mechanism contributing to the F0 onset height in the second syllable, which is determined by the ballistic movement of the entire word F0 contour. We propose that the timing of the F0 contour on prosodic words is programmed independently from the timing of the segments. As shown in Figure 21, when the first syllable's tone is T1, the F0 contour falls monotonically at a constant velocity over the prosodic word, regardless of segments' duration. When the vowel onset of S2 occurs earlier, the F0 onset in S2 is higher on the whole prosodic word ballistic F0 trajectory. This is what happens when the word-medial consonant is voiced because it is shorter in duration than its voiceless counterpart (see §4.1.6.4.3). Although the duration of the vowel of the first syllable may be somewhat longer when followed by a voiced than a voiceless consonant, the vowel duration difference seems smaller for the vowel than for the consonant (Zhu, 1999: 191ff). Hence, when the word-medial consonant is voiced, the vowel onset of the second syllable occurs earlier and thus tends to begin

with a higher F0 onset, inasmuch the "ballistic" F0 movement is concerned. We may assume that the presumed depressor effect of the word-medial voiced onset is not sufficient to overcome the ballistic effect so that, as a net result, the F0 onset of the following vowel tends to be slightly higher rather than lower than after a voiceless S2 onset. In contrast, when the first syllable's tone is T2, the ballistic effect and the depressor effect of voiced onsets add to one another instead of cancelling each other as in the post-T1 context, yielding a clearly lower F0 onset after voiced than voiceless medial consonants.

4.1.6.2 Voicing

Two measures were used to index voicing: (1) voice onset time (VOT) for word-initial unaspirated stops and (2) voicing ratio (henceforth, v-ratio) for word-medial unaspirated stops and for fricatives in all contexts.

4.1.6.2.1 Onsets in monosyllables

The first part of this section concerns stop onsets, and the second part fricative onsets in monosyllables.

- **Voicing of stop onsets**

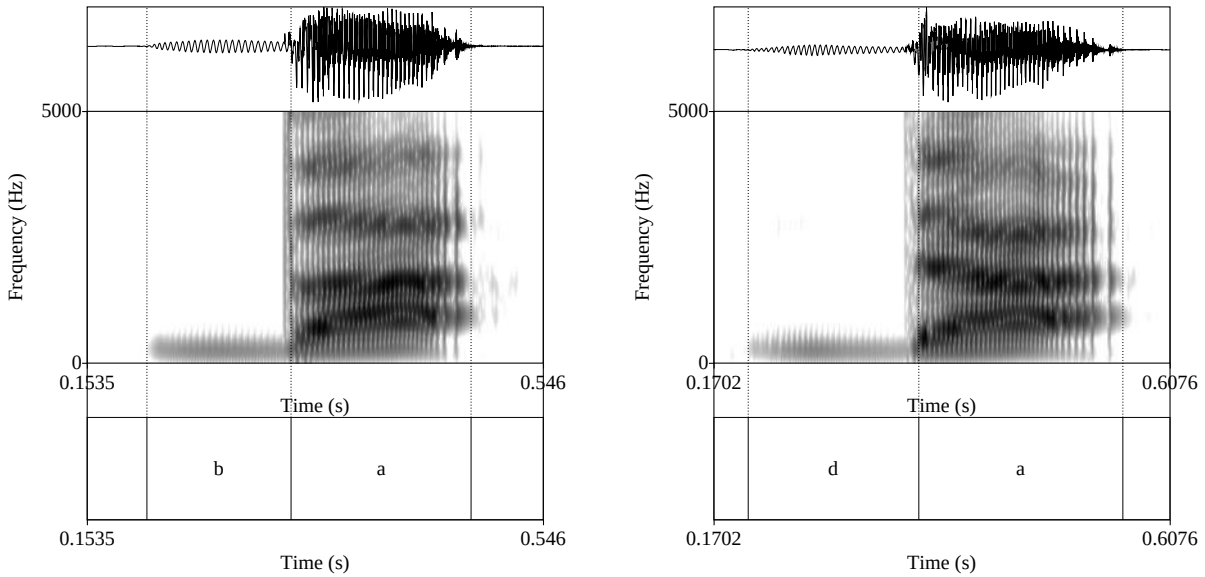


Figure 22. Speech signals and spectrograms of prevoiced word-initial stop onsets in checked syllables produced by a young female speaker. (CD: 4.1.6.2.1_fig23)

As described in the literature, phonologically voiced stops in word-initial position are pronounced without pre-voicing, with rare exceptions. Among all the 22 speakers, only one young female speaker aged 24 produced these stops with negative VOTs very occasionally (twice in monosyllables, and twice in the S1 context), and exclusively in checked syllables. Figure 22 shows the speech signals and spectrograms of two checked monosyllables, /ba?/ and /da?/, which she pronounced with pre-voiced stop onsets. The

VOTs of these [b] and [d] onsets were respectively –124 ms and –163 ms. This female speaker speaks English and German and has stayed in Bremen (Germany) for one year. Her pronunciation of Shanghai Chinese might be influenced by her exposure to the German voicing contrast, but it remains unclear why she produced pre-voiced stops only in checked syllables. The negative VOTs of this speaker were excluded from the VOT data analyses.

Figure 23 shows the average VOTs of stop onsets according to the syllable tone in the monosyllabic context, pooled across speaker groups and onset types; Figure 24 separately shows (A) the young speakers' and (B) the elderly speakers' data. The detailed data are reported in Appendix 2. In each figure, the dashed vertical line separates the three tones that apply to unchecked syllables (T1-3) on the left and the two tones that correspond to checked syllables (T4-5) on the right. Separate statistics were conducted for unchecked and checked syllables.

To summarize the main trends of the statistical analyses run on the data, overall, the VOTs of syllable-initial stops are significantly longer for T3 (*yang*) than T1 and T2 (*yin*) syllables, and do not differ between T1 and T2. VOTs are likewise longer for T5 (*yang*) than T4 (*yin*).

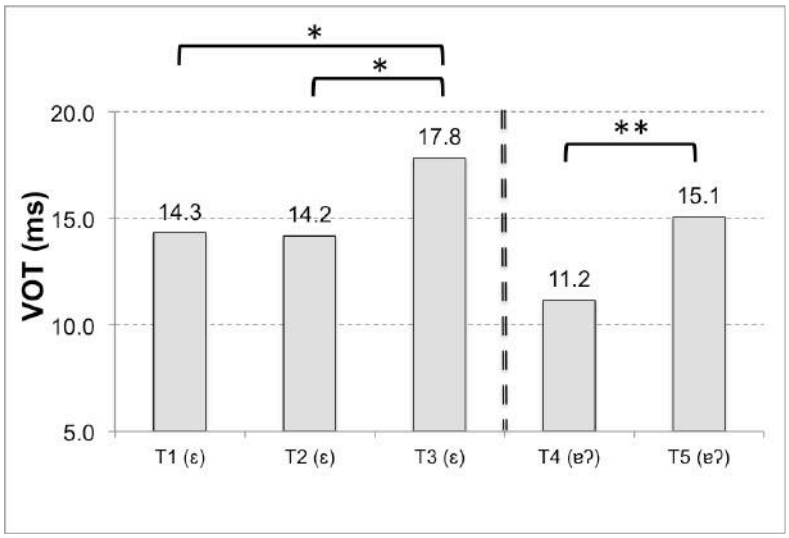


Figure 23. Average stop onset VOTs as a function of tone in monosyllables. Significance levels: * for $p<.01$; ** for $p<.001$.

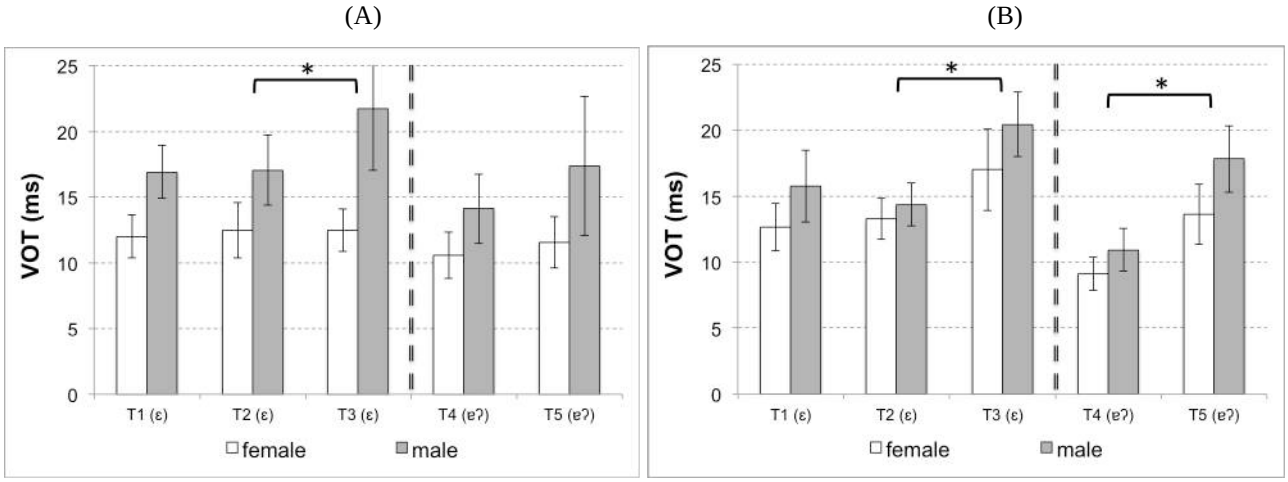


Figure 24. Average stop onset VOTs as a function of tone in monosyllables. (A) young speakers; (B) elderly speakers. Significance level: * for $p<.05$.

We ran two by-subject ANOVAs separately for the unchecked (T1-3) and checked (T4-5) syllable data. In both analyses, *VOT* was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Place of articulation* (labial vs. dental in both analyses) and *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) were within-subject factors.

For unchecked syllables, *Place* had no effect overall, $F(1,18)<1$, n.s., but *Tone* was significant, $F(2,36)=9.8$, $p<.0005$: *VOT* was the longest for *yang* stops (T3 vs. T1 stops: 17.8>14.3 ms, $F(1,18)=9.6$, $p<.01$; T3 vs. T2 stops: 17.8>14.2 ms, $F(1,18)=13.8$, $p<.005$). The *Tone* × *Place* interaction was significant, $F(2,36)=4.4$, $p<.05$, reflecting larger T3 vs. T1-2 differentials for dental (Δ=4.4 ms) than for labial stops (Δ=2.8 ms). *Gender* had a significant effect on *VOT*, $F(1,18)=11.2$, $p<.005$: male speakers had longer *VOTs* than female speakers (17.6>13.3 ms). *Age* had no effect, $F(1,18)<1$, and did not interact with *Gender*, $F(1,18)=2.1$, $p=.17$, n.s. Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(2,36)=1.7$, $p=.20$; *Tone* × *Age*: $F(2,36)<1$).

For checked syllables, *Tone* was also significant overall, $F(1,18)=19.3$, $p<.0005$, with longer *VOTs* for *yang* than *yin* stops (T5 vs. T4, 15.1>11.2 ms). *Place* had no effect, $F(1,18)=1.09$, $p=.31$, and did not interact with *Tone*, $F(1,18)<1$. *Gender* was, again, significant, $F(1,18)=6.0$, $p<.05$, with longer *VOTs* for male than female speakers (15.0>11.2 ms). *Age* was not significant, $F(1,18)<1$, n.s., and did not interact with *Gender*, $F(1,18)<1$. Again, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(2,36)=1.7$, $p=.20$; *Tone* × *Age*: $F(2,36)<1$).

- **Voicing fricative onsets**

The voicing of fricative onsets in monosyllables were measured by their voicing ratio (v-ratio), comprised between zero and one: Higher v-ratios indicate higher voicing degree. Figure 25 shows the average v-ratio of fricative onsets according to tone for the monosyllabic context, pooled across speaker groups and onset types; Figure 26 shows (A) the young and (B) the elderly speakers' data, separately for male and female speakers. The detailed data are reported in Appendix 2.

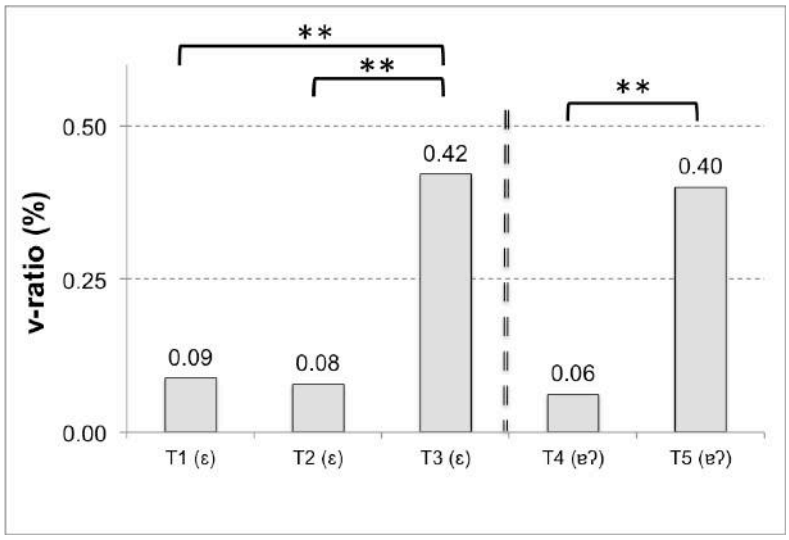


Figure 25. Average v-ratios of fricative onsets in monosyllables, according to tone. Significance levels: * for $p<.01$; ** for $p<.001$.

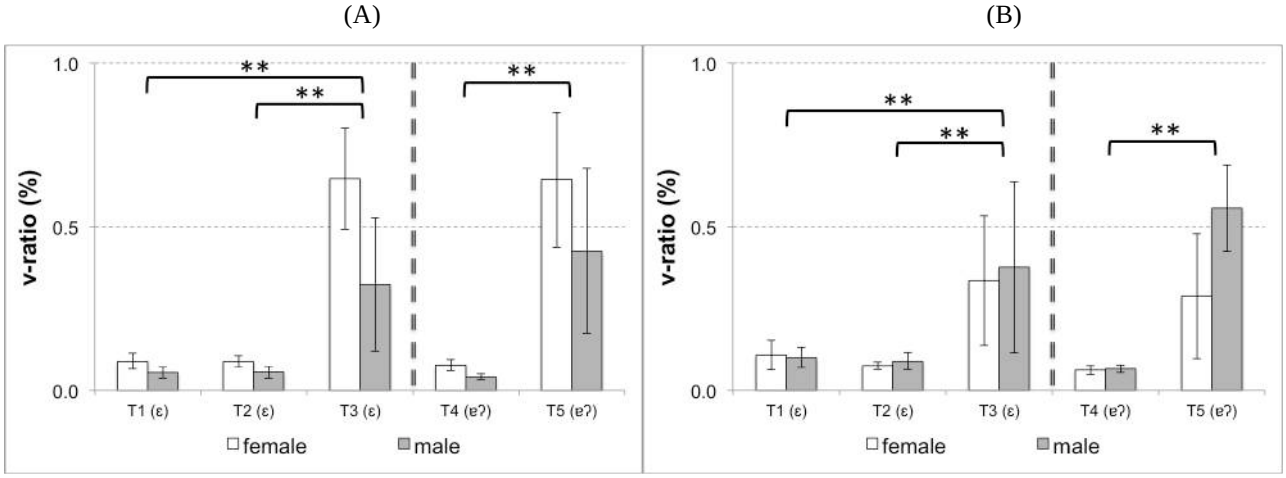


Figure 26. Average v-ratios of fricative onsets in monosyllables, according to tone, age, and gender. (A) young speakers; (B) elderly speakers. Significance level: ** for $p<.01$.

Fricative onsets do not behave exactly like stop onsets concerning their voiceless phonetic realization in word-initial position, as described in the literature. The *yin* fricative onsets are voiceless, with v-ratio close to zero. However, the phonetic realization of the *yang* fricative onsets is more variable, with a considerably higher average v-ratio, suggesting these onsets are partly voiced phonetically. Table 17 shows the number of occurrences of "substantially voiced" fricatives (with a v-ratio higher than 0.5) for each gender and age group. As can be seen, young speakers produce more substantially voiced fricatives than elderly speakers, and female speakers more than male speakers. Labial fricatives tend to be more often voiced than dental fricatives.

Table 17. Number of occurrence of "substantially voiced" fricative onsets (with a v-ratio higher than 0.5) in monosyllables, according to subject group and syllable type.

Syllable	vɛ	vaʔ	zɛ	zaʔ	Total
young female	10/11	10/11	6/11	5/11	**31/44 (70.5%)**
young male	4/12	8/12	3/12	2/12	**17/48 (35.4%)**
elderly female	7/10	6/10	0/10	0/10	**13/40 (32.5%)**
elderly male	3/7	1/7	0/7	0/7	**4/28 (14.3%)**

According to our results, the v-ratios of fricative onsets in monosyllables are overall higher for *yang* than for *yin* syllables. They do not differ between T1 and T2 syllables. To substantiate these observations, we ran two by-subject ANOVAs separately for the unchecked (T1-3) and checked (T4-5) syllable data. In both analyses, *v-ratio* was the dependent variable and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Place of articulation* (labial vs. dental in both analyses) and *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) were within-subject factors.

For unchecked syllables, *Tone* was significant overall, $F(2,36)=64.0$, $p<.0001$: *v-ratio* was significantly much higher for *yang* than *yin* onsets (T3 vs. T1: 0.42>0.09, $F(1,18)=62.3$, $p<.0001$; T3 vs. T2: 0.42>0.08, $F(1,18)=68.8$, $p<.0001$). *Place* was also significant, $F(1,18)=23.0$, $p<.0005$: *v-ratio* was higher for labial than dental fricatives (0.27>0.12). The *Tone* × *Place* interaction was significant, $F(2,36)=21.9$, $p<.0001$,

reflecting a larger T3 vs. T1-2 differential for labial (Δ=0.52) than dental (Δ=0.16) fricatives. Neither *Gender* nor *Age* was significant (*Gender*: $F(1,18)=2.9$, $p=.11$; Age: $F(1,18)<1$). The *Tone* × *Age* interaction was marginal, $F(2,36)=3.2$, $p=.052$, reflecting a larger T3 vs. T1-2 differential for young (Δ=0.41) than elderly (Δ=0.26) speakers. The *Tone* × *Gender* interaction was significant within young speakers, $F(2,20)=6.1$, $p<.01$, reflecting a larger T3 vs. T1-2 differential for female (Δ=0.56) than male (Δ=0.27) speakers, but *Tone* did not interact with *Gender* within elderly speakers, $F(2,16)<1$.

For checked syllables, *Tone* also had a significant effect overall on *v-ratio* $F(1,18)=52.1$, $p<.0001$: *v*-ratio was higher for *yang* than *yin* onsets (0.40>0.06). *Place* was also significant, $F(1,18)=36.7$, $p<.0001$: *v-ratio* was, again, higher for labial than dental fricatives (0.34>0.12). The *Tone* × *Place* interaction was highly significant, $F(1,18)=32.5$, $p<.0001$, reflecting a larger T5 vs. T4 differential for labial (Δ=0.53) than dental (Δ=0.13) onsets. *Age* had a significant effect on *v-ratio*, $F(1,18)=7.6$, $p<.05$: *v-ratio* was higher for young than elderly speakers overall (0.30>0.16). *Gender* was not significant, $F(1,18)=2.1$, $p=.17$, and did not interact with *Age*, $F(1,18)=1.3$, $p=.28$. *Tone* did not interact with *Gender*, $F(1,18)=1.5$, $p=.23$, but the *Tone* × *Age* interaction was significant, $F(1,18)=9.0$, $p<.01$, reflecting a larger T5 vs. T4 differential for young (Δ=0.49) than elderly (Δ=0.19) speakers.

4.1.6.2.2 First syllable in disyllables (S1)

The first part of this section concerns stop onsets, and the second part fricative onsets in the first syllable of disyllabic words.

- **Voicing of stop onsets**

Same as for the monosyllabic context, word-initial stops in disyllables (S1 context) were produced without pre-voicing, with rare exceptions, regardless of phonological voicing. Still, the same young female speaker who produced pre-voicing in the monosyllabic context also produced one occurrence of pre-voiced /d/ in the syllable /da?/, with a VOT of –119 ms. This value is excluded from the descriptions and statistical analyses.

Figure 27 shows the average VOTs of stop onsets according to the underlying tone of the first syllable (S1), pooled across subject groups and onset types; Figure 28 separately shows (A) the young speakers' and (B) the elderly speakers data. The detailed data are reported in Appendix 2.

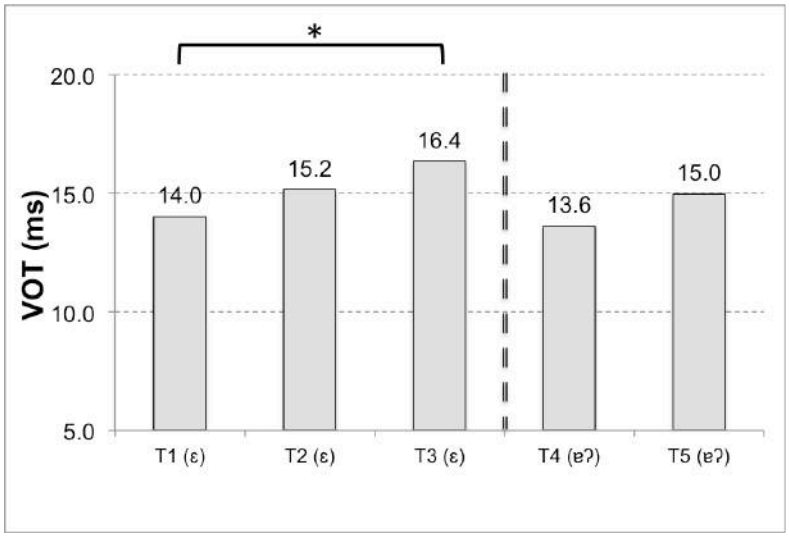


Figure 27. Average VOTs of stop onsets in the S1 context, according to underlying tone. Significance level: * for p<.01.

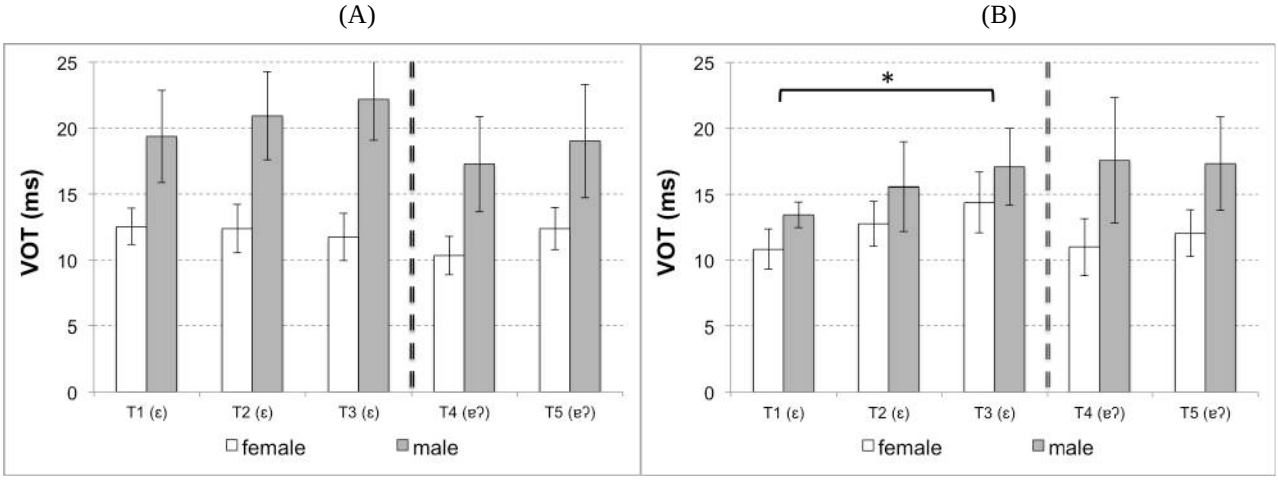


Figure 28. Average VOTs of stop onset in the S1 context, according to underlying tone, age, and gender. (A) young speakers; (B) elderly speakers. Significance levels: * for p<.05.

Numerically, the underlying tone of the syllable did not affect VOT in S1 stop onsets as much as in monosyllable stop onsets. VOTs were significantly longer for T3 than for T1 syllables and for T3 than T2 (although the difference was tiny) but not for T5 than T4 syllables. We ran two by-subject ANOVAs separately for the unchecked

(T1-3) and checked (T4-5) syllable data. In both analyses, *VOT* was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Place of articulation* (labial vs. dental in both analyses) and *Tone* (unchecked syllables: T1, T2, T3; checked syllables: T4 vs. T5) were within-subject factors.

For unchecked syllables, *Tone* was significant overall, $F(2,36)=6.4$, $p<.005$: *VOT* was longer for T3 than T1 syllables, (16.4>14.0, $F(1,18)=10.1$, $p<.01$), and for T3 than T2 syllables (16.4>15.2, $F(1,18)=5.4$, $p<.05$), but did not differ between T2 and T1, $F(1,18)=2.6$, $p=.12$. *Place* was not significant, $F(1,18)<1$, and did not interact with *Tone*, $F(2,36)<1$. *Gender* was significant, $F(1,18)=18.3$, $p=.0005$, but *Age* was not, $F(1,18)=3.51$, $p=.077$. The *Gender* × *Age* interaction was marginal, $F(1,18)=4.1$, $p=.058$: *VOT* was significantly higher for males than for females only within young speakers (20.8>12.1 ms). Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(2,36)<1$; *Tone* × *Age*: $F(2,36)=1.73$, $p=.19$.

For checked syllables, *Tone* was not significant, $F(1,18)=3.4$, $p=.081$; *Place* was not significant, $F(1,18)=2.7$, $p=.12$. *Gender* was significant, $F(1,18)=12.7$, $p<.005$: male speakers had longer VOTs than female speakers (17.2>11.4 ms). *Age* was not significant, $F(1,18)=.18$, $p=.68$, and did not interact with *Gender*, $F(1,18)=0.079$, $p=.78$.

- **Voicing of fricative onsets**

Fricative onsets in the S1 context shared the same voicing pattern as those in the monosyllabic context: *yin* onsets were voiceless, and the v-ratio of *yang* onsets was much higher and more variable for *yang* than *yin* onsets.

Figure 29 shows the average v-ratio of fricative onsets according to underlying tone for the first syllable (S1), pooled across speaker groups and onset types; Figure 30 separately shows the v-ratios for the four subject groups. The detailed data are shown in Appendix 2. As in the monosyllabic context, v-ratios were higher for *yang* than *yin* fricatives, suggesting that *yang* fricative onsets are partially voiced phonetically.

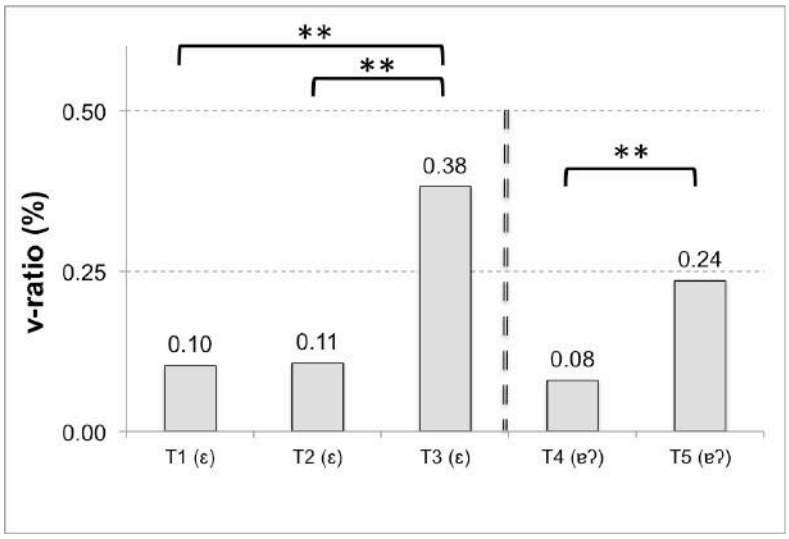


Figure 29. Average v-ratios of fricative onsets in the S1 context, according to underlying tone. Significance levels: * for $p<.01$; ** for $p<.001$.

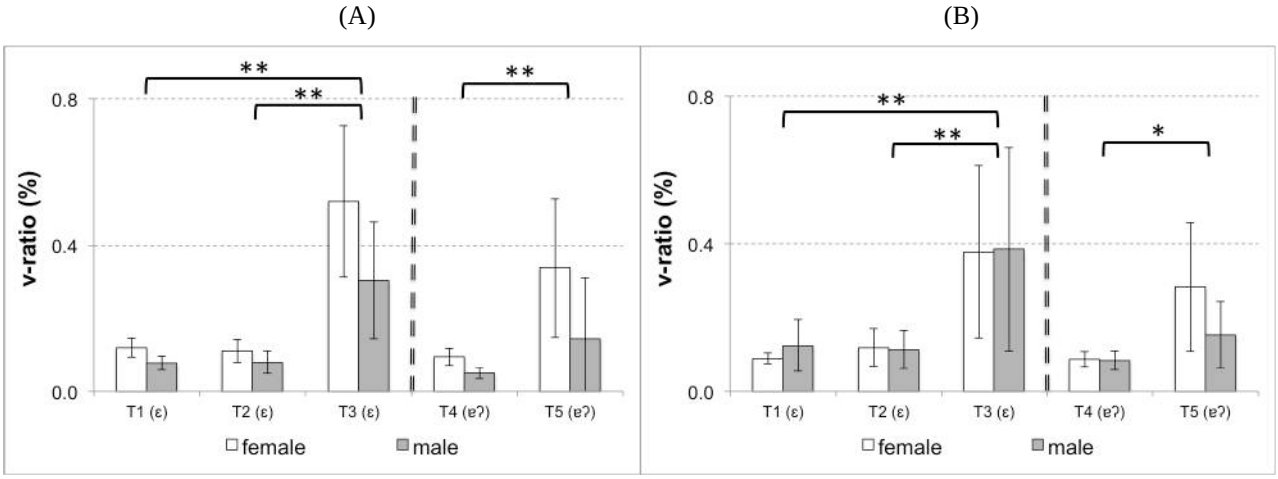


Figure 30. Average v-ratios of fricative onsets (S1 context), according to underlying tone, age, and gender. (A) young speakers; (B) elderly speakers. Significance levels: ** for $p<.05$.

Figure 29 shows the average v-ratio of fricative onsets according to underlying tone for the first syllable (S1), pooled across speaker groups and onset types; Figure 30 separately shows the v-ratios for the four speaker groups. The detailed data are shown in Appendix 2. As in the monosyllabic context, v-ratios were higher for *yang* than *yin* fricatives, suggesting that *yang* fricative onsets are partially voiced phonetically. To substantiate these observations, we ran two by-subject ANOVAs separately for the unchecked (T1-3) and checked (T4-5) syllable data. In both analyses, *v-ratio* was the dependent variable, and *Speaker* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors.

Place (labial vs. dental in both analyses) and *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) were within-subject factors.

For unchecked syllables, *Tone* was highly significant, $F(2,36)=43.4$, $p<.0001$: *v-ratio* was higher for *yang* than *yin* fricative onsets (T3 vs. T1: 0.38>0.10, $F(1,18)=47.0$, $p<.0001$; T3 vs. T2: 0.38>0.11, $F(1,18)=42.8$, $p<.0001$). *Place* was highly significant, $F(1,18)=42.3$, $p<0001$: *v-ratio* was higher for labial than dental fricatives (0.24>0.07). The *Tone* × *Place* interaction was significant, $F(2,36)=31.8$, $p<.0001$, reflecting a larger T3 vs. T1-2 differential for labial ($\Delta=0.40$) than dental ($\Delta=0.05$) fricatives. Neither *Gender* nor *Age* was significant (Gender: $F(1,18)=3.1$, $p=.10$; Age: $F(1,18)<1$). *Tone* did not interact with *Age*, $F(2,36)<1$. As for the monosyllabic context, the *Tone* × *Gender* interaction was significant only within young speakers, $F(2,20)=5.2$, $p<.05$, reflecting a larger T3 vs. T1-2 differential for females ($\Delta=0.40$) than males ($\Delta=0.16$). This interaction was not significant for elderly speakers, $F(2,16)<1$.

For checked syllables, *Tone* was highly significant, $F(1,18)=21.8$, $p<.0001$: *v-ratio* was higher for T5 than T4 fricative onsets (0.24>0.08). *Place* was highly significant, $F(1,18)=30.9$, $p<.0001$: *v-ratio* was again higher for labial than dental fricatives (0.26>0.07). The *Tone* × *Place* interaction was significant $F(1,18)=20.0$, $p<.0005$, reflecting a significant T5 vs. T4 differential for labial fricatives ($\Delta=0.32$) but no significant difference for dental fricatives ($\Delta=0.01$). *Age* was significant, $F(1,18)=8.2$, $p<.05$: *v-ratio* was higher overall for female than male speakers (0.20>0.11). *Gender* was not significant, $F(1,18)<1$, and did not interact with *Age*, $F(1,18)<1$. Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(1,18)=3.9$, $p=.063$; *Tone* × *Age*: $F(2,36)<1$).

4.1.6.2.3 Second syllable in disyllables (S2)

Voicing for the onsets (stops and fricatives) of the second syllable of disyllables was measured by the voicing ratio (v-ratio). As described in the literature, the voicing contrast is maintained phonetically in word-medial position. The v-ratio was higher for *yang* (T3 and T5) than *yin* (T1, T2 and T4) onsets, whether stops or fricatives. Figure 31 shows the average v-ratio of stop and fricative onsets according to the

underlying tone of the second syllable (S2), pooled across speaker groups, onset types, and the first syllable's tones; Figure 32 separately shows the same data for the four speaker groups. The detailed data are reported in Appendix 2. Since the preceding tone (T1 or T2) did not interact with the main factor Tone, the data are always pooled across these two preceding tone contexts.

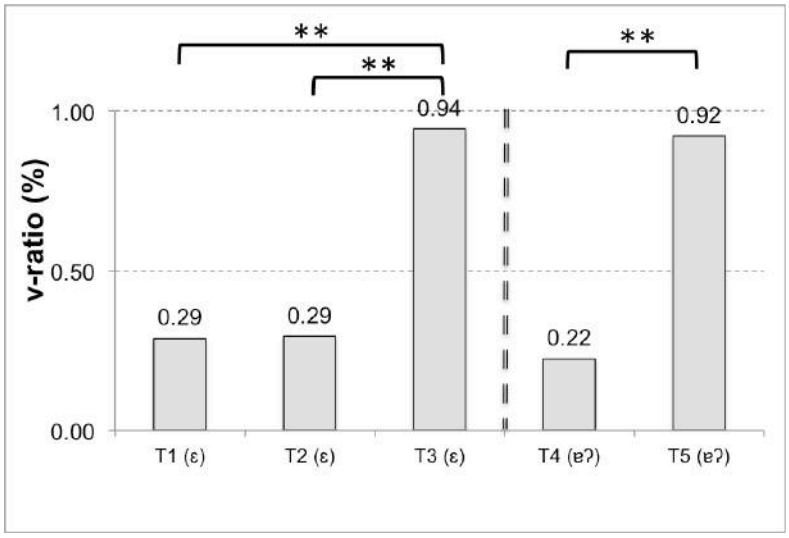


Figure 31. Average v-ratios of S2 obstruent onsets, according to underlying tone. Significance level: ** for $p<.001$.

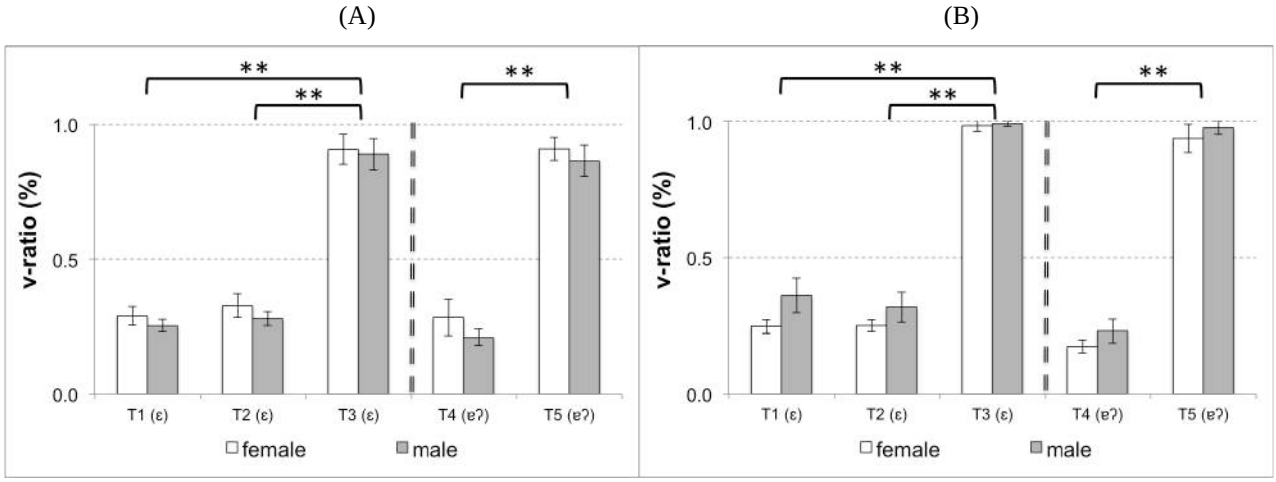


Figure 32. Average v-ratios of S2 obstruent onsets, according to underlying tone. (A) young speakers; (B) elderly speakers. Significance level: ** for $p<.01$.

To substantiate these observations, we ran two by-subject ANOVAs separately for the unchecked (T1-3) and checked (T4-5) syllable data. In both analyses, *v-ratio* was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male)

and *Age* group (young vs. elderly) were between-subject factors. *Manner* of articulation (stop vs. fricative in both analyses), *Place* of articulation (labial vs. dental in both), and *Tone* (unchecked syllables: T1, T2, T3; checked syllables: T4 vs. T5) were within-subject factors.

For unchecked syllables, *Tone* was highly significant, $F(2,36)=611.7$, $p<.0001$: *v-ratio* was higher for *yang* than *yin* onsets (T3 vs. T1: 0.94>0.29, $F(1,18)=680.8$, $p<.0001$; T3 vs. T2: 0.94>0.29, $F(1,18)=617.3$, $p<.0001$). *Place* was significant, $F(1,18)=8.6$, $p<.01$: *v-ratio* was higher for labial than for dental onsets (0.45>0.41). *Manner* was not significant, $F(1,18)=2.8$, $p=.11$. The *Tone* × *Place* interaction and the *Tone* × *Manner* interaction were both significant (*Tone* × *Place*: $F(2,36)=7.7$, $p<.005$; *Tone* × *Manner*: $F(2,36)=24.0$, $p<.0001$), with a larger T3 vs. T1-2 differential for dental ($\Delta=0.43$) than labial ($\Delta=0.38$) fricatives, and for fricatives ($\Delta=0.70$) than stops ($\Delta=0.59$). Neither *Gender* nor *Age* was significant (*Gender*: $F(1,18)=1.8$, $p=.20$; *Age*: $F(1,18)<1$, n.s.). Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(2,36)<1$; *Tone* × *Age*: $F(2,36)<1$).

For checked syllables, *Tone* was significant: *v-ratio* was significantly higher for *yang* than *yin* onsets (T5 vs. T4: 0.92>0.22, $F(1,18)=1347.4$, $p<.0001$). The *Tone* × *Manner* interaction was significant, $F(1,18)=22.4$, $p<.0005$, reflecting again a larger T5 vs. T4 differential for fricatives ($\Delta=0.76$) than stops ($\Delta=0.62$). The *Tone* × *Place* interaction was not significant, $F(1,18)<1$. Neither *Gender* nor *Age* was significant (*Gender*: $F(1,18)=1.8$, $p=.20$; *Age*: $F(1,18)<1$). Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(1,18)<1$; *Tone* × *Age*: $F(1,18)=3.4$, $p=.80$).

Our data also showed a kind of phonetic realization of word-medial voiced stops, which is rarely mentioned in the literature. Occasionally, the stops of T3 and T5 syllables were spirantized in word-medial position. As illustrated in Figure 33, word-medial /b/ was pronounced by one speaker (a) as a pre-voiced stop, as shown by the energy decrease during the closure portion, and by another speaker (b) as a [ɸ]- or [β]-like spirantized fricative, as suggested by the formant structure appearing between the two vowels. Moreover, intra-speaker variability was also observed concerning these two realizations of word-medial voiced stops.

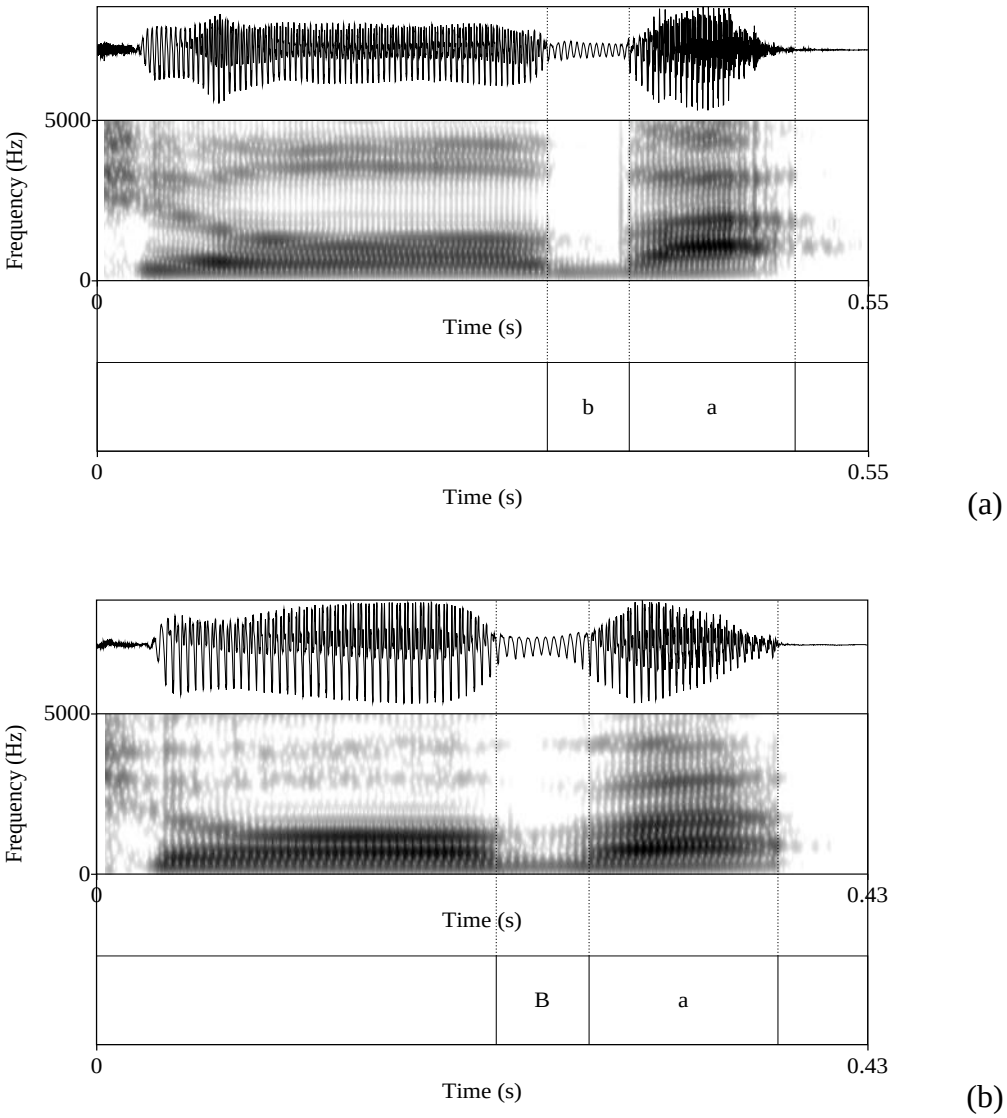


Figure 33. Speech waveforms and spectrograms of an intervocalic stop realized (a) as a stop by a 72 year-old female speaker, and (b) as a fricative-like intervocalic stop by a 69 year-old female speaker. (CD: 4.1.6.2.3_fig34)

Table 18 shows the number of occurrences of stop spirantization out of the total number of productions for each subject group. A noticeable point is that elderly female speakers were the most prone to produce spirantized stops, followed by elderly male speakers, then by young female speakers, and lastly by young male speakers. These spirantized realizations were not included in voicing or duration analyses.

Table 18. Number of occurrences of stops' spirantization according to syllable type, preceding tone, and speaker group.

syllable	**bɛ**		**baʔ**		**dɛ**		**daʔ**		**Total**
S1 tone	T1-	T2-	T1-	T2-	T1-	T2-	T1-	T2-	
young female	0/11	1/11	0/11	0/11	0/11	1/11	0/11	0/11	**2/88**
young male	1/12	2/12	0/12	1/12	1/12	1/12	0/12	1/12	**7/96**
elderly female	6/10	0/10	2/10	2/10	2/10	2/10	0/10	0/10	**14/80**
elderly male	2/7	1/7	2/7	1/7	1/7	0/7	0/7	0/7	**7/56**

4.1.6.2.4 Discussion

In this section, we investigated the phonetic voicing of phonologically voiced onsets in word-initial and word-medial context. According to what is described in the literature, the voicing contrast is neutralized word-initially and is realized word-medially. Our results are in partial conformity with this description: voicing contrast is neutralized only for stops in word-initial position, but not for fricatives.

Word-initially, *yang* stops are realized predominantly without pre-voicing, with extremely rare exceptions, but *yin* and *yang* stops nevertheless differ in terms of VOT, and most clearly in the monosyllabic context. *Yang* stops have longer VOT than *yin* stops. Many studies on Shanghai stops insisted on the positive VOT values of phonologically voiced stops but attach less importance to their VOT values *relative* to those of voiceless stops (e.g., Ren, 1988). However, longer VOT values for *yang* stops are also found in 苏州 Suzhou (Iwata, Hirose, Niimi & Horibuchi, 1991) and in 镇海 Zhenhai (Rose, 1982b) (but see Shi, 1983, for contradictory results). In these Wu dialects, *yang* stops are all reported to be "breathy" or "whispery." It would thus seem that breathy phonation is correlated with longer VOTs. Similar evidence can also be found in other languages with distinctive phonation differences: breathy/slack/lax stops have longer VOTs than their modal phonation counterparts (Maddieson & Ladefoged, 1985, for Kawa and Jingpho; Andruski & Ratliff, 2000, for Green Mong). This said, the VOT difference we found between *yin* and *yang* word-initial stops is less pronounced in S1 context than in monosyllabic context.

Yang fricatives in word-initial position are not as *qingyin* 清音 'clear/voiceless' as *yang* stops. They are quite often, although not systematically, realized as voiced. Labial fricatives are more prone to phonetic voicing than dental ones. Our data also show large inter- and intra-speaker variability in the realization of fricatives' voicing. Moreover, young speakers tend to realize *yang* fricatives as voiced much more often than elderly speakers, and female speakers more often than male speakers.

It should be noted that in Shanghai Chinese, the /v/ evolved mainly from two forms of Middle Chinese[13]: *bɦ (which became *fɦ later), as in the syllable 饭 /vɛ/ (T3) (< MC *bjonH)[14]; and *m/w, as in the syllable 万 /vɛ/ (T3) (< MC *mjonH). Probably the /v/ originating from *m/w would tend to be produced as voiced. In the two /v/-initial morphemes we used, the origin of the /v/ onset was the *bɦ form of Middle Chinese. So these /v/s would normally be comparable to the other obstruent onsets. But this particular origin might still have some consequences on speakers' production (for more detail, see General Discussion, p. 266).

Word-medially, *yang* onsets are phonetically voiced as opposed to *yin* onsets, which are phonetically voiceless, as described in the literature. One new finding of our study is the spirantized realization of word-medial voiced stops, in free variation with the full stop realization in this position. Cross-age variation is also found: spirantized voiced stops occur more often among elderly than young speakers.

---

[13] I thank Laurent Sagart for this comment at the 27èmes Journées de Linguistique d'Asie Orientale.

[14] The Middle Chinese (MC) reconstructed forms come from Baxter and Sagart (2011): the final H stands for the MC rising tone (上声).

4.1.6.3 Phonation type

Several acoustic measurements, though not all possible measurements, were used as indicators of phonation difference between *yin* and *yang* syllables: for the consonant onset, HNR (harmonics-to-noise ratio); for the vowel part: H1–H2 (amplitude difference between the first and second harmonics), H1–A1 (amplitude difference between the first harmonic and the first formant), H1–A2 (amplitude difference between the first harmonic and the second formant), CPP (Cepstral Peak Prominence) and F1 (value of the first formant).

4.1.6.3.1 Word-initial onsets (stops and nasals)

We measured HNR in the word initial nasals as well as in the release part of the word-initial stops, to evaluate the harmonic structure relative to the noise of the spectrum. The source of the noise is presumed to be the breathiness in our data; hence, lower HNR values indicate breathier voice, other things being equal. We will not report HNR values on fricatives, because *yang* fricatives that were realized as voiced (see §4.1.6.2) presented a strong harmonic structure, thus yielding higher HNR values than voiceless fricatives, even when they were breathier. For the same reason, we excluded four occurrences of word-initial phonetically voiced [d] (see §4.1.6.2) from HNR analyses. Similar patterns were found in monosyllables and word-initial syllables (S1 syllables).

Figure 34 shows the average HNR values of stop onsets according to syllable tone in the monosyllabic and S1 contexts, pooled across speaker groups and onset types; Figure 35 separately shows (A) the young speakers' and (B) the elderly speakers' data for stop onsets. The detailed data are shown in Appendix 2. Figure 36 shows the average HNR values of nasal stops according to syllable tone in the monosyllabic context, pooled across speaker groups and places of articulation (the S1 context was not examined and nasal onsets do no co-occur with T4). Figure 37 separately shows (A) the young speakers' and (B) the elderly speakers' data for nasal onsets.

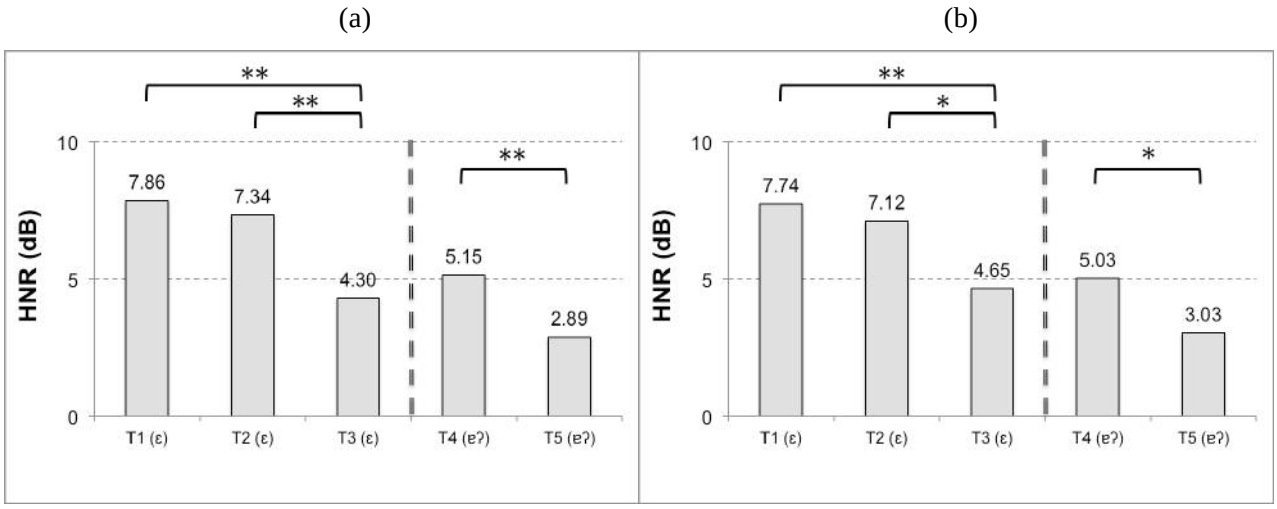


Figure 34. Average HNR (dB) of stop onsets according to tone in (a) monosyllables and (b) S1 syllables. Significance levels: * for $p<.01$; ** for $p<.001$.

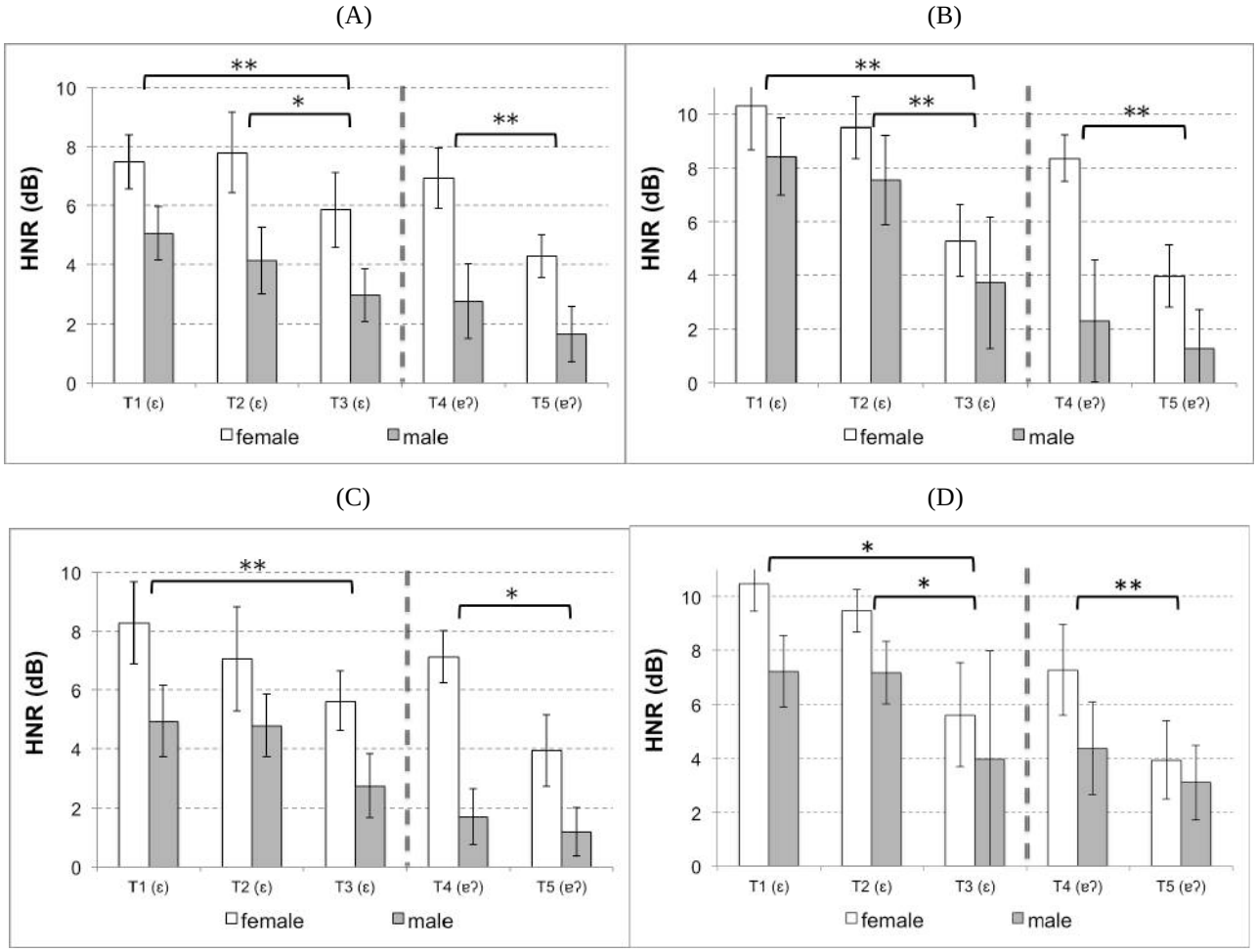


Figure 35. HNR of stop onsets according to tone in (A-B) monosyllables and (C-D) S1 syllables. (A) and (C): young speakers; (B) and (D): elderly speakers. Significance levels: * for $p<.05$; ** for $p<.01$.

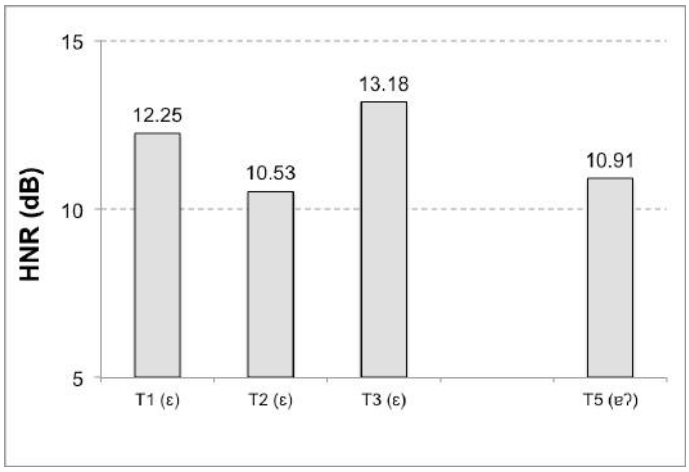


Figure 36. Average HNR (dB) of nasal onsets according to tone in monosyllables.

(A) (B)

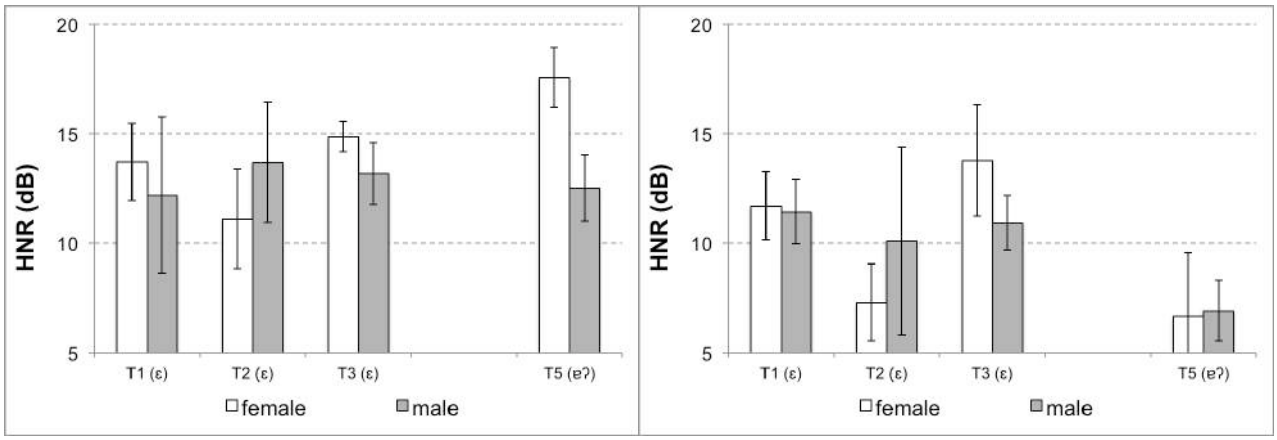


Figure 37. HNR of nasal onsets according to tone in monosyllables. (A) young speakers; (B) elderly speakers.

According to our results, in both the monosyllabic and S1 contexts, the HNRs of syllable-initial stops are lower for T3 (*yang*) than T1 or T2 (*yin*) syllables, and T1 and T2 do not differ. HNRs are likewise lower for T5 (*yang*) than T4 (*yin*). Nonetheless, this is not the case for nasal onsets. Globally, HNR is not lower for T3 or T5 than the other tones, and is even numerically higher for T3 than the others. This indicates that *yang* nasal onsets are not breathier than *yin* onsets.

Statistical analyses were run only on stop onsets. We ran four by-subject ANOVAs separately for the unchecked (T1-3) and checked (T4-5) syllable data, and for the monosyllabic and S1 contexts. In all four analyses, *HNR* was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Place of articulation* (labial vs. dental in both analyses) and *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) were within-subject factors.

For unchecked monosyllables, *Tone* was highly significant overall, $F(2,36)=28.7$, $p<.0001$: *HNR* was significantly lower for *yang* than *yin* tones (T3 vs. T1: 4.3<7.9, $F(1,18)=44.8$, $p<.0001$; T3 vs. T2: 4.3<7.3, $F(1,18)=27.2$, $p=.0001$) but did not differ between T1 and T2, $F(1,18)<1$. *Place* was not significant, $F(1,18)<1$, and did not interact with *Tone,* $F(1,18)<1$. Both *Gender* and *Age* were significant (*Gender*: $F(1,18)=15.3$, $p=.001$; *Age*: $F(1,18)=9.1$, $p<.01$): *HNR* was higher for female than male speakers (7.7>5.3), and higher for elderly than young speakers (7.4>5.6). Neither the *Gender* × *Age* nor the *Tone* × *Gender* interaction was significant, *Fs*<1. The *Tone* × *Age* interaction was significant, $F(2,36)=6.1$, $p=0.005$, reflecting a higher T1-2 vs. T3 differential for elderly ($\Delta=4.9$, $F(2,16)=22.0$, $p<.0001$) than young speakers ($\Delta=1.7$, $F(2,20)=6.8$, $p<.01$).

For checked monosyllables, *Tone* was highly significant, $F(1,18)=47.1$, $p<.0001$: *HNR* was significantly lower for *yang* than *yin* onsets (T5 vs. T4: 2.9<5.2). *Place* was again not significant, $F(1,18)=3.2$, $p=.09$, and did not interact with *Tone,* $F(1,18)=2.8$, $p=.11$. *Gender* was significant, $F(1,18)=41.0$, $p<.0001$: *HNR* was higher for female than male speakers (5.9>2.5). *Age* was not, $F(1,18)<1$, and did not interact with *Gender,* $F(1,18)<1$. The *Tone* × *Gender* interaction was significant, $F(1,18)=11.3$, $p<.005$, reflecting a higher T4 vs. T5 differential for female ($\Delta=3.5$, $F(1,10)=41.9$, $p=.0001$) than male ($\Delta=1.0$, $F(1,8)=6.6$, $p<.05$) speakers. The *Tone* × *Age* interaction was marginal, $F(1,18)=4.2$, $p=.055$, reflecting a higher T4 vs. T5 differential for elderly ($\Delta=2.6$, $F(1,10)=21.1$, $p<.005$) than young ($\Delta=1.9$, $F(1,8)=18.7$, $p<.005$) speakers.

For unchecked S1 syllables, *Tone* was highly significant, $F(1,18)=15.4$, $p<.0001$: *HNR* was significantly lower for *yang* than *yin* onsets (T3 vs. T1: 4.7<7.7, $F(1,18)=26.3$, $p=.0001$; T3 vs. T2: 4.7<7.1, $F(1,18)=12.1$, $p<.005$) but did not differ between T1 and T2, $F(1,18)<1$. *Place* was not significant, $F(1,18)=1.7$, $p=.21$, and did not interact with *Tone,* $F(1,18)<1$. Both *Gender* and *Age* were significant (*Gender*: $F(1,18)=25.6$, $p=.0001$; *Age*: $F(1,18)=13.7$, $p<.005$): *HNR* was higher for female than male speakers (7.7>5.3), and higher for elderly than young speakers (7.4>5.6), with the same mean values as for monosyllables. Neither the *Tone* × *Gender* nor the *Tone* × *Age* interaction was significant (*Tone* × *Gender*: $F(2,36)<1$; *Tone* × *Age*: $F(2,36)=1.2$, $p=.32$).

For checked S1 syllables, *Tone* was highly significant, $F(1,18)=23.0$, $p=.0001$: *HNR* was significantly lower for *yang* than *yin* onsets (T5 vs. T4: 3.0<5.0). *Place* was not significant, $F(1,18)=1.7$, $p=.21$, and did not interact with *Tone,* $F(1,18)<1$. *Gender* was significant, $F(1,18)=25.2$, $p=.0001$: *HNR* was higher for female than male speakers (5.6>2.5). *Age* was marginal, $F(1,18)=4.1$, $p=.058$: *HNR* tended to be higher for elderly than young speakers (4.5>3.5). The *Tone* × *Gender* interaction was significant, $F(1,18)=7.3$, $p<.05$: the T4 vs. T5 difference was significant only for female speakers ($\Delta=3.2$, $F(1,10)=26.5$, $p<.0005$), not for male speakers ($\Delta=0.8$, $F(1,8)=1.9$, $p<.20$). The *Tone* × *Age* interaction was not significant, $F(1,18)<1$.

4.1.6.3.2 Vocalic part of monosyllables

This section is composed of three parts, corresponding to measurements of three indicators of phonation types: spectural tilt, cepstral peak prominence (CPP) and the value of the first formant (F1).

- **Spectral tilt measures**

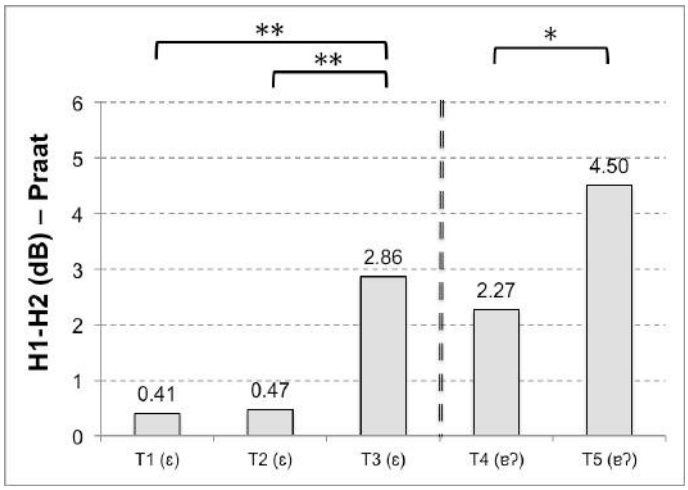
We used H1–H2, H1–A1, and H1–A2 as possible estimates of spectral tilt. For the three measures, higher values indicate steeper negative spectral tilt, that is, breathier phonation.

First, we compared the uncorrected H1–H2 values measured with Praat (on a 30 ms window at five time points in the rime), uncorrected H1–H2 values measured with Voicesauce (averaged over five consecutive time intervals of the rime) (Shue, Keating, Vicenik & Yu, 2011), and corrected H1–H2 values (noted H1*–H2*) measured with Voicesauce (averaged over five time intervals of the rime). We also compared the uncorrected H1–A1 and H1–A2 values with the corrected H1–A1 and H1–A2 values (noted H1*–A1*, H1*–A2*), both measured with Voicesauce (averaged over five consecutive time intervals of the rime). The corrections implemented in Voicesauce correct H1, H2, and formants' amplitudes in attempting to remove the influence of the vocal tract resonances (which indeed may vary, mainly according to vowel), using an algorithm developed by Iseli & Alwan (2004). Figure 38, Figure 39 and Figure 40 plot respectively the results of H1–H2, H1–A1, and H1–A2 values in the vowel pooled

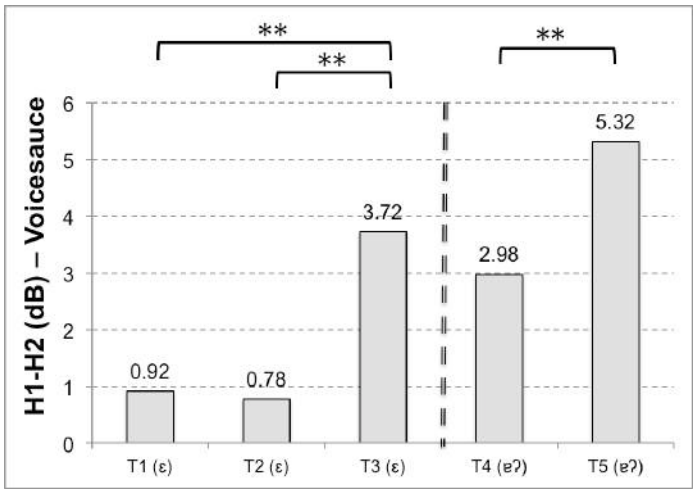
across time points or time intervals, speaker groups, and onset types, measured in the different ways mentioned above.

The two measurements without correction on harmonics (H1–H2) showed little difference. Both measurements showed higher H1–H2 values for vowels carrying *yang* tones than *yin* tones, indicating breathier voice on *yang* than *yin* vowels. On the other hand, the corrected H1–H2 values were much higher than the uncorrected ones, and the difference between *yin* and *yang* tones was highly reduced. Both the corrected H1–A1 and H1–A2 values are much higher than their uncorrected counterparts due to the removal of the influence of the vocal tract, but the *yin–yang* difference is still found.

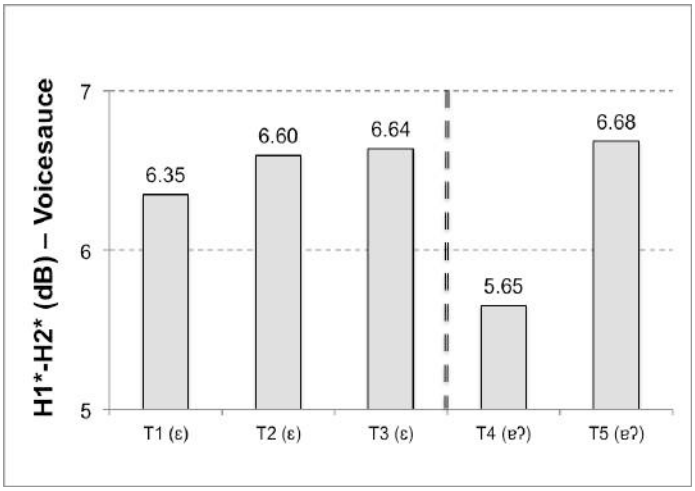
Since we use the same vowels for *yin/yang* comparison (/ɛ/ for unchecked and /aʔ/ for checked, with at most slightly different vowel quality between *yin* and *yang*), the normalization between vowels with corrected measurements was not useful. We thus choose to report the detailed results of the uncorrected H1–H2, H1–A1, and H1–A2 measurements obtained with Voicesauce for monosyllables.



(a)

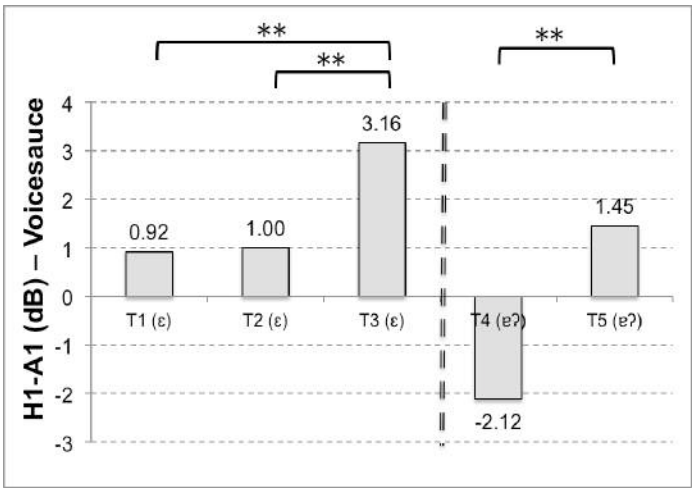


(b)

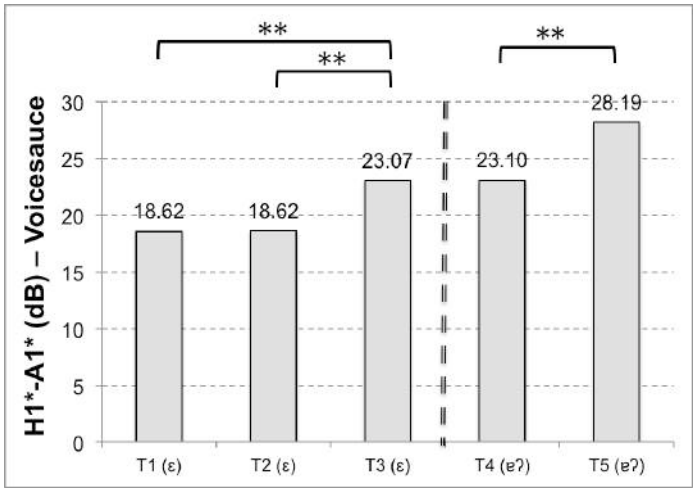


(c)

Figure 38. Average H1–H2 values in monosyllables' rimes, pooled across onset types, vowel time points or intervals, and speaker groups: (a) measured with Praat; (b) measured with Voicesauce on uncorrected H1 and H2 values; (c) measured with Voicesauce on corrected H1 and H2 values. Significance levels: * for $p<.01$; ** for $p<.001$.



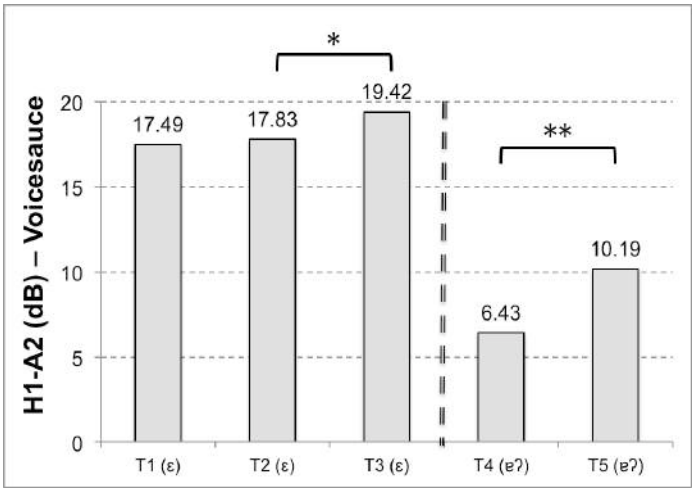
(a)



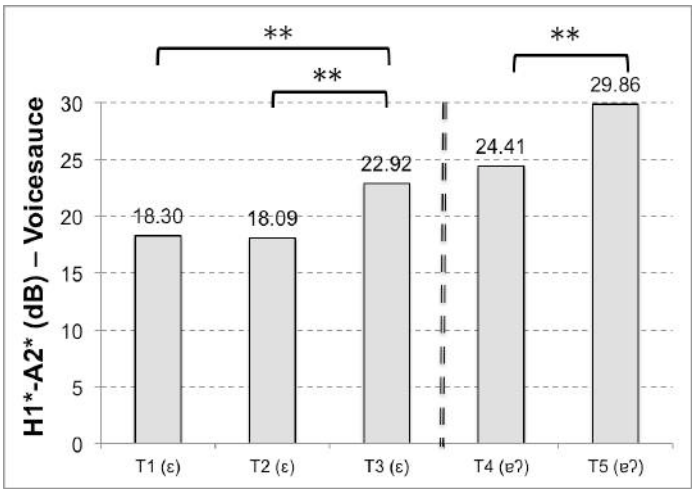
(b)

Figure 39. Average H1-A1 values in monosyllables' rimes, pooled across onset types, vowel time points or intervals, and speaker groups, measured with Voicesauce on (a) uncorrected H1 and A1 values and (b) corrected H1 and A1 values.

Significance levels: * for $p<.01$; ** for $p<.001$.



(a)



(b)

Figure 40. Average H1-A2 values in monosyllables' rimes, pooled across onset types, vowel time points or intervals, and speaker groups, measured with Voicesauce on (a) uncorrected H1 and A2 values and (b) corrected H1 and A2 values.

Significance levels: * for $p<.01$; ** for $p<.001$.

The following figures show H1–H2 (Figure 41 for unchecked and Figure 42 for checked syllables), H1–A1 (Figure 43 for unchecked and Figure 44 for checked syllables) and H1–A2 (Figure 45 for unchecked and Figure 46 for checked syllables) measured with Voicesauce at each of the five time intervals (*yang* tones in muddy color), for (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers.

A general trend for greater *yang* than *yin* values can be seen for all of the three measures, indicating steeper spectral tilt for vowels in *yang* syllables and thus breathier phonation. The difference is largest at the first and second time points and decreases along the vowel. At first view, the *yin–yang* difference is greater for elderly than for young speakers.

In order to substantiate these observations, we conducted by-subject ANOVAs for each of the three measures, separately for unchecked and checked syllables. The dependent variable of these ANOVAs were *H1–H2*, *H1–A1,* or *H1–A2*, and *Subject* was the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5), *Onset* (unchecked: Ø, /p/, /t/, /f/, /s/, /m, n/; checked: Ø, /p/, /t/, /f/, /s/) and *Time point* (P1, P2, P3, P4, P5) were within-subject factors. Because nasal onsets do not co-occur with all tones, we pooled together the /m/ and /n/ onsets into /m, n/ for unchecked syllables so that /m, n/ co-occurred with all three unchecked tones; because, there is no nasal onset syllable in checked tone T4, we excluded nasal onset syllables from analyses for checked syllables. In the following paragraphs, we only report the overall effect of *Tone*, as well as the interaction between *Tone* and *Gender* or *Tone* and *Age*. The detailed *yin–yang* difference data with statistical significance are reported from Table 19 to Table 24.

For H1–H2 in unchecked syllables, *Tone* was highly significant, $F(2,36)=36.6$, $p<.0001$. *H1–H2* was higher for T3 than T1 (3.91>0.82 dB, $F(1,18)=31.6$, $p<.0001$) or T2 (3.91>0.83 dB, $F(1,18)=62.2$, $p<.0001$). *Tone* did not interact with *Gender*, $F(2,36)=2.3$, $p=.11$, and the *Tone* × *Age* interaction was marginal, $F(2,36)=2.8$, $p=.074$, reflecting a greater T3 vs. T1-2 differential for elderly, ($\Delta=3.77$, $F(2,16)=77.4$, $p<.0001$) than young ($\Delta=2.40$, $F(2,20)=6.7$, $p<.01$) speakers.

For H1–H2 in checked syllables, *Tone* was highly significant, $F(1,18)=57.0$, $p<.0001$. *H1–H2* was higher for T5 than T4 (5.45>2.98 dB). Again *Tone* did not

interact with *Gender*, $F(1,18)=2.0$, $p=.18$, but the *Tone* × *Age* interaction was significant, $F(1,18)=5.0$, $p<.05$, again reflecting a greater T5 vs. T4 differential for elderly ($\Delta=3.14$, $F(1,8)=39.3$, $p<.0005$) than young ($\Delta=1.81$, $F(1,10)=18.3$, $p<.005$) speakers.

For H1–A1 in unchecked syllables, *Tone* was highly significant, $F(2,36)=19.4$, $p<.0001$. *H1–A1* was higher for T3 than T1 (3.16>0.92 dB, $F(1,18)=23.7$, $p=.0001$) and T2 (3.16>1.00 dB, $F(1,18)=23.9$, $p=.0001$). *Tone* did not interact with *Age*, $F(2,36)=2.7$, $p=.080$, but the *Tone* × *Gender* interaction was significant, $F(2,36)=5.3$, $p<.01$, reflecting a greater T3 vs. T1-2 differential for male ($\Delta=2.90$, $F(2,16)=26.8$, $p<.0001$) than female ($\Delta=1.73$, $F(2,20)=5.1$, $p<.05$) speakers.

For H1–A1 in checked syllables, *Tone* was highly significant, $F(1,18)=32.9$, $p<.0001$. *H1–A1* was higher for T5 than T4 (1.45>–2.12 dB). *Tone* did not interact with *Gender*, $F(1,18)<1$, but the *Tone* × *Age* interaction was significant, $F(1,18)=6.7$, $p<.05$, again reflecting a greater T5 vs. T4 differential for elderly ($\Delta=4.39$, $F(1,8)=32.2$, $p=.0005$) than young ($\Delta=2.88$, $F(1,10)=6.4$, $p<.05$) speakers.

For H1–A2 in unchecked syllables, *Tone* was significant, $F(1,18)=6.0$, $p<.01$. *H1-A2* was higher for T3 than T1 (19.42>17.49 dB, $F(1,18)=5.5$, $p<.05$) or T2 (19.42>17.83 dB, $F(1,18)=8.9$, $p<.01$). *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(2,36)=2.67$, $p=.08$; *Tone* × *Age*: $F(2,36)=1.8$, $p=.18$).

For H1–A2 in checked syllables, *Tone* was highly significant, $F(1,18)=40.2$, $p<.0001$. *H1–A2* was higher for T5 than T4 (10.19>6.43 dB). *Tone* did not interact with *Gender*, $F(1,18)<1$, but the *Tone* × *Age* interaction was significant, $F(1,18)=10.9$, $p<.005$, reflecting a greater T5 vs. T4 differential for elderly ($\Delta=7.89$, $F(1,8)=34.4$, $p<.0005$) than young ($\Delta=3.24$, $F(1,10)=7.4$, $p<.05$) speakers.

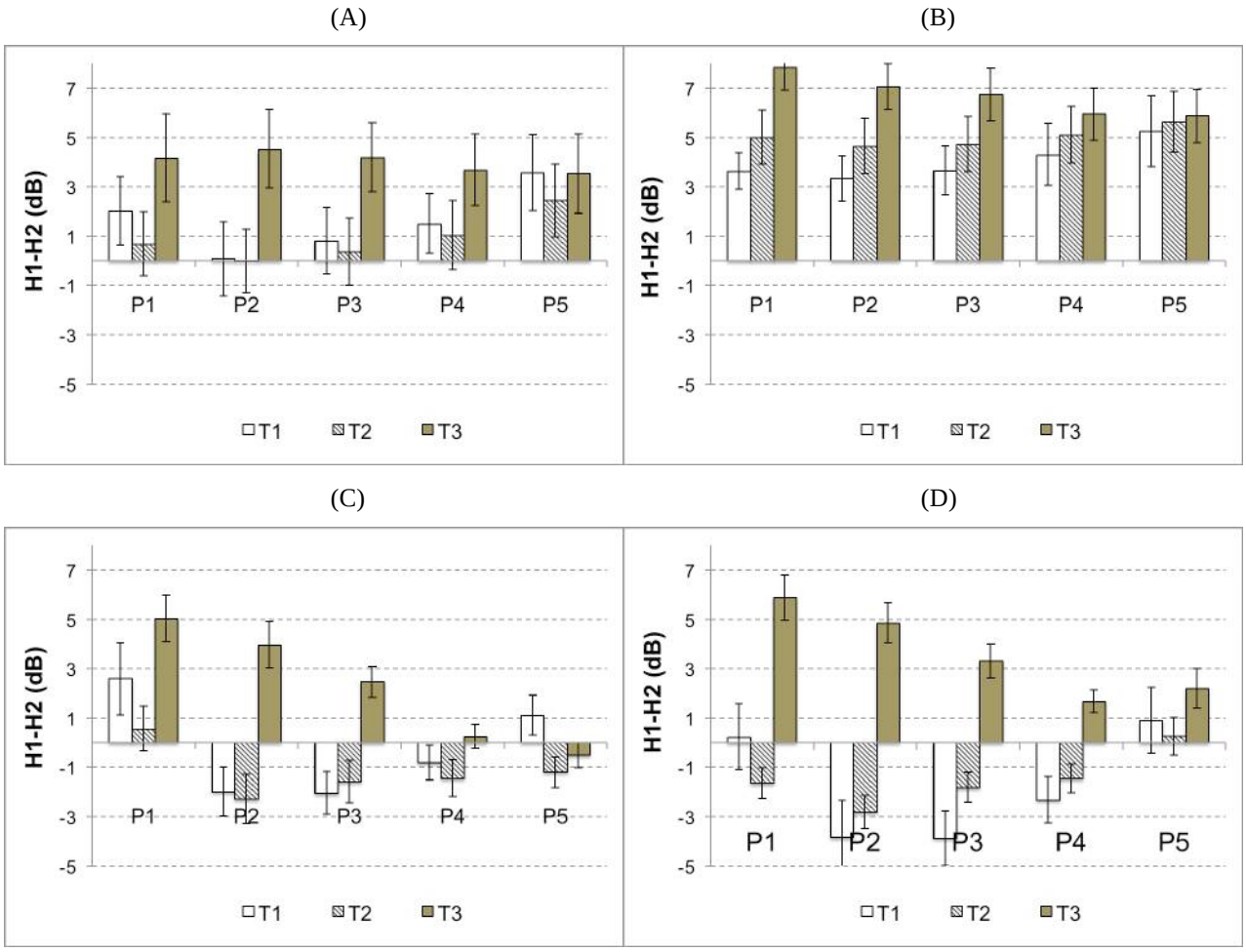


Figure 41. Average H1-H2 in monosyllables (computed using Voicesauce) for T1-3 (unchecked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

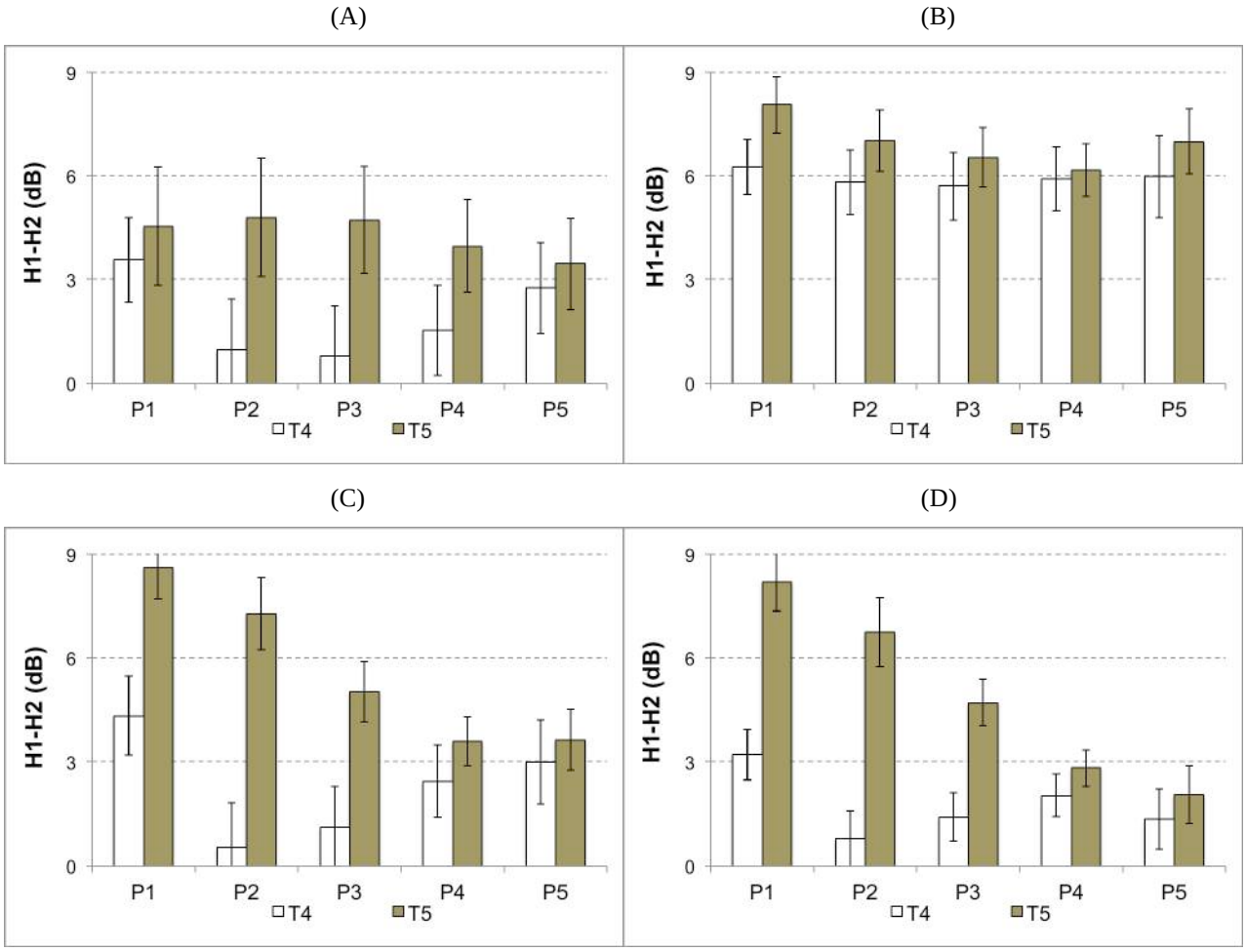


Figure 42. Average H1-H2 in monosyllables (computed using Voicesauce) for T4-5 (checked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

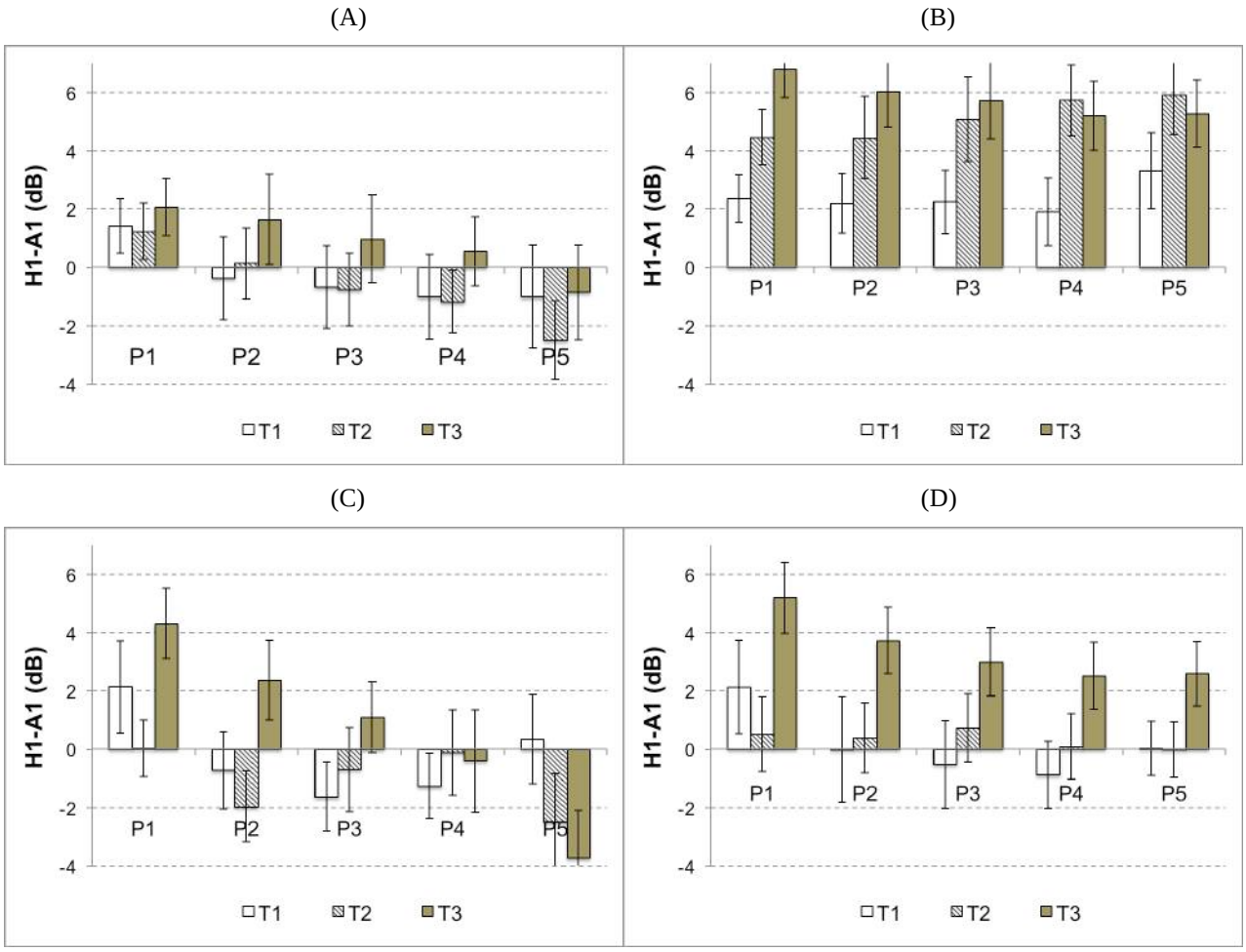


Figure 43. Average H1-A1 in monosyllables (computed using Voicesauce) for T1-3 (unchecked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

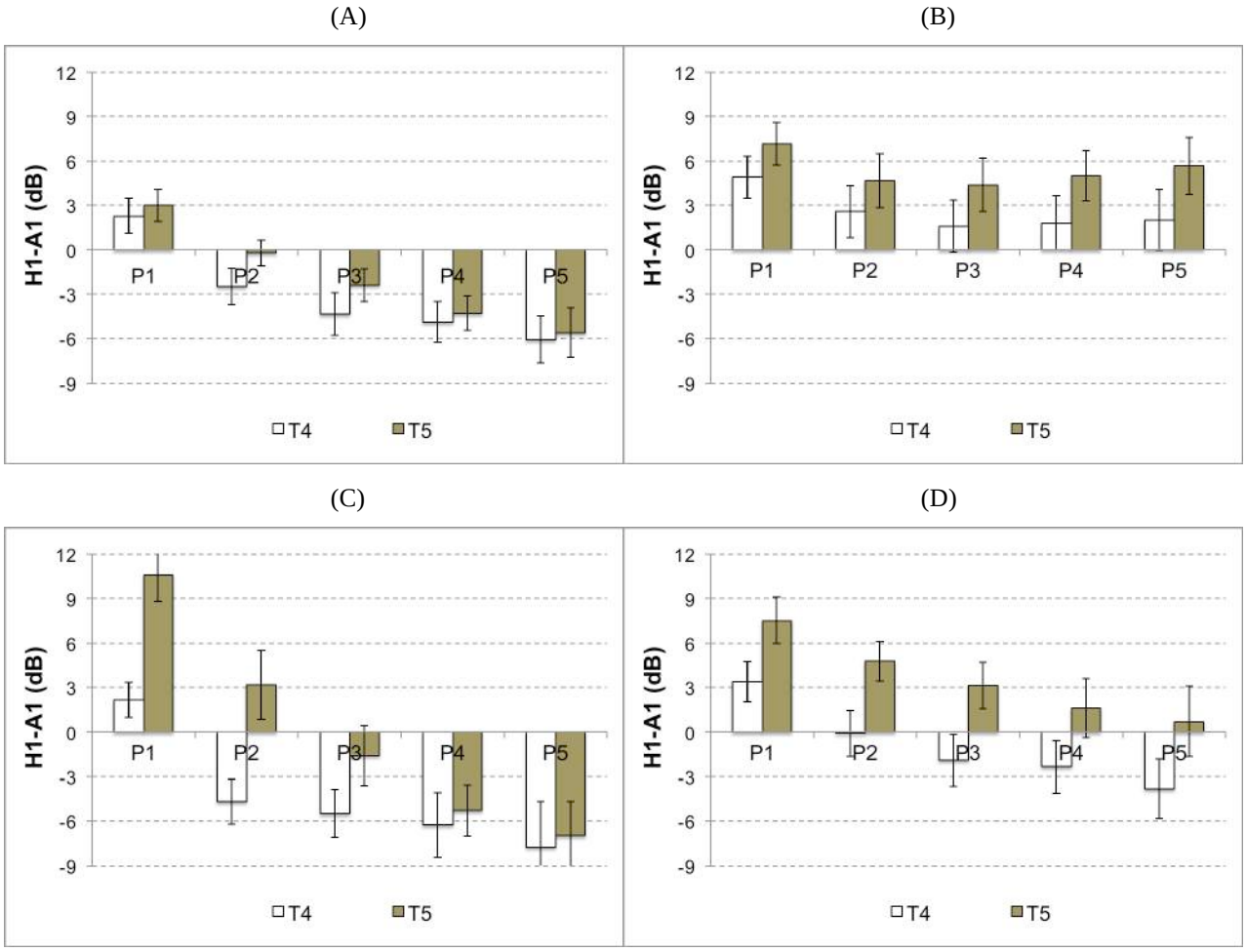


Figure 44. Average H1-A1 in monosyllables (computed using Voicesauce) for T4-5 (checked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

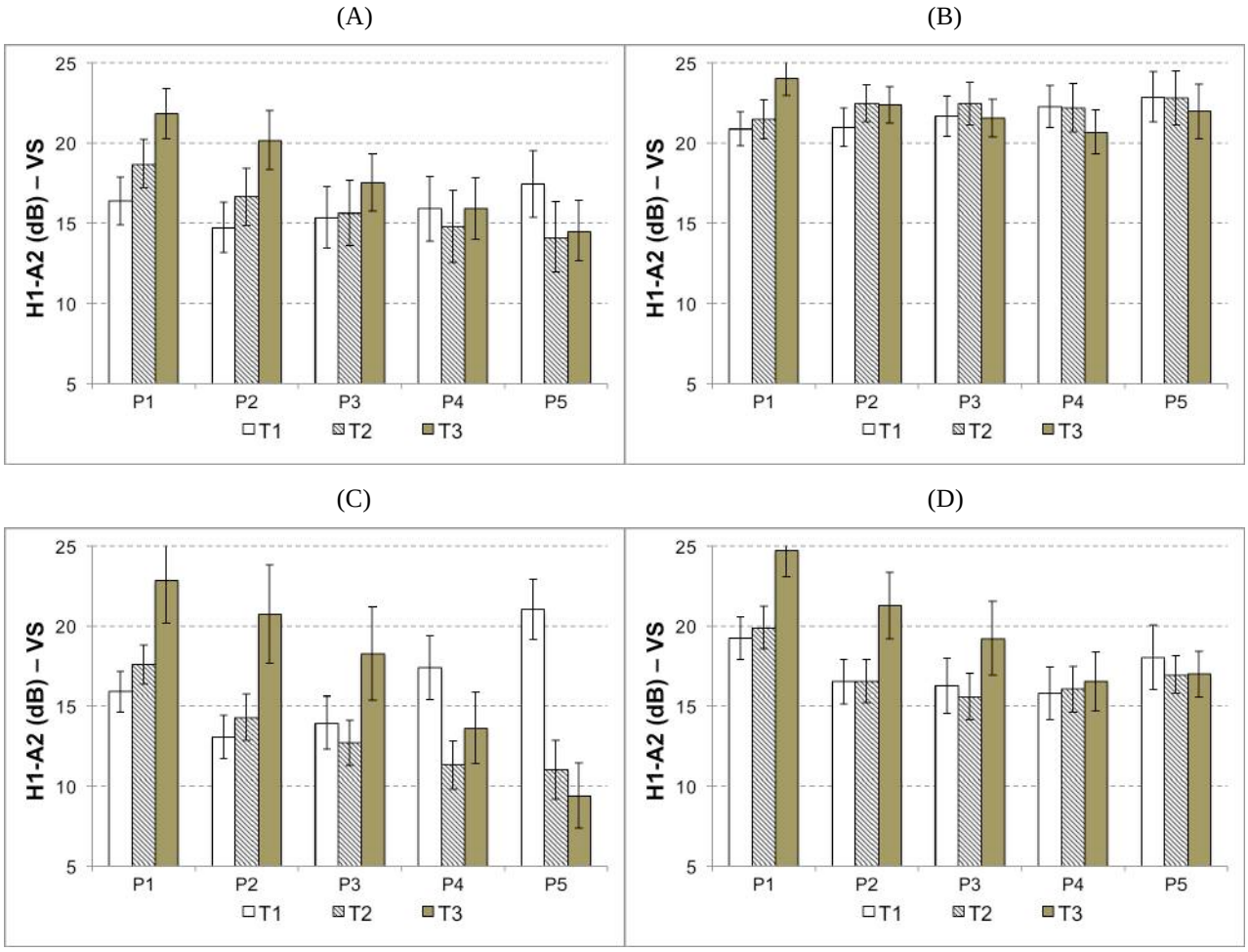


Figure 45. Average H1-A2 in monosyllables (computed using Voicesauce) for T1-3 (unchecked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

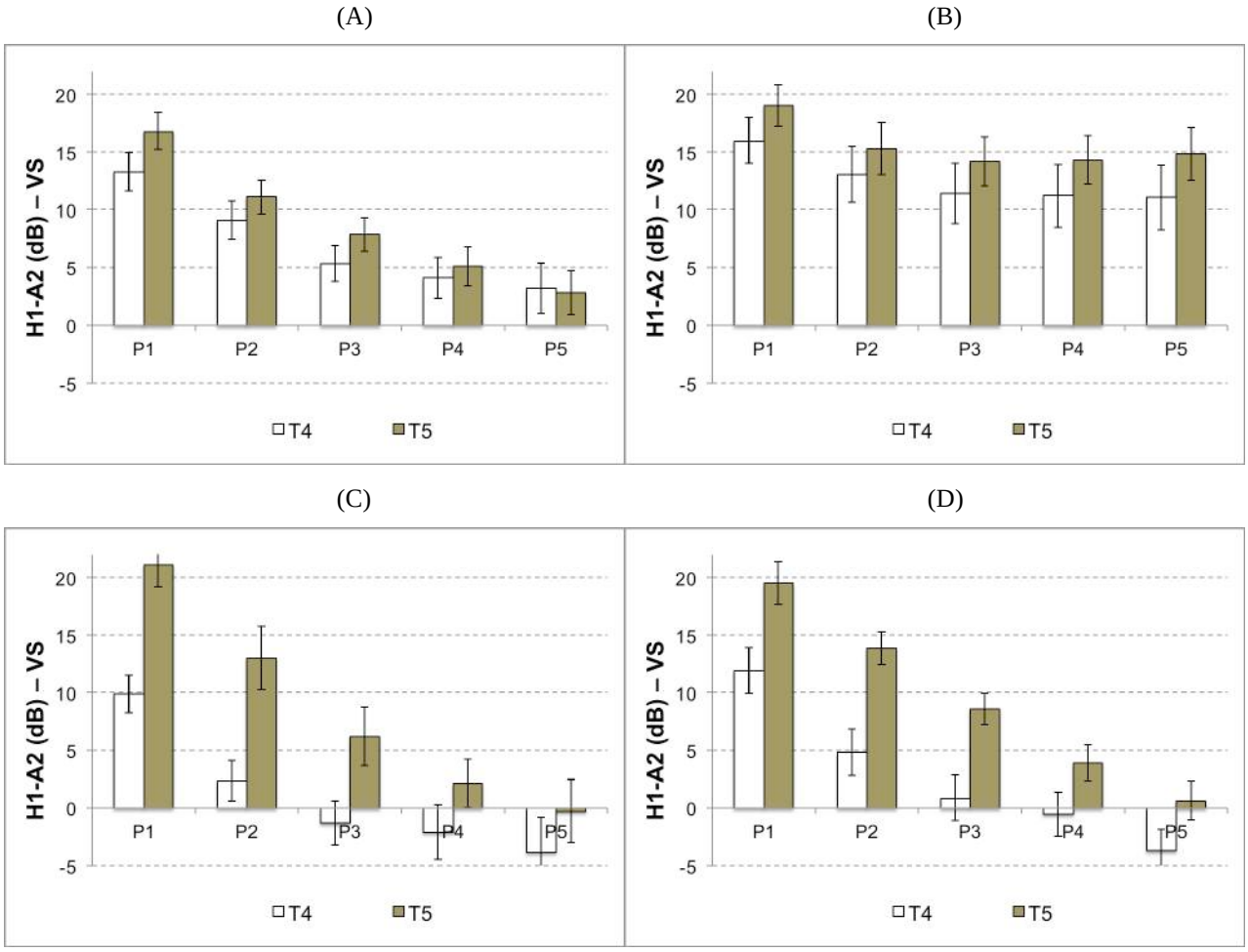


Figure 46. Average H1-A2 in monosyllables (computed using Voicesauce) for T4-5 (checked) rimes at five time points (P1-P5), according to Tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

Table 19. H1–H2 differentials between *yang* (T3) and *yin* (T1-2) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher H1–H2 for *yang* than *yin* syllable.

H1–H2		T3 vs. T1-2 differentials				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
	Ø	* 2.95	3.35	2.03	-0.27	-2.36
	/p, b/	1.66	2.89	2.32	1.60	0.97
young	**/t, d/**	0.91	* 3.80	* 4.06	2.24	-0.15
female	**/f, v/**	10.99	9.05	7.72	7.77	3.91
	/s, z/	-2.40	3.37	1.90	0.76	0.25
	/m, n/	0.69	3.14	2.15	0.06	0.32
	Ø	* 3.63	3.22	** 2.21	-0.21	-0.62
	/p, b/	** 3.11	** 2.66	** 2.43	* 1.78	0.93
young	**/t, d/**	** 3.64	* 3.54	* 2.45	1.27	* 1.89
male	**/f, v/**	4.79	3.66	4.49	2.59	0.33
	/s, z/	* 2.51	* 2.25	1.20	0.81	-0.34
	/m, n/	1.94	1.26	1.14	0.81	-0.22
	Ø	** 5.99	** 4.70	** 3.16	0.87	-1.10
	/p, b/	** 6.43	** 8.04	** 4.99	* 1.55	0.36
elderly	**/t, d/**	* 4.76	** 6.46	** 4.60	* 2.31	0.61
female	**/f, v/**	-0.62	** 4.13	** 3.83	0.76	-1.12
	/s, z/	0.70	** 7.11	* 4.80	1.44	-1.08
	/m, n/	-2.27	0.67	0.25	0.59	1.05
	Ø	** 9.92	** 8.17	* 6.28	3.03	0.19
	/p, b/	** 9.12	* 8.46	** 6.80	** 4.70	* 2.95
elderly	**/t, d/**	** 8.86	** 9.01	** 7.13	** 5.11	2.69
male	**/f, v/**	2.22	* 7.90	4.49	2.09	0.89
	/s, z/	* 2.85	** 7.41	** 6.05	* 2.87	1.34
	/m, n/	1.64	4.63	* 4.59	2.18	0.21

Table 20. H1–H2 differentials between *yang* (T5) and *yin* (T4) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher H1–H2 for *yang* than *yin* syllable.

H1–H2		T5 vs. T4 differentials				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
young female	**Ø**	2.35	2.56	2.33	0.72	0.35
	/p, b/	* 2.39	** 4.41	** 3.51	* 2.07	0.62
	/t, d/	1.95	** 5.97	** 3.73	* 1.73	-0.34
	/f, v/	0.26	2.38	4.60	3.65	1.07
	/s, z/	-2.13	* 3.89	** 5.60	* 4.03	1.86
young male	**Ø**	2.26	1.17	0.53	-0.66	-0.70
	/p, b/	* 2.72	* 2.77	2.02	* 1.95	* 4.26
	/t, d/	2.58	1.87	1.76	0.41	0.81
	/f, v/	0.71	-0.18	-0.39	0.37	0.31
	/s, z/	0.76	0.38	0.19	-0.78	0.38
elderly female	**Ø**	** 5.08	2.52	0.80	0.47	0.00
	/p, b/	** 7.40	** 6.84	** 3.61	1.14	0.16
	/t, d/	** 6.37	** 8.91	** 4.37	0.90	0.20
	/f, v/	0.81	* 7.42	* 4.94	1.83	3.59
	/s, z/	1.82	** 8.16	* 5.89	1.38	-0.76
elderly male	**Ø**	* 3.69	3.62	1.98	0.53	1.09
	/p, b/	7.08	* 6.39	2.69	0.20	-0.57
	/t, d/	* 6.52	* 5.99	* 5.64	1.74	0.15
	/f, v/	2.43	** 5.19	* 2.50	1.02	* 4.33
	/s, z/	** 5.26	* 8.69	** 3.75	0.52	-1.48

Table 21. H1–A1 differentials between *yang* (T3) and *yin* (T1-2) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher H1–A1 for *yang* than *yin* syllable.

H1–A1		T3 vs. T1-2 differentials				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
	Ø	1.97	1.40	1.19	1.81	0.83
	/p, b/	0.73	1.74	* 2.79	1.06	-0.46
young	**/t, d/**	1.39	2.96	2.21	2.65	0.75
female	**/f, v/**	0.41	0.71	0.89	0.97	2.18
	/s, z/	-0.81	1.99	1.42	1.76	1.19
	/m, n/	* 3.69	* 4.19	2.72	1.16	0.97
	Ø	* 5.55	3.15	* 3.12	1.97	0.67
	/p, b/	** 3.22	* 3.43	* 3.29	2.68	1.04
young	**/t, d/**	** 3.22	2.55	1.32	0.50	0.78
male	**/f, v/**	** 2.53	* 2.10	0.92	0.83	0.81
	/s, z/	2.48	2.26	1.62	0.94	-0.01
	/m, n/	** 4.08	3.34	2.76	* 1.88	0.67
	Ø	4.64	1.99	-0.12	-0.73	** -5.17
	/p, b/	4.34	* 5.11	2.37	0.15	-2.43
elderly	**/t, d/**	3.75	* 4.84	2.46	1.24	-0.64
female	**/f, v/**	2.14	1.82	2.51	0.13	-2.61
	/s, z/	1.28	* 4.85	4.06	0.64	-2.49
	/m, n/	3.78	* 4.30	1.71	1.05	** -2.68
	Ø	** 10.98	* 6.90	6.30	* 4.10	* -2.84
	/p, b/	* 7.45	* 6.00	* 4.83	** 3.60	3.20
elderly	**/t, d/**	* 7.78	* 6.67	* 4.79	3.48	3.06
male	**/f, v/**	3.28	* 5.77	1.56	0.55	1.18
	/s, z/	** 5.43	* 6.60	* 6.34	* 4.20	2.11
	/m, n/	1.85	2.13	* 1.53	0.98	-0.60

Table 22. H1–A1 differentials between *yang* (T5) and *yin* (T4) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher H1–A1 for *yang* than *yin* syllable.

H1–A1		**T5 vs. T4 differentials**				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
young female	**Ø**	3.94	* 6.15	*4.97	2.15	1.01
	/p, b/	0.38	2.90	2.36	0.03	-0.12
	/t, d/	-0.65	2.70	0.80	-0.58	-0.31
	/f, v/	1.15	-0.75	0.38	0.95	1.87
	/s, z/	-1.20	-0.03	1.25	0.42	-0.15
young male	**Ø**	** 6.88	3.80	2.59	3.22	4.46
	/p, b/	* 2.27	* 3.37	* 5.21	** 5.75	** 7.06
	/t, d/	1.69	1.59	2.51	2.27	2.64
	/f, v/	0.53	0.00	0.15	-0.16	0.63
	/s, z/	-0.01	1.69	* 3.34	* 4.86	3.40
elderly female	**Ø**	** 13.95	** 10.79	5.63	1.66	0.62
	/p, b/	** 7.87	** 7.84	3.21	* 2.94	3.50
	/t, d/	** 7.27	** 9.67	4.82	1.02	1.00
	/f, v/	6.73	5.39	1.79	-1.13	-0.73
	/s, z/	** 6.32	5.60	4.10	0.42	-0.54
elderly male	**Ø**	** 14.5	* 11.07	* 6.39	* 2.40	1.09
	/p, b/	* 7.23	* 9.25	6.37	1.67	2.47
	/t, d/	*9.98	* 11.00	* 8.98	3.13	-0.21
	/f, v/	2.73	6.82	* 5.89	2.42	6.88
	/s, z/	* 4.64	** 7.30	** 6.36	4.82	2.31

Table 23. H1–A2 differentials between *yang* (T3) and *yin* (T1-2) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher H1–A2 for *yang* than *yin* syllable.

H1–A2		**T3 vs. T1-2 differentials**				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
young female	**Ø**	* 5.61	5.22	2.67	0.60	-1.86
	/p, b/	1.98	2.26	0.70	-1.21	-3.24
	/t, d/	3.32	4.64	1.00	-0.61	* -3.88
	/f, v/	** 6.40	* 6.72	* 3.91	* 2.63	2.55
	/s, z/	** 4.12	3.60	1.85	1.43	0.11
	/m, n/	4.91	5.22	4.30	2.90	-0.51
young male	**Ø**	* 5.32	2.00	-0.06	0.01	0.14
	/p, b/	2.55	0.48	-0.78	-1.64	-0.91
	/t, d/	2.86	0.31	-1.24	** -3.50	* -2.58
	/f, v/	* 2.29	0.35	-0.37	-0.25	-0.18
	/s, z/	1.30	0.14	-0.08	* -2.29	-0.79
	/m, n/	1.85	0.44	1.03	1.38	1.71
elderly female	**Ø**	8.72	4.80	2.55	-1.48	** -5.58
	/p, b/	6.87	9.20	5.50	-2.14	-8.00
	/t, d/	5.46	5.08	0.46	-3.79	* -9.29
	/f, v/	3.96	7.48	6.75	1.65	* -5.25
	/s, z/	** 5.59	** 8.79	** 9.54	2.11	** -4.99
	/m, n/	** 6.88	** 8.85	* 5.96	0.77	-2.56
elderly male	**Ø**	* 10.74	7.72	7.85	2.93	-1.61
	/p, b/	4.36	5.89	3.69	-0.05	-1.00
	/t, d/	* 7.26	* 4.14	1.30	-1.39	-0.93
	/f, v/	0.03	1.52	0.10	-0.32	0.52
	/s, z/	3.48	4.37	6.50	1.93	0.58
	/m, n/	-4.42	-4.38	* -2.96	* -2.08	** -4.80

Table 24. H1–A2 differentials between *yang* (T5) and *yin* (T4) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher H1–A2 for *yang* than *yin* syllable.

H1–A2		**T5 vs. T4 differentials**				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
young female	**Ø**	9.16	* 11.16	* 6.84	3.25	0.54
	/p, b/	0.20	0.15	2.85	1.28	-0.64
	/t, d/	2.90	2.38	1.90	-0.57	-2.59
	/f, v/	2.13	-1.46	0.70	0.92	1.00
	/s, z/	3.15	-2.22	0.50	0.26	-0.26
young male	**Ø**	** 8.56	6.15	4.22	3.73	4.23
	/p, b/	* 3.08	* 3.72	* 4.61	** 6.58	** 8.39
	/t, d/	0.60	0.58	2.10	1.11	0.43
	/f, v/	1.37	-0.38	0.79	0.53	0.50
	/s, z/	1.57	1.13	* 2.24	* 3.60	5.14
elderly female	**Ø**	** 17.34	** 14.58	7.47	2.97	1.12
	/p, b/	** 11.51	** 10.01	* 7.72	4.66	5.63
	/t, d/	** 10.30	* 11.75	9.23	4.37	4.21
	/f, v/	7.91	9.42	5.55	4.39	4.41
	/s, z/	** 8.94	7.42	7.67	4.86	2.51
elderly male	**Ø**	** 14.05	* 12.36	* 9.33	* 7.26	5.94
	/p, b/	* 5.93	* 9.37	6.73	2.80	3.03
	/t, d/	* 8.89	* 9.67	* 9.71	3.72	1.30
	/f, v/	2.71	6.11	* 6.47	3.47	8.53
	/s, z/	* 6.41	** 7.66	** 6.50	5.04	3.02

- **Cepstral Peak Prominence**

Figure 48 shows the CPP values (computed using Voicesauce) according to tone, pooled across time points, speaker groups, and onset types. Recall that breathier phonation should be indexed by lower CPP values. CPP is highest overall for T2 syllables, followed by T3 syllables, which are followed by T1 syllables. CPP is numerically higher for T4 than T5 syllables.

Figure 48 (unchecked syllables) and Figure 49 (checked syllables) further show CPP values at each of the five time points (*yang* tones in muddy color) for (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers. As can be seen in the figures, CPP is systematically the highest, among unchecked syllables, for T2 syllables. It is higher for T1 than T3 only at the beginning of the vowel and only for elderly speakers, whereas for young speakers, the pattern seems to be reversed to the advantage of T3. Likewise, for checked syllables, CPP is higher for T4 than T5 vowels at the first three time points only for elderly speakers whereas no noticeable difference is observed for young speakers.

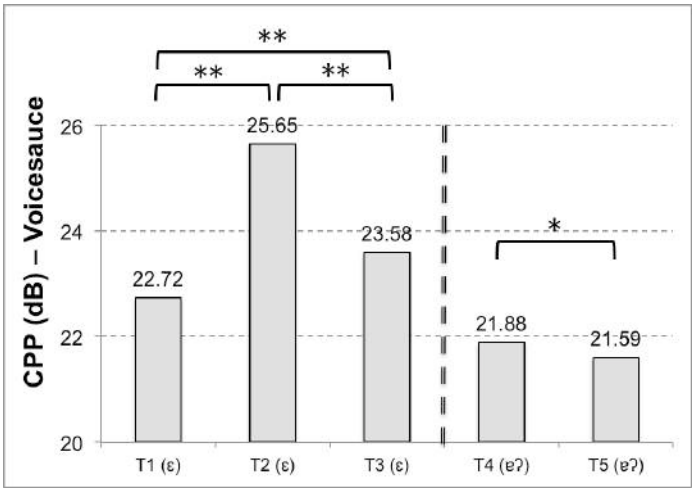


Figure 47. Average CPP values pooled over all onset types, all five vowel time points, and all subject groups in monosyllables measured in Voicesauce.
Significance levels: * for $p<.05$; ** for $p<.01$.

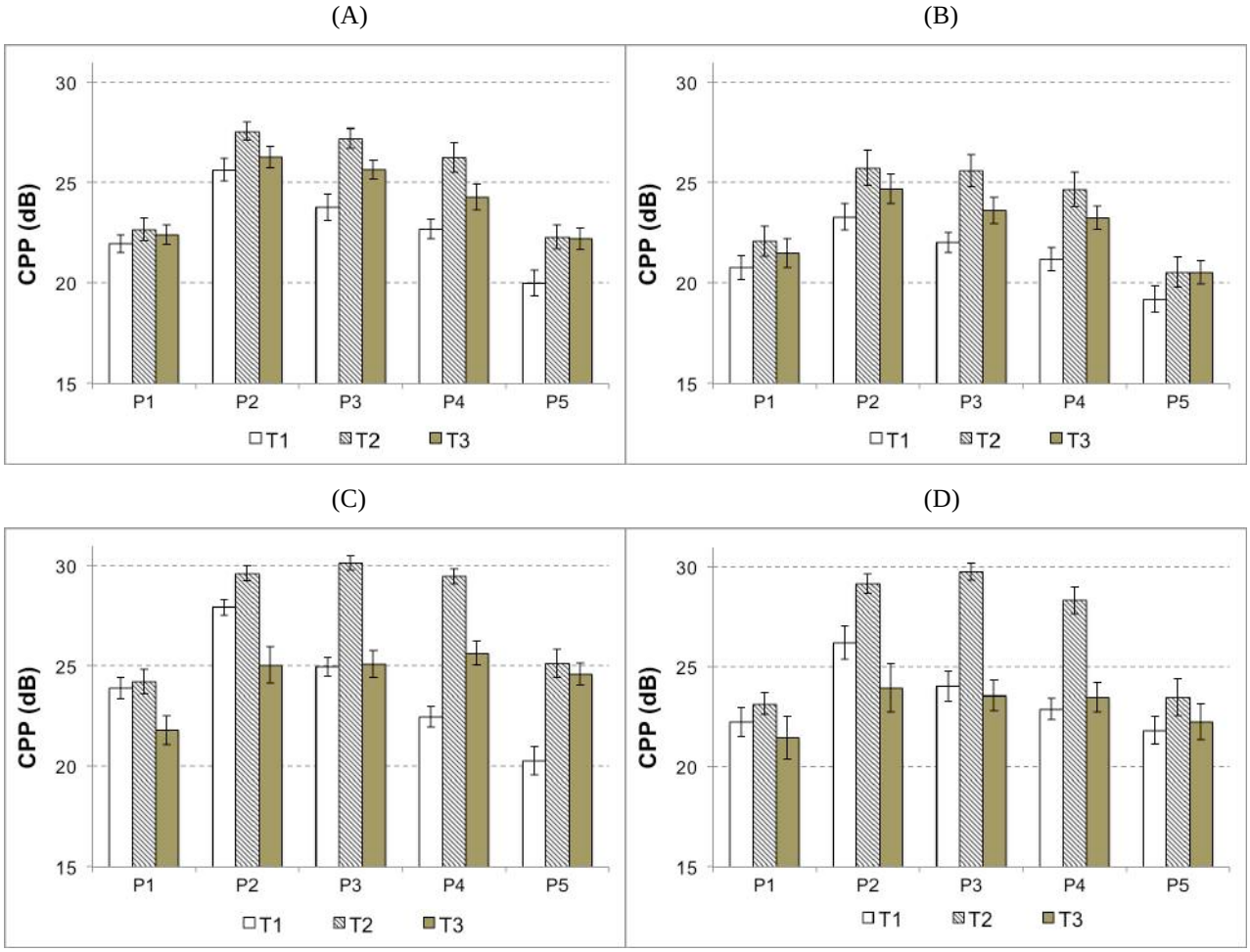
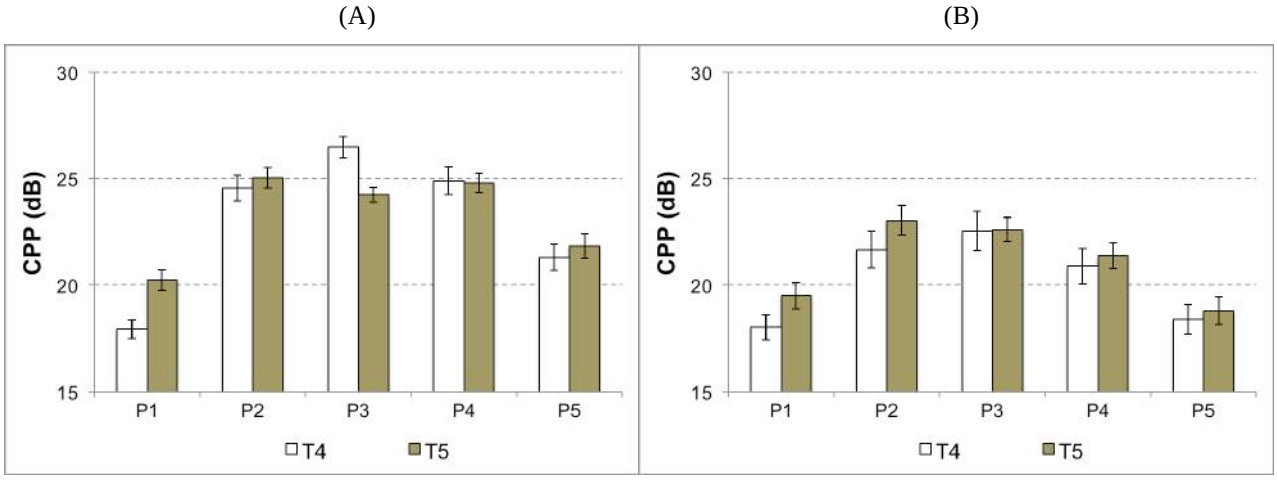


Figure 48. Average CPP in monosyllables (computed using Voicesauce) for T1-3 (unchecked) rimes at five time points (P1-P5), according to Tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.



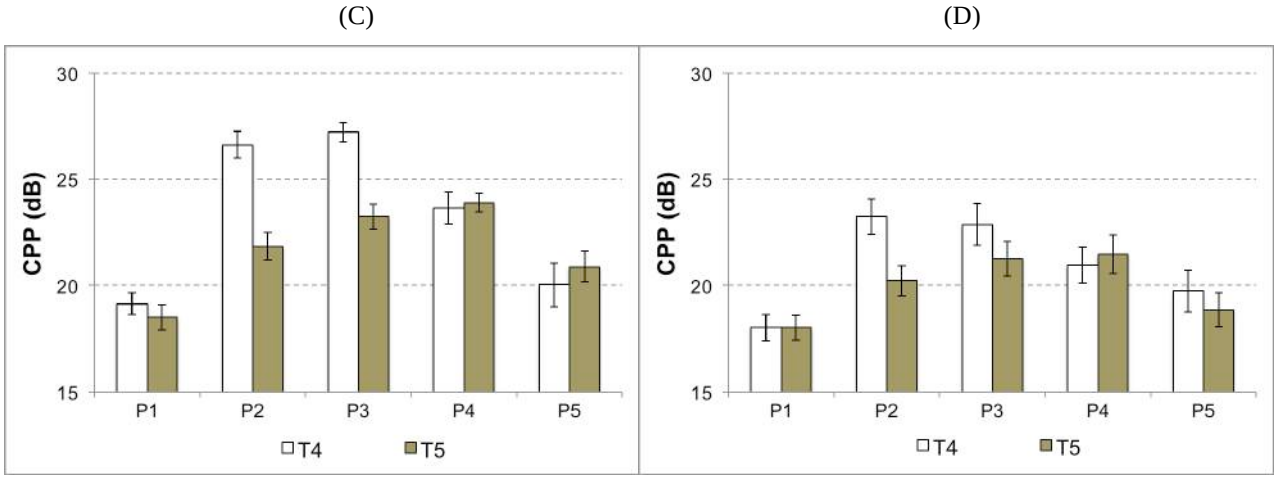


Figure 49. Average CPP in monosyllables (computed using Voicesauce) for T1-3 (unchecked) rimes at five time points (P1-P5), according to Tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

In order to substantiate these observations, we conducted by-subject ANOVAs separately for unchecked and checked syllables. *CPP* was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5), *Onset* (unchecked: Ø, /p/, /t/, /f/, /s/, /m, n/; checked: Ø, /p/, /t/, /f/, /s/) and *Time point* (P1, P2, P3, P4, P5) were within-subject factors. As explained for the spectral tilt measures, we pooled /m/ and /n/ onsets for unchecked syllables and excluded nasal onset syllables from analyses for checked syllables. In the following paragraphs, we report only the overall effect of *Tone*, as well as the interaction between *Tone* and *Gender* or *Tone* and *Age*. The detailed *yin–yang* difference data with statistical significance are reported in Table 25 and Table 26.

For unchecked syllables, *Tone* was highly significant, $F(2,36)=72.3$, $p<.0001$. *CPP* was lower for T3 than T2 syllables (23.58<26.65 dB, $F(1,18)=56.1$, $p<.0001$), but higher for T3 than T1 syllables (23.58>22.72, $F(1,18)=12.6$, $p<.005$). *Tone* did not interact with *Gender*, $F(2,36)<1$, but the *Tone* × *Age* interaction was significant, $F(2,36)=12.5$, $p<.0001$. *CPP* was significantly lower for T3 than other tones only for elderly speakers ($\Delta=1.75$, $F(1,8)=21.0$, $p<.005$), not for young speakers ($\Delta=-0.19$, $F(1,10)<1$).

For checked syllables, *Tone* was significant, $F(1,18)=5.1$, $p<.05$. *CPP* was lower for T5 than T4 syllables (21.59<21.88 dB). Again, *Tone* did not interact with *Gender*,

$F(1,18)=2.1$, $p=.16$, but the *Tone* × *Age* interaction was significant, $F(1,18)=14.9$, $p<.005$. *CPP* was significantly lower for T5 than T4 syllables only for elderly speakers ($\Delta=1.32$, $F(1,8)=10.6$, $p<.05$), not for young speakers ($\Delta=-0.50$, $F(1,10)=1.7$, $p=.22$).

Table 25. CPP differentials between *yang* (T3) and *yin* (T1-2) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly lower CPP for *yang* than *yin* syllable.

CPP		**T1-2 vs. T3 differentials**				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
	Ø	-0.33	0.01	-0.89	-0.17	* -1.38
	/p, b/	0.03	0.77	-0.16	-0.08	-0.20
young	**/t, d/**	0.16	0.52	-0.66	0.32	* -1.90
female	**/f, v/**	-0.45	-0.22	-0.25	-0.37	-1.23
	/s, z/	-0.54	-0.08	-0.26	0.45	-1.08
	/m, n/	0.29	0.27	-0.37	-0.36	-1.29
	Ø	-0.18	0.08	0.18	-0.33	0.19
	/p, b/	-0.36	-0.11	1.00	-0.51	* -1.35
young	**/t, d/**	* 0.66	0.42	0.21	-0.25	-0.77
male	**/f, v/**	-0.35	-0.68	0.16	0.61	0.21
	/s, z/	0.01	-0.26	-0.91	-1.16	-1.51
	/m, n/	0.06	-0.79	-0.43	-1.23	-1.10
	Ø	1.49	* 4.57	* 3.09	1.00	-1.76
	/p, b/	2.91	4.06	2.22	-0.92	-1.99
elderly	**/t, d/**	* 2.90	3.27	1.69	-0.41	-2.52
female	**/f, v/**	1.38	* 2.31	2.11	0.31	-2.35
	/s, z/	* 2.88	* 5.50	2.96	0.75	-1.93
	/m, n/	2.33	2.85	1.35	-0.92	* -2.56
	Ø	** 4.57	* 7.70	* 6.20	3.06	-0.94
	/p, b/	3.91	** 6.97	4.57	2.31	0.54
elderly	**/t, d/**	2.70	* 4.49	3.29	1.53	0.04
male	**/f, v/**	-0.21	1.20	1.48	0.95	0.64
	/s, z/	-0.04	3.90	** 4.38	2.68	0.68
	m/n	-1.21	0.28	0.00	0.79	0.85

Table 26. CPP differentials between *yang* (T5) and *yin* (T4) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences: * for $p<.05$; ** for $p<.01$. Shaded cells: significantly higher CPP for *yang* than *yin* syllable.

CPP		T4 vs. T5 differentials				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
young female	**Ø**	-0.11	0.16	1.60	0.11	-2.49
	/p, b/	-1.14	-0.09	* 1.95	0.55	0.66
	/t, d/	-1.98	0.30	** 2.74	0.70	0.04
	/f, v/	** -3.61	-1.24	* 1.86	-0.56	1.04
	/s, z/	** -3.16	-2.18	** 2.40	0.42	0.90
young male	**Ø**	-1.86	-2.47	0.29	* 1.91	0.34
	/p, b/	0.41	-0.21	0.45	0.19	0.33
	/t, d/	-0.83	-0.14	1.81	-0.09	-1.40
	/f, v/	* -2.23	-2.40	-2.03	* -3.09	* -1.38
	/s, z/	-0.77	-0.47	0.09	-0.41	-0.21
elderly female	**Ø**	1.34	* 4.27	* 5.03	2.84	-1.11
	/p, b/	1.66	** 4.70	* 2.90	-1.51	-1.75
	/t, d/	* 2.27	** 6.90	** 4.55	-0.75	-0.97
	/f, v/	0.04	* 6.36	** 4.25	-0.91	-0.47
	/s, z/	0.66	4.69	4.16	-0.40	0.23
elderly male	**Ø**	0.92	3.60	* 3.75	1.36	0.59
	/p, b/	-0.47	** 4.90	2.35	-1.18	-0.86
	/t, d/	0.45	2.75	* 3.18	-0.60	0.43
	/f, v/	0.51	2.68	0.66	-3.59	0.12
	/s, z/	0.54	* 4.10	0.44	-0.98	0.84

- **First formant (vowel quality)**

Figure 50 shows the overall F1 values in the vowel (/ɛ/ for unchecked syllables, /a/ for checked syllables) as a function of tone, pooled across the five time intervals, speaker groups and onset types. Previous studies showed a F1 lowering with breathy voice but this tendency was not systematic (e.g., Thongkum, 1988). In our study, F1 is higher in /ɛ/ for T3 (*yang*) than T1 or T2 (*yin*), but no difference is observed in /a/ between T4 and T5. Figure 51 shows the time course of F1 in the vowel as a function of tone. F1 increases along the vowel for all tones except for T1. This is probably because some speakers produced several syllables in tone T1 with a slightly diphthongized vowel [e^{i}] , whereas no such diphthongization was produced in tone T2 or T3. Many young speakers produce the Shanghai /ɛ/ vowel with diphthongization when it corresponds to the [ei] rime in Standard Chinese, as explained in §2.1.2. For this reason, we excluded T1 syllables from the detailed figures and from statistical analyses.

Figure 52 (unchecked syllables) and Figure 53 (checked syllables) further show F1 values at each of the five time points (*yang* tones in muddy color) for (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers. For unchecked syllables, F1 is numerically higher for T2 than T1 across the entire vowel for all speaker groups, except elderly female speakers. For checked syllables, the relation between T4 and T5 vowels is more variable and little or no difference in F1 can be observed between the two tones.

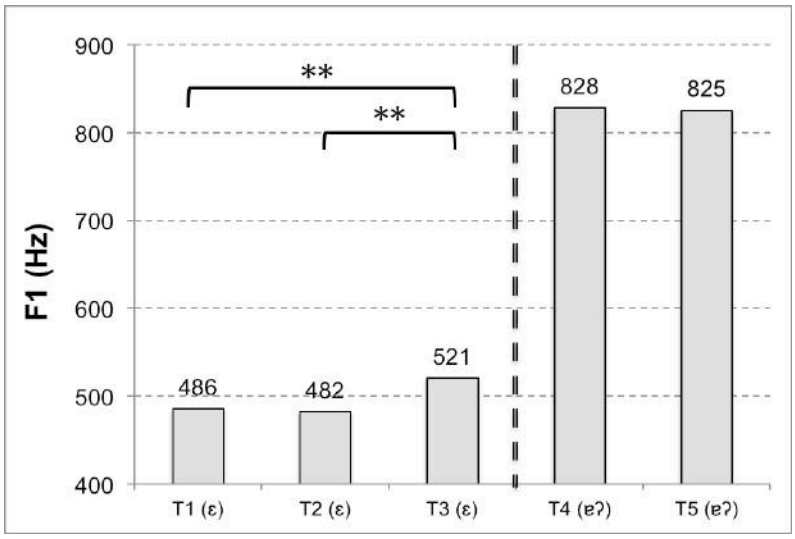


Figure 50. F1 averaged across the monosyllables' vowel, as a function of tone.

Significance level: ** for p<.001.

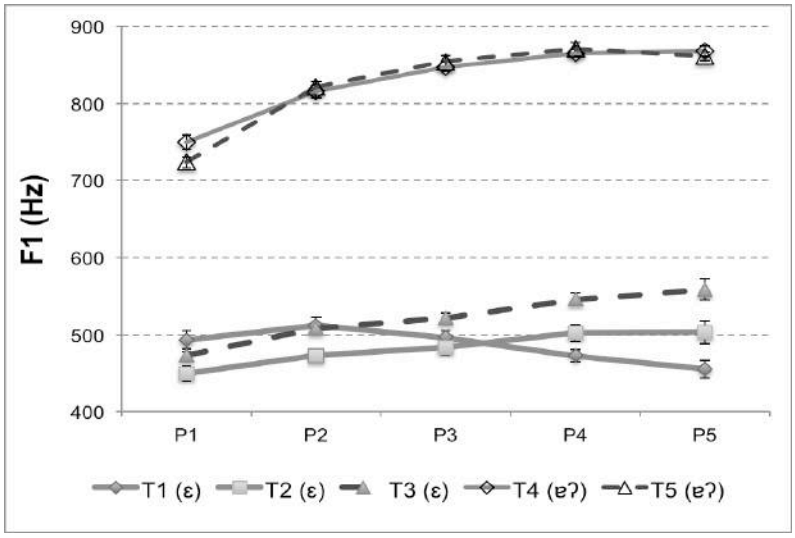


Figure 51. F1 at the five time points along the monosyllables' vowel, as a function of tone.

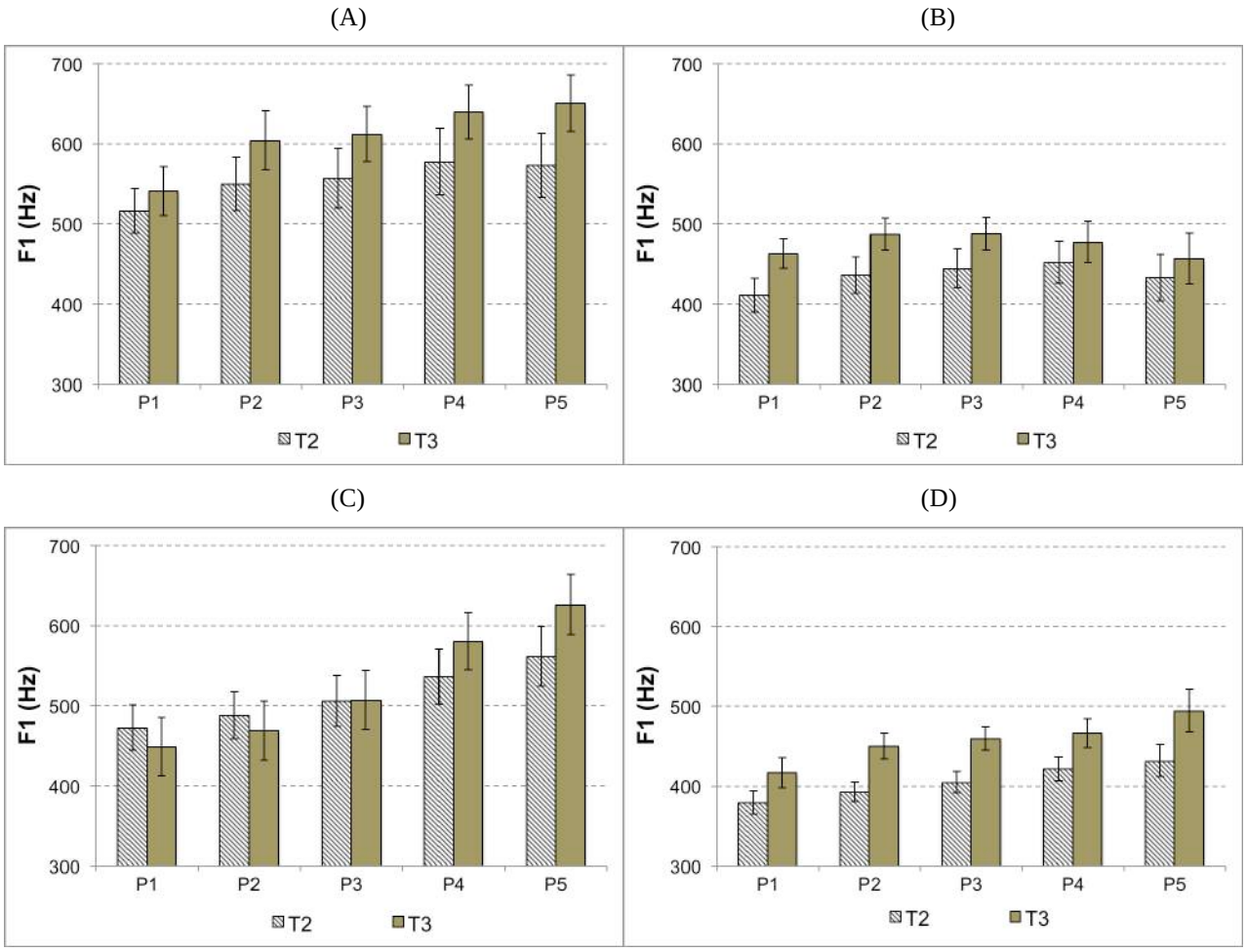


Figure 52. Average F1 in monosyllables for T1-3 (unchecked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

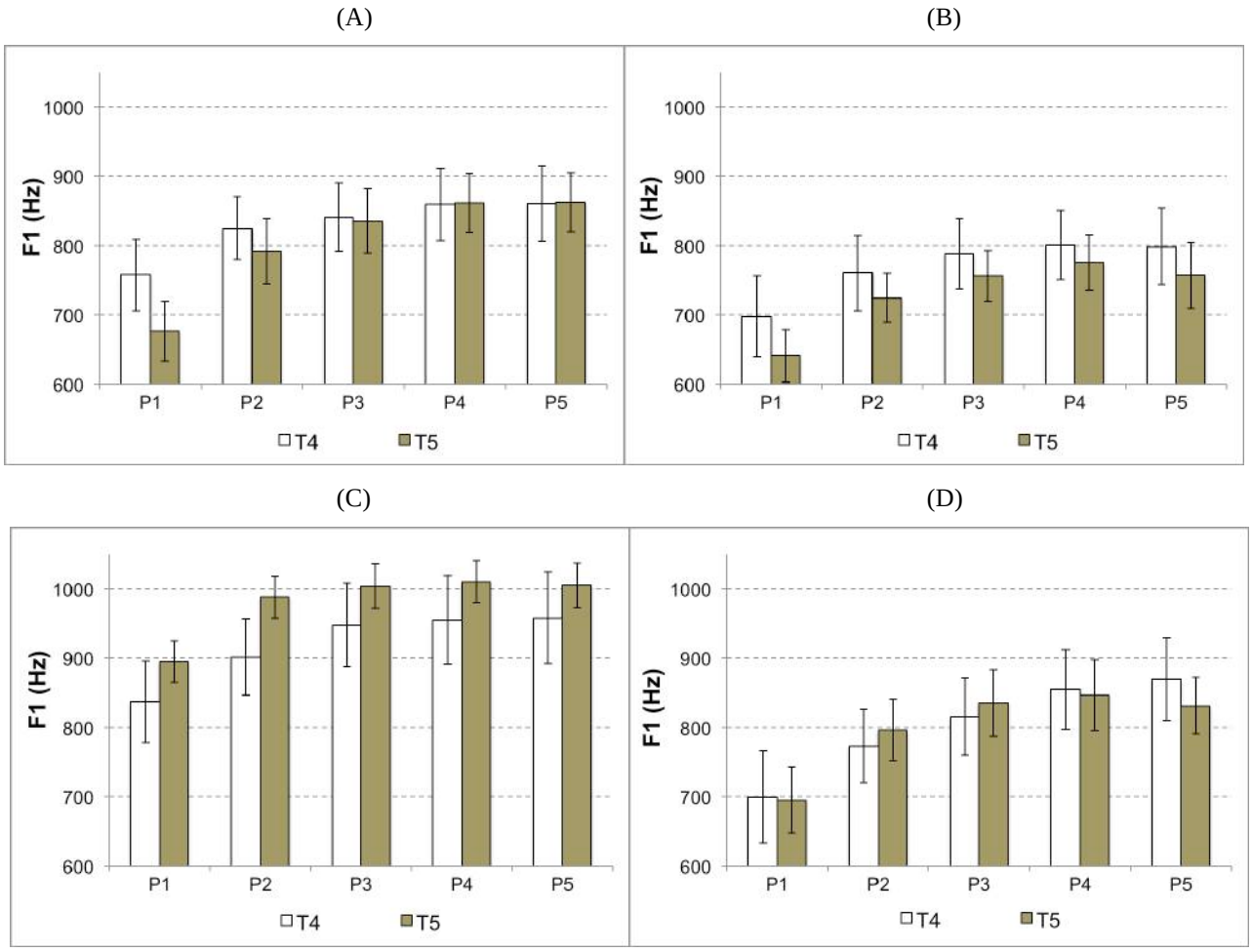


Figure 53. Average F1 in monosyllables for T4-5 (checked) rimes at five time points (P1-P5), according to tone and speaker group. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

We again conducted by-subject ANOVAs separately for unchecked and checked syllables. *F1* was the dependent variable, and *Subject* the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Tone* (unchecked: T2 vs. T3; checked: T4 vs. T5), *Onset* (zero, /p/, /t/, /f/, /s/) and *Time point* (P1, P2, P3, P4, P5) were within-subject factors. We excluded nasal onset syllables because T2 does not co-occur with /n/ and because /ɛ/ in both T2 /mɛ/ and T3 /mɛ/ or /nɛ/ can be diphthongized into [e^{i}]. Moreover, T4 does not co-occur with nasal onsets. In the following paragraphs, we only report the overall effect of *Tone*, as well as the interaction between *Tone* and *Gender* or *Tone* and *Age*.

For unchecked syllables (T2 and T3), *Tone* was significant, $F(1,18)=12.6$, $p<.005$. *F1* was higher for T3 than T2 syllables (521>482 Hz, $F(1,18)=19.0$, $p<.0005$), and for T3 than T1 syllables (521>486 Hz, $F(1,18)=31.1$, $p<.0001$). We expected, however, lower *F1* for T3 than T1 syllables. *Tone* did not interact with either *Gender* or *Age*, Fs$(1,18)<1$. As for checked syllables, *Tone* was not significant, $F(1,18)<1$. Thus, we do not report detailed data for checked syllables. Detailed *yin–yang* difference data with statistical significance for T2 and T3 syllables are reported in Table 27.

Table 27. F1 differentials between *yang* (T3) and *yin* (T1-2) monosyllables, according to speaker group, syllable onset, and time point. Significance levels for *yin–yang* differences (all higher for *yang* than *yin* syllables): * for $p<.05$; ** for $p<.01$.

F1		**T3 vs. T2 differentials**				
speaker	**onset**	**P1**	**P2**	**P3**	**P4**	**P5**
young female	**Ø**	67	68	70	93	159
	/p, b/	13	43	45	26	40
	/t, d/	10	48	71	89	110
	/f, v/	39	86	86	110	100
	/s, z/	-18	38	50	65	65
young male	**Ø**	** 60	19	29	28	30
	/p, b/	14	32	** 41	9	21
	/t, d/	** 44	* 43	42	34	34
	/f, v/	* 30	** 43	** 34	* 24	24
	/s, z/	45	* 51	38	46	23
elderly female	**Ø**	-32	-64	57	9	99
	/p, b/	-6	5	32	69	71
	/t, d/	-35	-17	-12	38	* 71
	/f, v/	-44	-39	24	54	73
	/s, z/	53	** 72	* 61	* 86	59
elderly male	**Ø**	45	49	55	47	69
	/p, b/	* 60	* 87	76	64	102
	/t, d/	** 50	* 76	66	67	76
	/f, v/	** 47	* 68	34	46	47
	/s, z/	35	57	65	50	69

4.1.6.3.3 Discriminant analyses for vowels in monosyllables

Linear discriminant analyses (LDA) were conducted to determine which measures among those reported in the preceding sections were the most efficient to discriminate breathy from modal phonation. We used the MASS package (Venables & Ripley, 2002) and the klaR package (Weihs, Ligges, Luebke, & Raabe, 2005) in R (R Core team, 2014).

The factors included in the LDA were the following five acoustic measures: H1–H2, H1–A1, H1–A2, CPP, and F1, all pooled across the five time points in each vowel. Analyses were conducted separately for each speaker group and for checked and unchecked syllables to discriminate between *yin* and *yang* syllables (unchecked syllables: T1 and T2 vs. T3; checked syllables: T4 vs. T5). For most measures, discrimination was more efficient in the first half of the vowel. For F1, however, F1 was numerically higher for *yang* than *yin* syllables over the entire vowel. For this reason we used the averaged values over the entire vowel instead of solely the first half. In order to prevent the discriminant analysis to be affected by outliers, we removed the data beyond two standard deviations from the means, for each tone category and for each speaker group.

The results of the linear discriminant analyses include:

(1) the overall Wilks' Lambda values (from 0 to 1) (output of the *manova* function) for the model including all the acoustic measures as variables; smaller values indicate higher significance;

(2) the coefficients of the linear discriminant functions (from –1 o 1) (output of the *lda* function), indicating the relative importance of each variable in the discriminant model;

(3) the LDA classification confusion matrices, indicating how well the discriminant functions predict each category;

(4) the *F* and *p* values for each variable of interest, converted from partial Wilks' Lambda values, which result from stepwise discriminant analyses, evaluating the significance of each variable in discriminating between the tone categories (output of the *greedy.wilks* function: partial Wilks' Lambda values are not provided by R).

Globally, the results show that the discriminant analysis combining all the five acoustic measures is the most efficient for elderly male speakers, followed by elderly

female, young male, and, lastly, by young female speakers, for both unchecked and checked tones. (This variation between speaker groups was also shown in the ANOVAs reported in the previous sections.) This ordering is supported by the overall Wilk's Lambda of the LDA model for each speaker group (lower values indicate higher significance) (Table 28) and the classification error rates of the discriminant functions (lower error rates for higher discriminant accuracy) (Table 29). Note that the classification confusion matrices in Table 29 were based on the predictions using a cross-validated dataset, which are more conservative and yield higher error rates than the predictions without cross-validation. Not using cross-validation, however, yielded the same ordering of speaker groups in terms of increasing prediction error rates: elderly male, elderly female, young male, and young female (unchecked syllables: 5%<16.3%<25.3%<28.5%); checked syllables: 8.7%<17.5%<18.6%<20.0%).

As for the relative importance of each acoustic measure in the discriminant analysis, the most important turns out to be the widely accepted H1–H2 measure. The importance of the measures is indicated by the coefficients of the discriminant functions and by the partial Wilks' Lambdas. When the model is significantly predictive, these two indicators converge. When the model is less predictive, the two indicators sometimes suggest different rankings of the measures in terms of their importance. The statistically important measures are shown in Table 30. When several measures appear, they are listed in order of decreasing importance. Detailed results for each speaker group and for unchecked and checked tone are reported in Appendix 2.

Scatterplot matrices (without outliers) are shown in Figure 54 (unchecked tones) and Figure 55 (checked tones). The labels in the diagonal are the labels for both the Y axis of the plots in the same row and the X axis of the plots in the same column. For example, the plot in the top row, fourth column is a scatterplot of H2–H2 (Y axis) × CPP (X axis). From the figures, it can again be seen that the separation between *yin* and *yang* tones is better achieved for the elderly male speakers than the other three speaker groups. Moreover, the separation is better achieved for unchecked than checked tones.

Table 28. Overall Wilk's Lambda of the LDA model for each speaker group (lower values indicate higher significance).

	young female	**young male**	**elderly female**	**elderly male**
unchecked	0.88	0.74	0.59	0.28
checked	0.72	0.68	0.58	0.33

Table 29. LDA classification confusion matrices with overall error rates on cross-validated datasets: the column labels indicate the actual categories and the row labels indicate the categories predicted by the LDA model, using all five acoustic measures.

(young female)

	T1-2	T3
T1-2	103	39
T3	18	26

error rate: 30.6%

	T4	T5
T4	27	10
T5	14	54

error rate: 22.9%

(young male)

	T1-2	T3
T1-2	106	33
T3	22	33

error rate: 28.4%

	T4	T5
T4	30	7
T5	17	64

error rate: 20.3%

(elderly female)

	T1-2	T3
T1-2	100	15
T3	14	43

error rate: 16.9%

	T4	T5
T4	29	7
T5	13	54

error rate: 19.4%

(elderly male)

	T1-2	T3
T1-2	70	5
T3	3	36

error rate: 7.0%

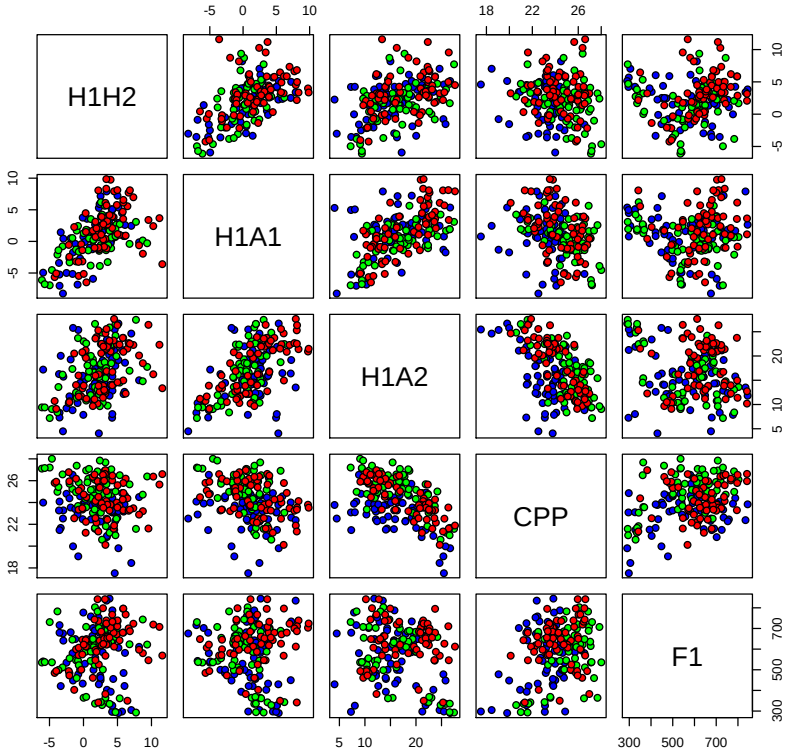
	T4	T5
T4	23	4
T5	5	37

error rate: 13.0%

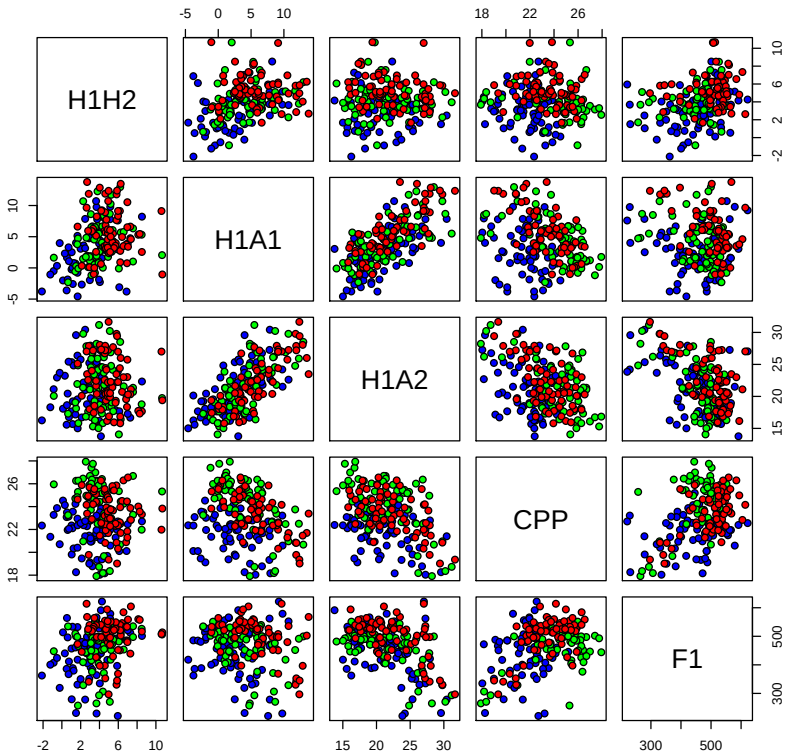
Table 30. Statistically significant acoustic measures in the discriminant function for each speaker group as shown by the two indicators.

	T1-2 vs. T3		**T4 vs. T5**	
	Coeff.	**Wilks' Lambda**	**Coeff.**	**Wilks' Lambda**
young female	CPP	H1–H2, F1	H1–H2, F1	
young male	H1–H2, H1–A1, F1		CPP	H1–A1, CPP, H1–H2
elderly female	H1–H2, H1–A2		CPP	H1–H2
elderly male	H1–H2, CPP		H1–H2, H1–A1, H1–A2	

Young female: Tones 1 - 3



Young male: Tones 1 - 3



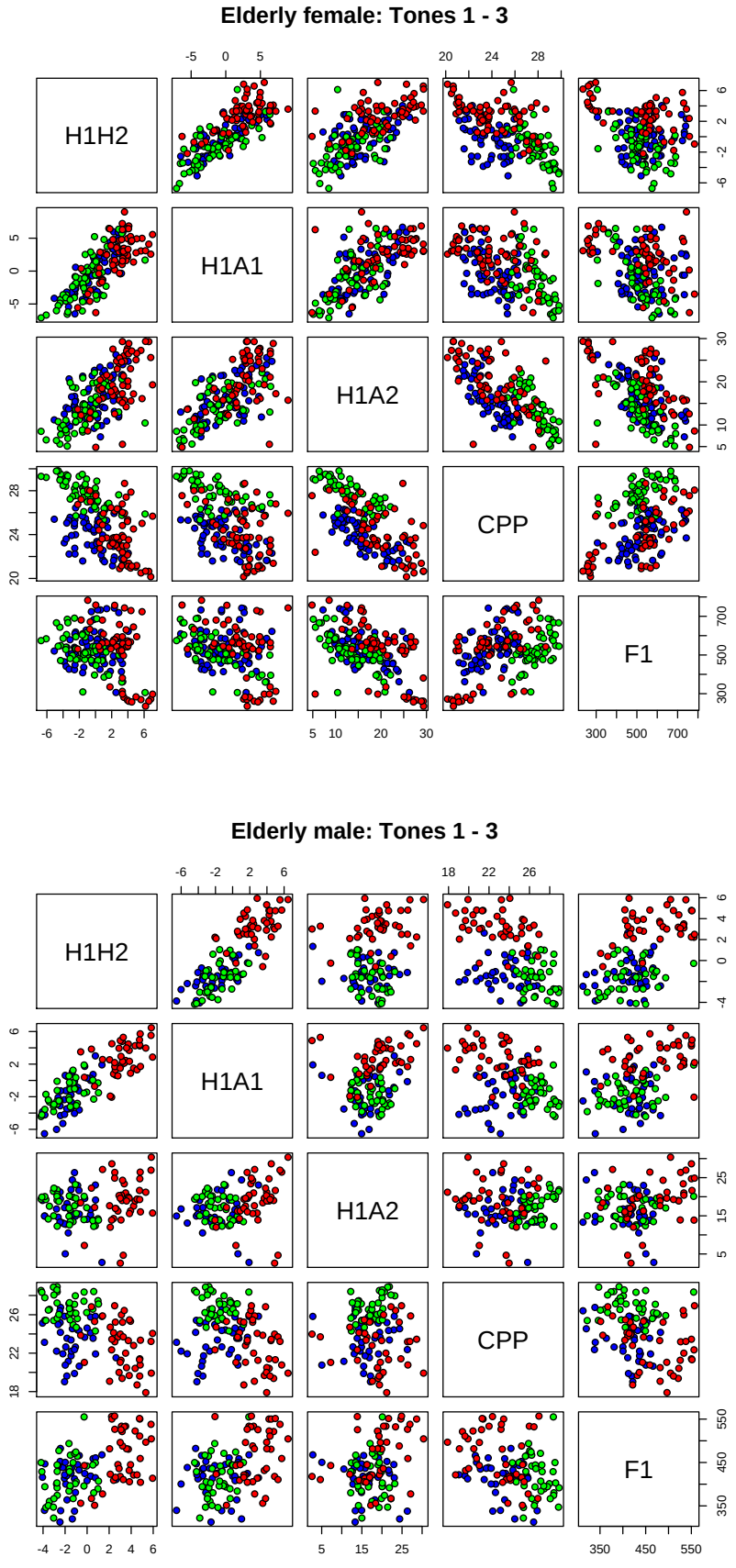
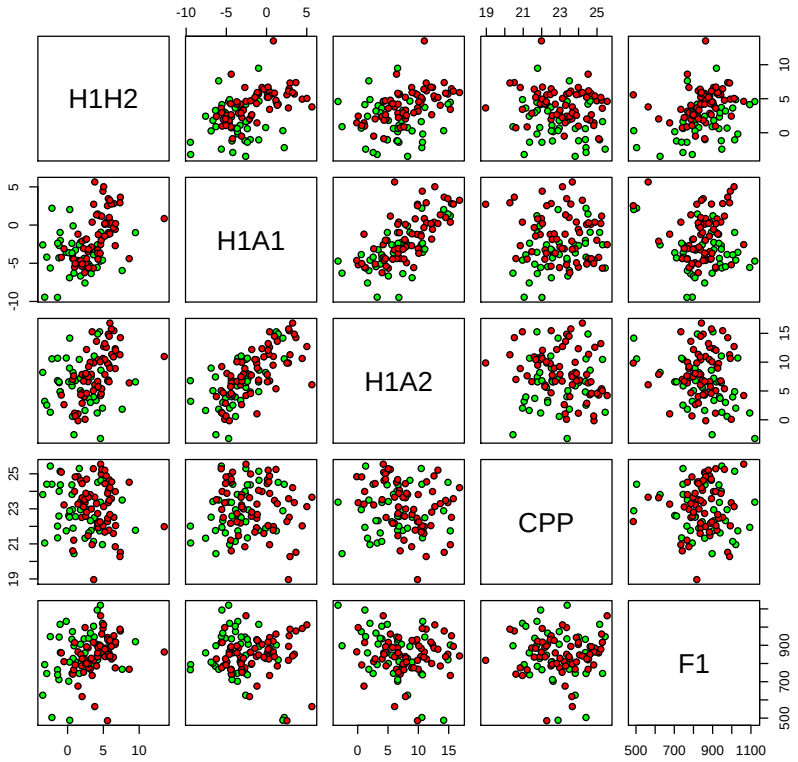
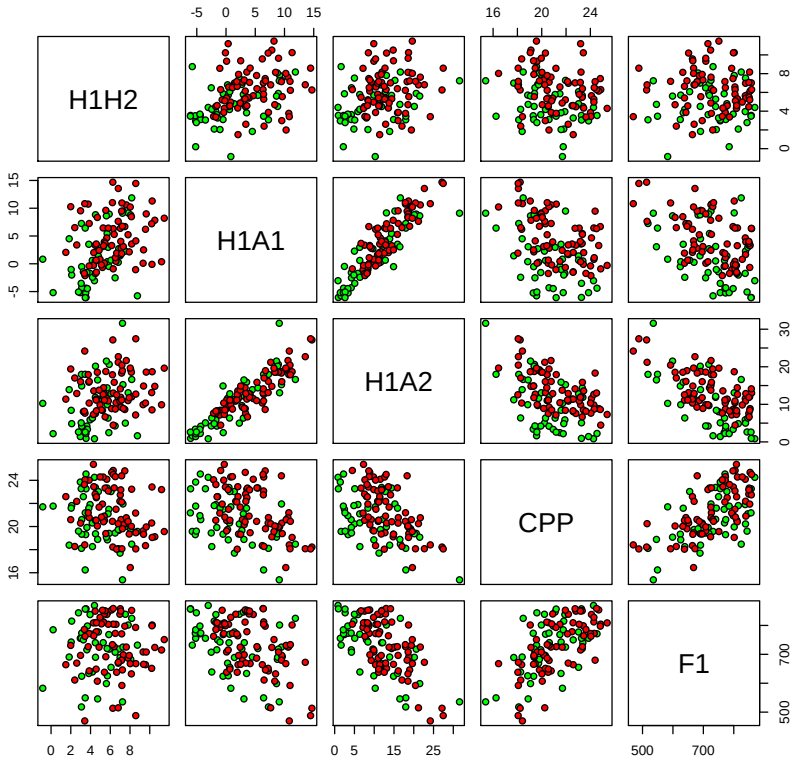


Figure 54. Scatterplot matrices for the three unchecked tone categories for each speaker group: blue for T1, green for T2, and red for T3.

Young female: Tones 4 - 5



Young male: Tones 4 - 5



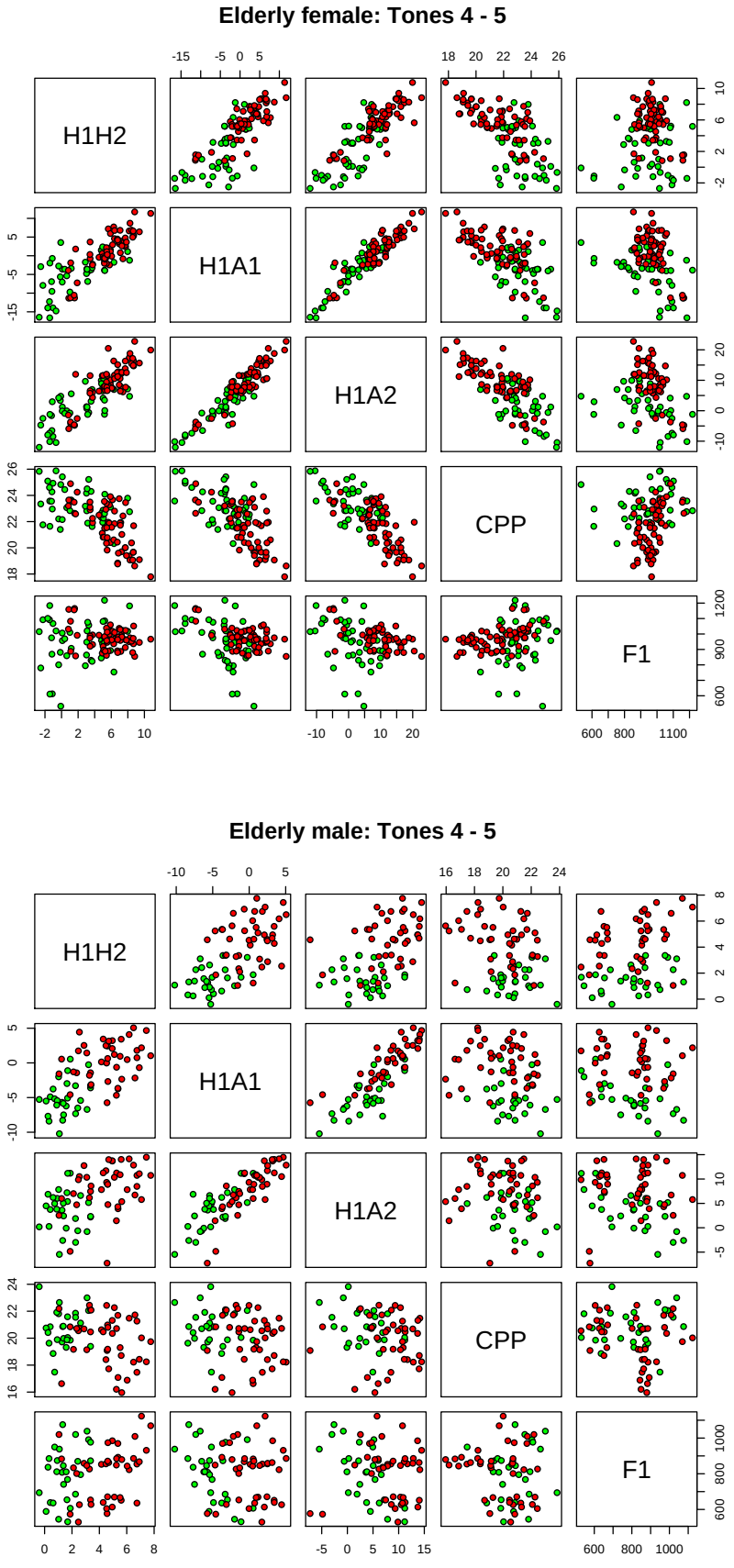


Figure 55. Scatterplot matrices for the two checked tone categories for each speaker group: green for T4 and red for T5.

4.1.6.3.4 Vocalic part of the first syllable in disyllables (S1 context)

In the preceding section on monosyllables, H1–H2 was found to be the most robust measure to distinguish breathy from modal phonation in Shanghai Chinese. In consequence, we only report this measure for the disyllabic contexts.

Figure 56 shows the average H1–H2 values of the vowel in the S1 context, computed with Praat, pooled across speaker groups and onset types, across all five time points for unchecked vowels and three time points for checked vowels. For checked syllables, in the S1 context just like in monosyllables, H1–H2 is higher for T5 than for T4 vowel with an average difference of about 2.3 dB. For unchecked syllables, H1–H2 is highest for T1 vowel, and higher for T3 than T2 vowel. The high value for T1 vowel is probably due to the nature of the following segment. In the two words 担心 [tɛ.ɕiŋ] and 三鲜 [sɛ.ɕi], with S1 in tone T1, /ɛ/ is followed by the fricative [ɕ], whose friction might make it breathier, whereas the /ɛ/ of S1 syllables in tone T2 are only followed by stops, affricates or sonorants.

Figure 57 (unchecked syllables) and Figure 58 (checked syllables) further show the H1–H2 at each time point (*yang* tones in muddy color) for (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers.

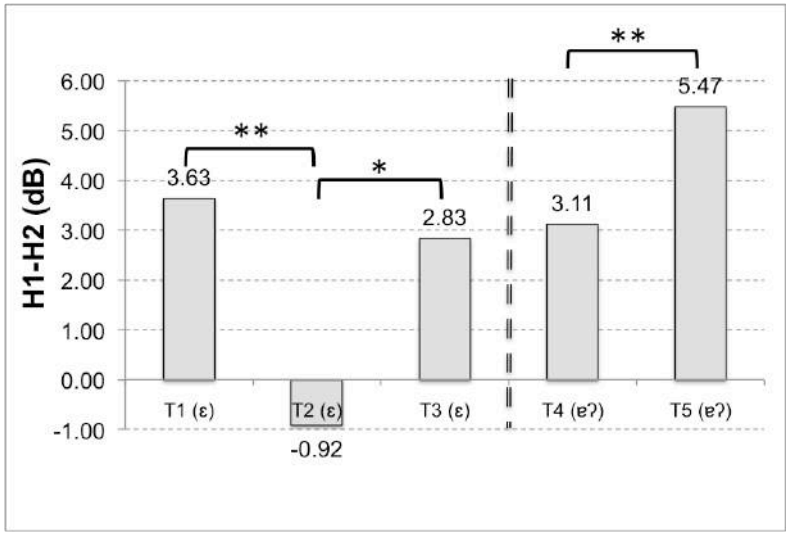


Figure 56. Average H1-H2 values in the S1 context according to underlying tone, pooled across onset types, time points, and speaker groups.
Significance levels: ** for $p<.01$.

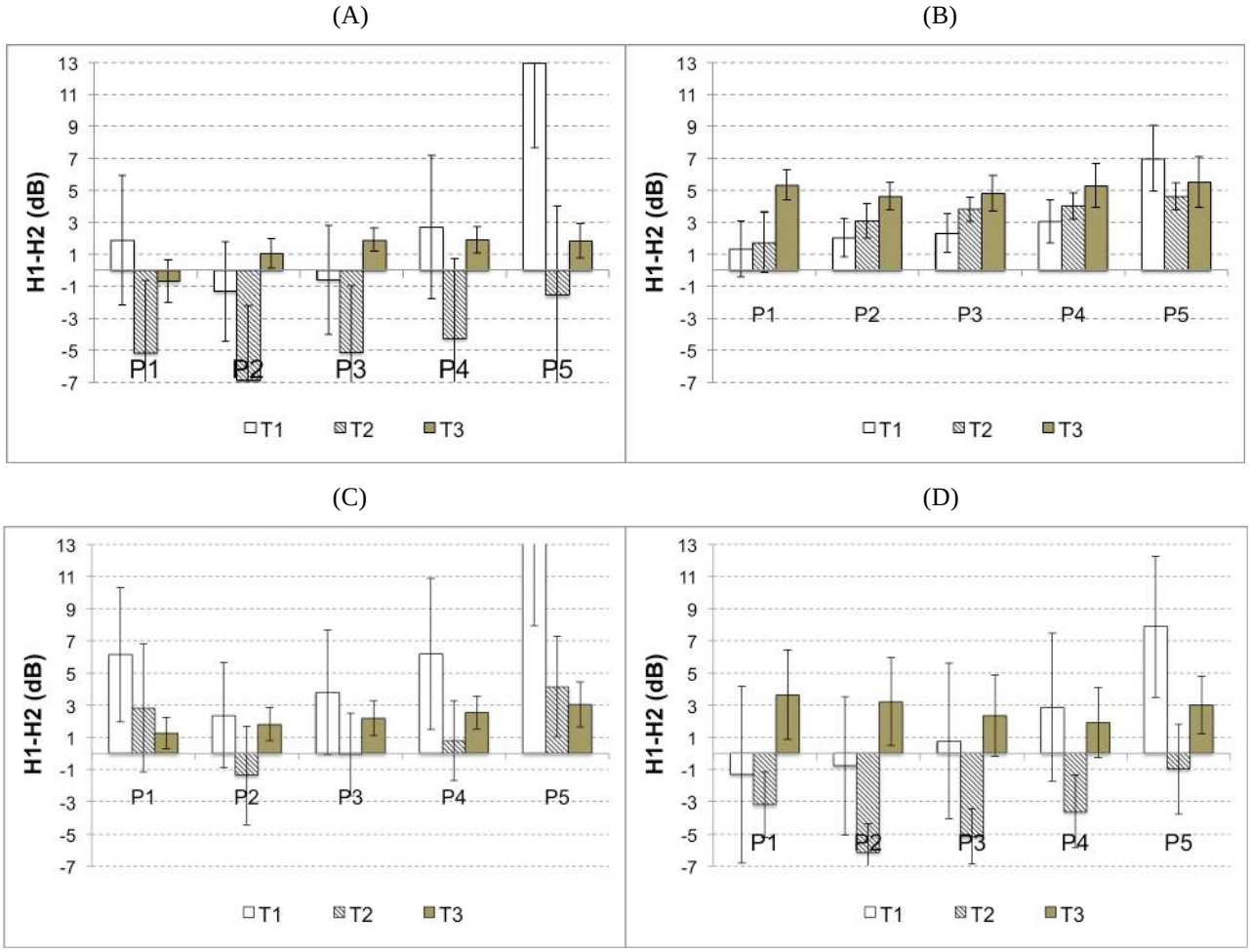
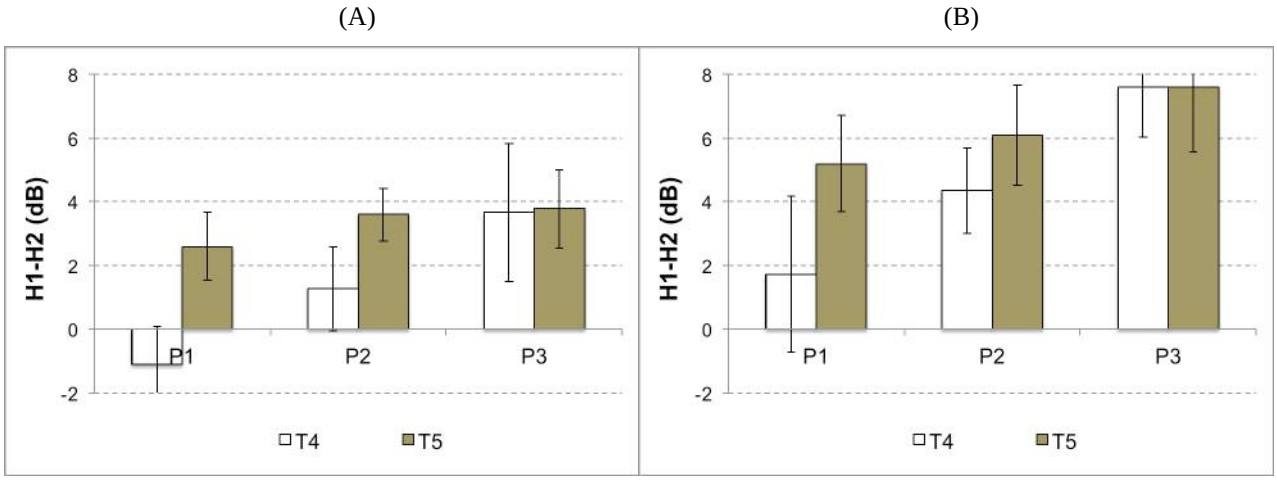


Figure 57. Average H1–H2 values for T1-3 unchecked vowel at five time points (P1-P5) in S1 according to underlying tone. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.



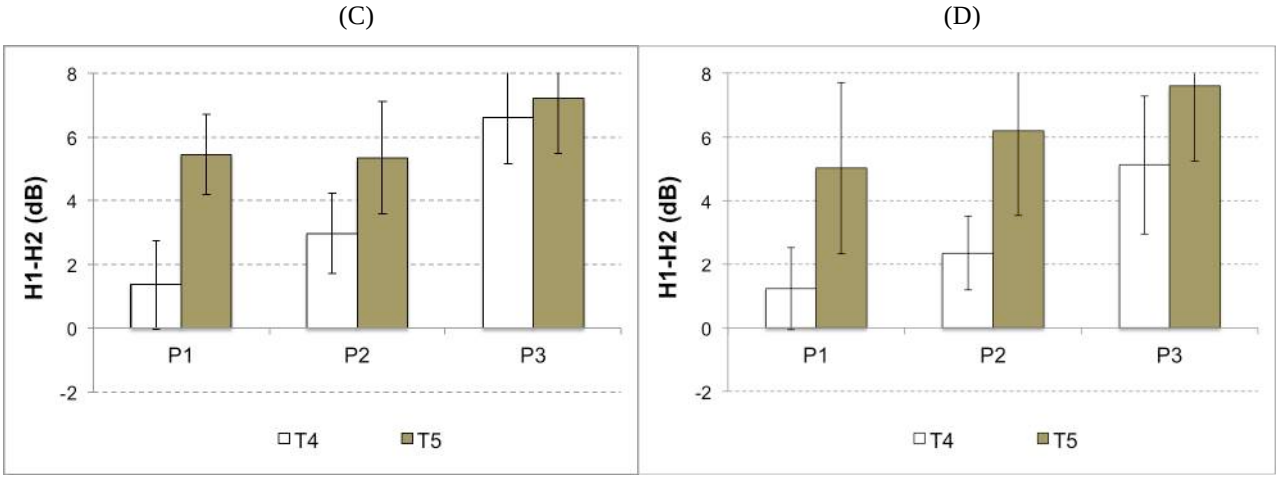


Figure 58. Average H1-H2 values for checked S1 syllables (T4-5) at five time points (P1-5), according to underlying tone: (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers.

We ran two by-subject ANOVAs separately for the unchecked (T1-3) syllable and the checked (T4-5) syllable data. In both analyses, *H1–H2* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. Underlying *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5), *Manner* of articulation (stop vs. fricative), *Place* of articulation (labial vs. dental) and *Time point* (unchecked: P1, P2, P3, P4, P5; checked: P1, P2, P3) were within-subject factors.

For unchecked syllables, *Tone* was globally significant, $F(2,36)=7.8$, $p<.005$. *H1–H2* was higher for T3 than T2 syllables (2.8>–0.9 dB, $F(1,18)=8.4$, $p<.01$), but numerically lower for T3 than T1 syllables (2.8 vs. 3.6 dB, $F(1,18)<1$). Limited to T1-2 vs. T3 difference, neither the *Tone* × *Place* ($F(1,18)=2.6$, $p=.12$) nor *Tone* × *Manner* ($F(1,18)<1$) interaction was significant. But the *Tone* × *Time point* interaction was, $F(4,72)=36.9$, $p<.0001$, reflecting larger T3 vs. T1-2 differentials at the second and third time points (P2: Δ=3.5 dB; P3: Δ=2.7 dB) than the other time points (P1: Δ=1.5 dB; P4: Δ=1.4 dB; P5: Δ= –2.8 dB). Neither either *Gender* nor *Age* was significant, Fs$(1,18)<1$, n.s. Limited to T1-2 vs. T3 difference, *Tone* did not interact with either *Gender* ($F(1,18)=1.2$, $p=.29$) or *Age* ($F(1,18)<1$).

For checked syllables, *Tone* was highly significant, $F(1,18)=25.4$, $p<.0005$: *H1–H2* was higher for T5 than T4 syllables (5.5>3.1 dB). *Tone* did not interact *Manner*, $F(1,18)<1$, but the *Tone* × *Place* interaction was, $F(1,18)=5.9$, $p<.05$, reflecting larger

T5 vs. T4 differentials for labial (Δ=3.3 dB) than dental (Δ=1.2 dB) onsets. The *Tone* × *Time point* interaction was significant, $F(2,36)=18.1$, $p<.0001$, reflecting larger T5 vs. T4 differentials for time points P1 than P2, for P2 than P3 (P1: Δ=3.8 dB; P2: Δ=2.6 dB; P3: Δ=0.8 dB). Neither either *Gender* nor *Age* was significant (*Gender*: $F(1,18)=2.1$, $p=.17$; *Age*: $F(1,18)<1$). Finally, *Tone* did not interact with either *Gender* or *Age* Fs$(1,18)<1$.

4.1.6.3.5 Vocalic part of the second syllable in disyllables (S2 context)

Figure 59 shows the average H1–H2 values in the vowel of S2 context syllables, pooled across time points, speaker groups and onset types, according to the preceding tone of S1 and the underlying tone of S2. Figure 60 (unchecked syllable) further shows the H1–H2 at each of five time points and Figure 61 (checked syllable) at each of three time points (*yang* tones in muddy color) for (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers.

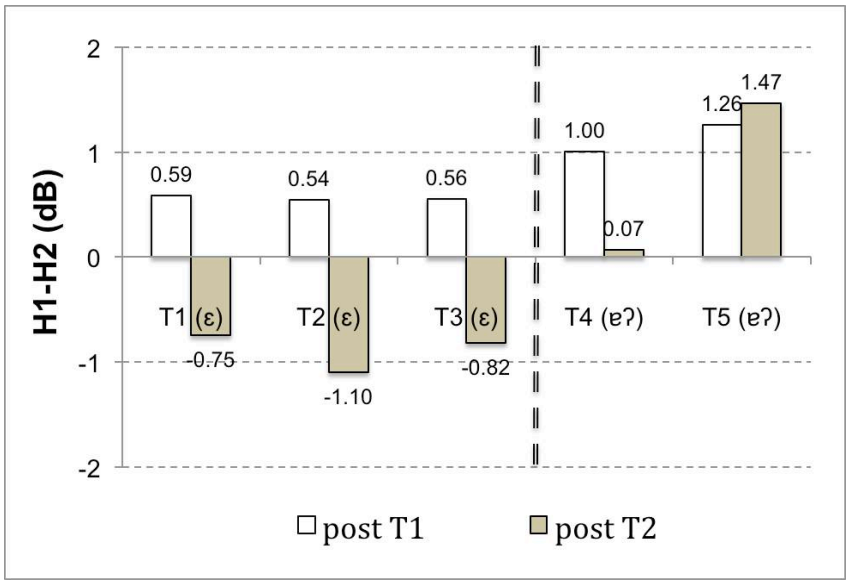


Figure 59. Average H1–H2 data in S2 syllables' rimes according to their underlying tone, after T1 vs. T2, pooled across onsets, time points, and speaker groups.

H1–H2 values were close to zero for each context. In contrast with the monosyllabic and S1 contexts, H1–H2 in the S2 context did not depend on the

underlying tone. Rather, it was affected by the preceding tone. For unchecked syllables, H1–H2 was higher after T1 than after T2. The H1–H2 differential, however, was quite small (<2 dB,. The detailed data (with the following additional factors: onsets, speaker groups, and time points) are reported in Appendix 2, but for unchecked syllables only, because neither underlying tone nor sandhi pattern affected H1–H2 in checked syllables.

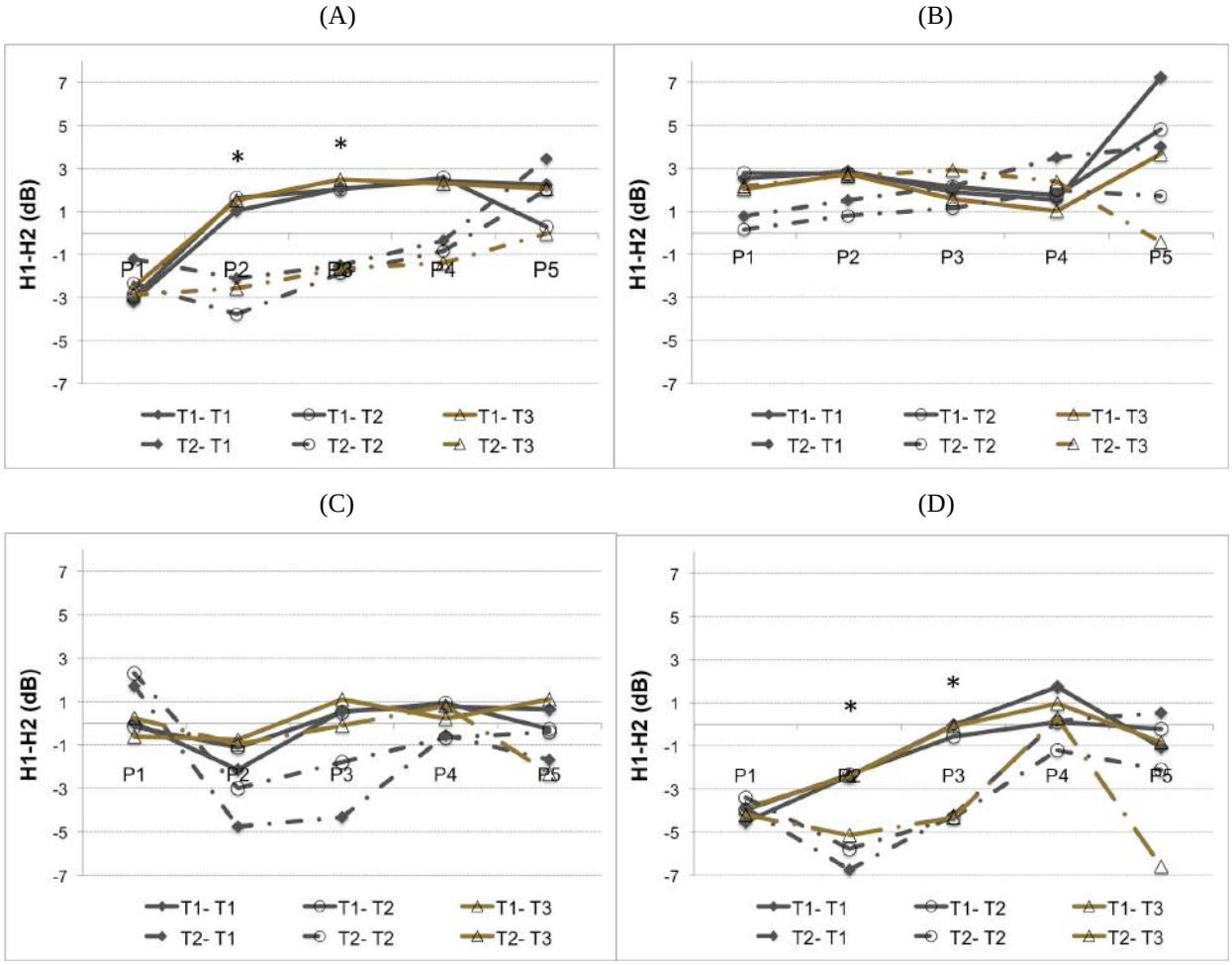


Figure 60. Average H1–H2 values in S2 unchecked syllables （T1-3） at five time points (P1-P5) according to preceding and underlying tone. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers. '*' ($p<.05$) shows higher values after T1 than T2.

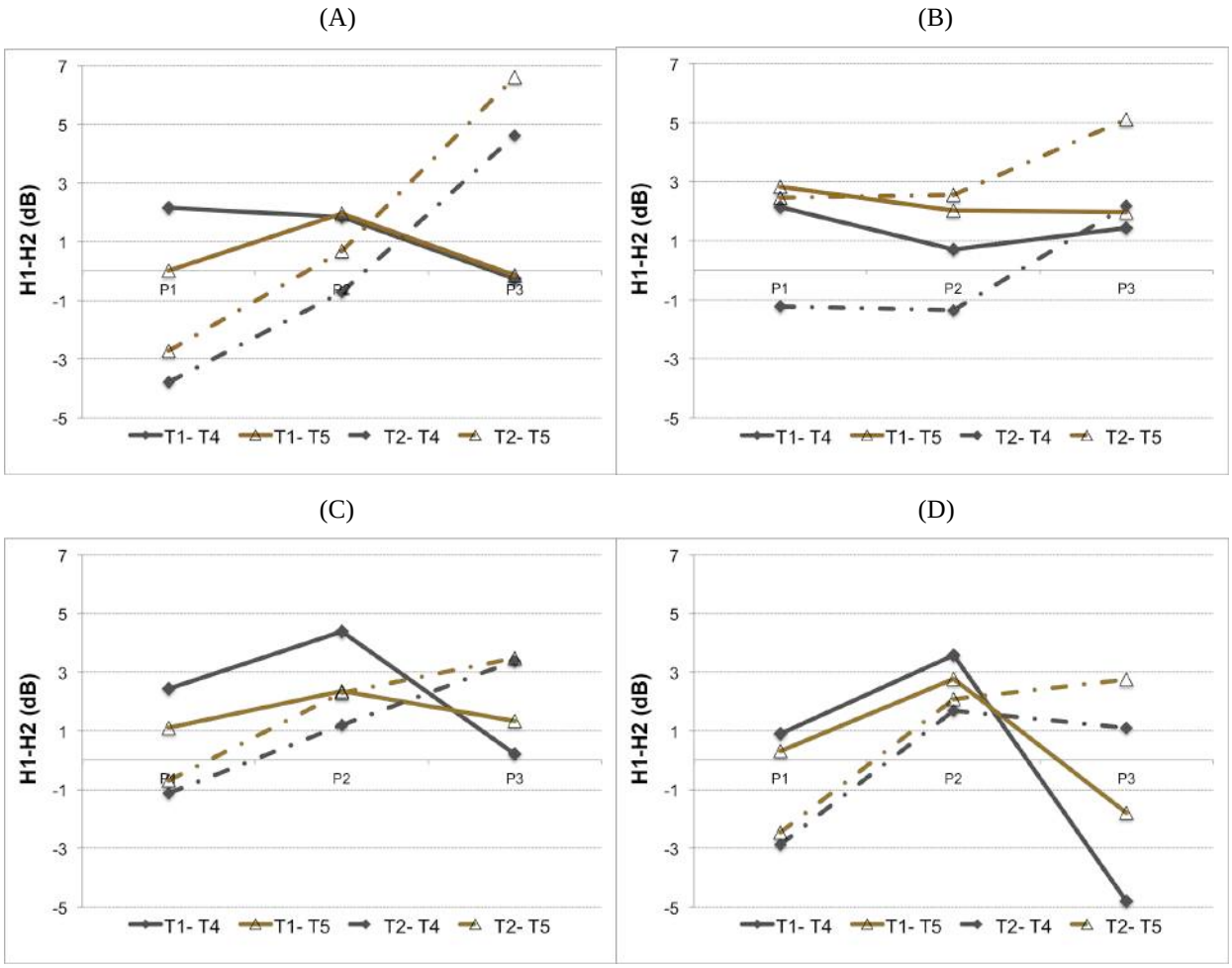


Figure 61. Average H1–H2 values in S2 checked syllables （T4-5） at five time points (P1-P3) according to preceding and underlying tone. (A) young female; (B) young male; (C) elderly female; (D) elderly male speakers.

To substantiate these observations, we ran two by-subject ANOVAs for the unchecked (T1-3) and checked (T4-5) syllable data. In both analyses, *H1–H2* was the dependent variable, and *Subject* was the random factor; *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Manner* of articulation (stop vs. fricative in both analyses), *Place* of articulation (labial vs. dental in both analyses), *Sandhi pattern* (post-T1 vs. post-T2 in both analyses), *Tone* (unchecked syllables: T1, T2, T3; checked syllables: T4 vs. T5) and *Time point* (unchecked syllables: P1, P2, P3, P4, P5; checked syllables: P1, P2, P3) were within-subject factors.

For unchecked syllables, *Sandhi pattern* was significant overall, $F(1,18)=5.0$, $p<.05$. *H1–H2* was higher for post-T1 than post-T2 sandhi (0.56>–0.89 dB). *Manner*

was also significant, $F(1,18)=23.1$, $p=.0001$. Overall, *H1–H2* was higher for fricative-onset syllables than for stop-onset syllables (0.50>–0.44 dB). Neither *Tone* nor *Place* was significant, Fs(1,18)<1. The *Sandhi pattern* × *Time point* interaction was significant, $F(4,72)=2.9$, $p<.05$: *Sandhi pattern* was significant only for Time points 2 and 3 (P2: Δ=2.62 dB; P3: Δ=2.63 dB, ps<.05), but not for the others (P1: Δ= –0.29 dB; P4: Δ=1.23 dB; P5: Δ=1.71 dB, Fs(1,18)<2, n.s.). *Gender* was not significant, $F(1,18)<1$, and *Age* was marginal, $F(1,18)=4.1$, $p=.059$. Finally, *Sandhi pattern* did not interact with either *Gender* or *Age*, Fs(1,18)<1.

For checked syllables, among all the within-subject factors, only *Time point* was significant, $F(2,36)=5.5$, $p<.01$. Overall, *H1–H2* was higher at the last two Time points than the first one (P1: -0.06 dB; P2: 1.81 dB; P3: 1.80 dB). None of the other within-subject factors was significant, Fs<1. Neither *Gender* nor *Age* was significant, Fs(1,18)<1.

4.1.6.3.6 Discussion

In this section, we investigated the phonation difference between *yin* and *yang* syllables. We used the HNR measure in the release part of word-initial stops. For the vocalic part in the monosyllabic context, we used several spectral tilt measures (i.e., H1–H2, H1–A1 and H1–A2), the CPP measure, and the F1 measure. In the S1 and S2 contexts, we limited ourselves to the most discriminant measure, H1–H2.

During the release of word-initial stops, HNR was lower for *yang* than *yin* stops, indicating that the *yang* stops are noisier, and the noise is probably related to the breathy phonation. This pattern was found in both the monosyllabic and the S1 contexts, and for all speaker groups. Worth to note, the *yin–yang* difference was greater overall for the elderly than young speakers. No difference was found, however, for nasal stops.

In the vocalic part of monosyllables, all spectral tilt measures showed overall higher (and positive) values for *yang* than *yin* vowels, indicating steeper spectral tilt, hence more breathiness in *yang* than *yin* syllables. The *yin–yang* difference was larger at rime beginning, then decreased but generally lasted at least until the midpoint of the rime, then disappeared at rime offset. This is not exactly what has been reported

in Cao and Maddieson's (1992) or Ren's (1992) studies, according to which the *yin–yang* difference is restricted to the beginning of the rime. Moreover, the *yin–yang* difference was, again, greater overall for elderly than young speakers, on most measures excepted H1–A1 and H1–A2 in unchecked vowels. Besides, the H1–A1 measure in unchecked vowels showed greater differences for male than female speakers. The CPP measure was a good indicator of breathiness for elderly speakers, but not for young speakers. CPP was lower for *yang* than *yin* vowels only for elderly speakers, and the *yin–yang* difference lasted until the midpoint of the rime. The F1 measure showed a global difference between *yin* and *yang* unchecked vowels, but not in the direction that was described in the literature (e.g., Thongkum, 1988).

Linear discriminant analyses (LDA) were conducted on monosyllables firstly in order to compare the overall efficiency of the combined five acoustic measures across the speaker groups, and secondly to determine the relative importance of the five phonation measures we used in the discrimination between *yin* and *yang* syllables, for each speaker group. The discriminant analyses with the combined five measures were the most efficient for elderly male, followed by elderly female, then young male, and lastly young female speakers. The most important acoustic measure, common to all speaker groups, and the most important in most contexts, turned out to be the widely used H1–H2 measure.

H1–H2 was thus used as a measure of breathy phonation for the vowels in the S1 and S2 contexts. In the S1 context, the *yin–yang* difference was somewhat smaller than in the monosyllabic context. For the vowels of checked syllables, this difference maintained over the entire vowel. Two explanations for this are possible: a physiological versus a phonological explanation. The short duration of S1 checked syllables might make it difficult for voice quality to change for physiological reasons. On a phonological viewpoint, the time domain of breathy voice might be determined by the duration of the prosodic word. In *yang* monosyllables, breathy voice is maintained until the midpoint of the rime, in the first half of the (monosyllabic) prosodic word, whereas in *yang* S1 syllables, breathy voice is maintained also in the first half of the (disyllabic) prosodic word, which is the entire first syllable in this case. If the second explanation is correct, the same pattern should be observed in unchecked syllables. However, in our data, the H1–H2 difference for unchecked

syllables was not significant overall. It was found significant only for the first two time points. But we cannot exclude the influence of the friction noise of the following consonant in some *yin* S1 contexts. Therefore, more data with better controlled segmental contexts for S1 syllables are needed to further examine the physiological and phonological explanations.

In the S2 context, the underlying *yin* or *yang* syllables can hardly be qualified as breathy, since for both, H1–H2 values are located near zero. H1–H2 did not differ significantly between the underlyingly *yin* and *yang* syllables. There were slight differences, however, between the post-T1 and post-T2 sandhi patterns. H1–H2 was slightly larger overall in post-T1 than post-T2 contexts, that is, larger for the lower F0 contour following T1 than T2. We may conclude that the breathy phonation cooccuring with word-initial *yang* syllables is absent in non-initial syllables.

Finally, all our measurements consistently showed that Shanghai male speakers produce breathier voice than female speakers, contrary to most the current findings in English on voice quality according to gender (British English: Henton & Bladon, 1985; American English: Klatt & Klatt, 1990; Hanson & Chuang, 1999). More investigations are needed to find out whether the difference we found is related to the inherent voice qualities of Shanghai male vs. female speakers or, for example, to reading style. Possibly, Shanghai male speakers have a less careful reading style compared to female speakers (as suggested, for example, by shorter syllable durations (see §4.1.6.4)) thereby using more effortless, slacker articulation, in particular for laryngeal articulation.

4.1.6.4 Durations

We measured the acoustic duration of onsets and rimes. Rime duration was measurable in all contexts, as well as the duration of fricative and nasal onsets. The closure duration of stop onsets in word-initial position was, however, impossible to measure in the absence of voicing (i.e., glottal pulsing) during stop closure, which is the rule for all the Shanghai Chinese stops utterance-initially. Based on the acoustic data, we therefore can only report stop closure duration in word-medial position. Preliminary results on stop closure duration based on visible articulation are reported in §4.3.

4.1.6.4.1 Monosyllables

4.1.6.4.1.1 Onset duration

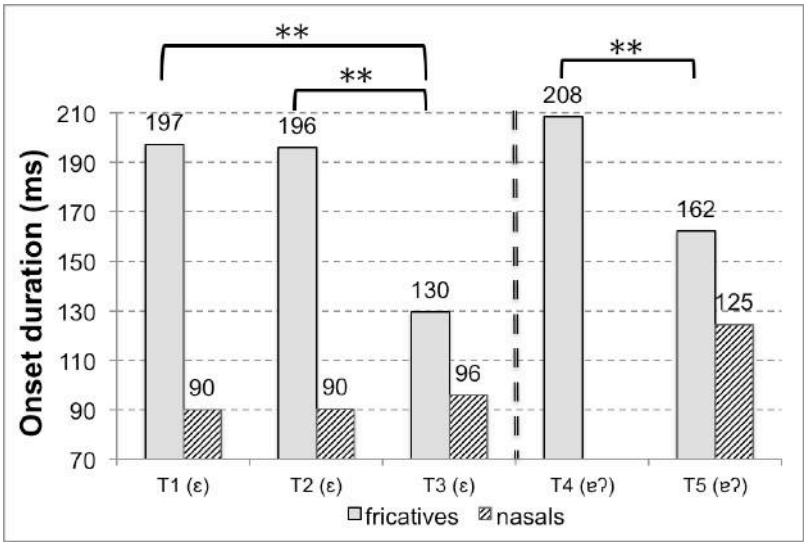


Figure 62. Average onset durations for fricative and nasal onsets in monosyllables as a function of tone. Significance level (for fricatives only): ** for $p<.001$.

Figure 62 shows the average onset durations as a function of the syllable tone in monosyllabic context, for fricatives and nasals, across all speaker groups. Note that the /nɛ/ syllable does not co-occur with T2 and /maʔ naʔ/ syllables do not co-occur with T4. Details are listed in Appendix 2.

Since nasal onset durations virtually do not vary according to syllable tone, Figure 63 only shows the average fricative onset durations (A) for young speakers and (B) for elderly speakers, with separate analyses for unchecked and checked syllables.

To summarize the main trends of the statistical analyses run on the data, we find that, for all speaker groups, T3 (*yang*) fricative onsets were significantly shorter overall than T1 and T2 (*yin*) onsets. Likewise, T5 (*yang*) fricative onsets were shorter overall than T4 (*yin*) onsets, but the difference was significant only for young speakers. No difference was found between T1 and T2 fricative onsets. Nasal onsets did not differ according to tone.

Yang fricative onsets were often realized as fully or partly voiced, as reported in §4.1.6.2. In order to eliminate the possible contribution of phonetic voicing to the *yin–yang* duration difference, we separately show in Figure 64 the fricative duration data restricted to the fricatives whose v-ratio is lower than 0.2, which may be considered as voiceless. This, however, resulted in very unbalanced data across tone categories. Hence, we did not run statistical analyses on these restricted data. As can be seen from Figure 63 and Figure 64, the duration patterns according to tone were very similar in the unrestricted vs. restricted data, especially for young speakers (except for greater variability in *yang* fricative durations in the restricted data). For the unchecked syllables of elderly speakers, a similar pattern of shorter *yang* than *yin* fricatives can also be seen in both the unrestricted and restricted data, although the durations of *yang* fricatives were longer and more variable overall in the restricted than unrestricted data.

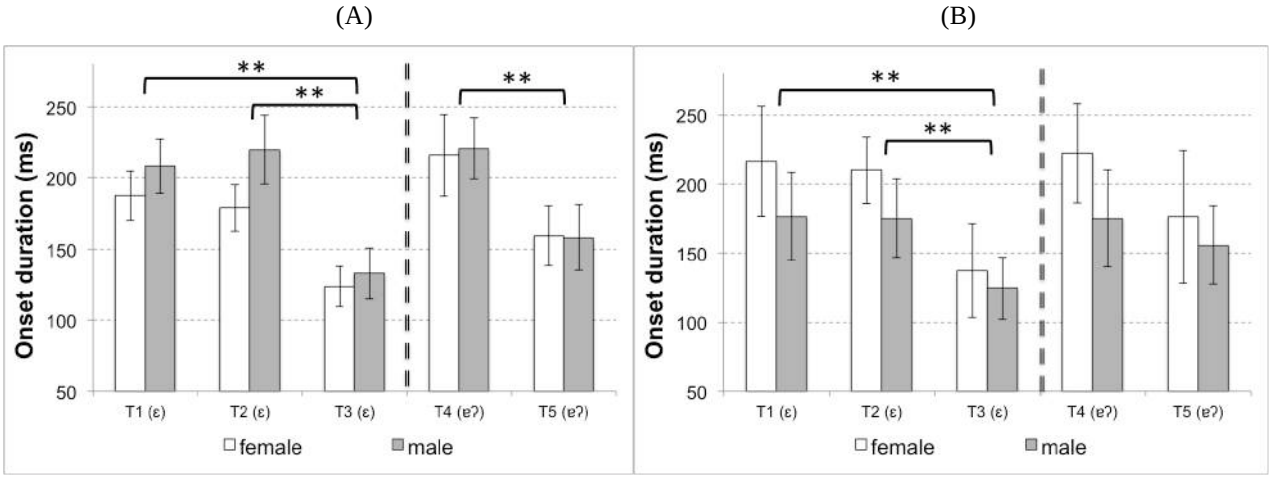


Figure 63. Average fricative onset durations in monosyllables according to Tone. (A) young speakers; (B) elderly speakers. Significance levels: * for $p<.05$; ** for $p<.01$.

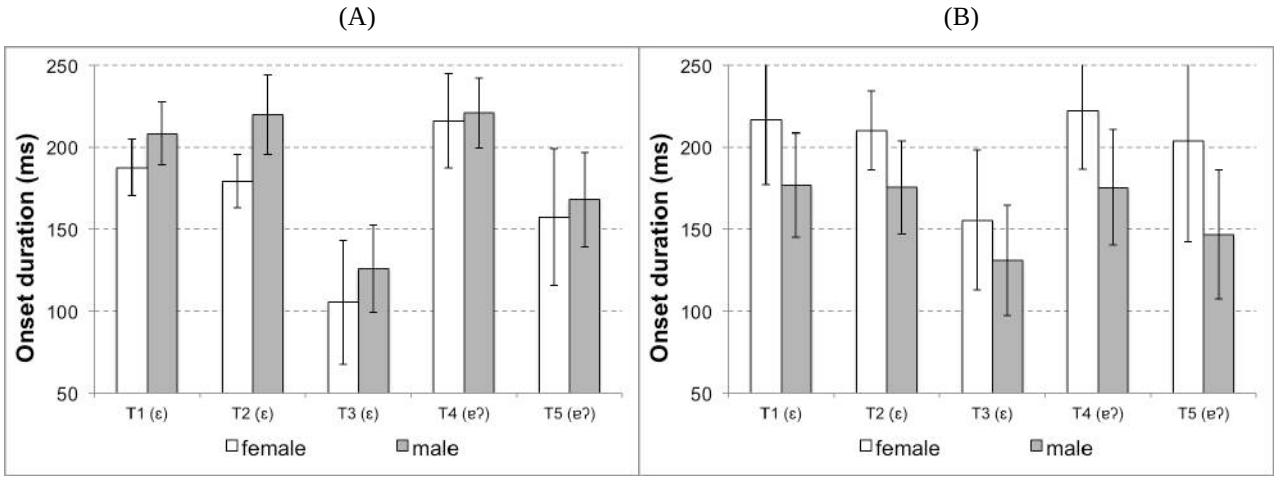


Figure 64. Average fricative onset durations in monosyllables, phonetically voiced fricatives excluded, according to Tone. (A) young speakers; (B) elderly speakers.

We ran separate by-subject ANOVAs for fricative onsets in unchecked (T1-3) syllables, for fricative onsets in checked (T4-5) syllables, and for nasal onsets in unchecked syllables. Recall that nasal-onset syllables were restricted to /mɛ/ in T2 unchecked syllables, and absent altogether from the checked syllable subset.. In all the analyses, *Duration* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Tone* (unchecked fricatives: T1, T2, T3; unchecked nasals: T1 vs. T3; checked fricatives: T4 vs. T5) and *Place* of articulation (labial vs. dental) were within-subject factors.

For fricatives in unchecked syllables, *Tone* was highly significant globally, $F(2,36)=58.8$, $p<.0001$: *Duration* was the shortest for *yang* fricatives (T3 vs. T1: 130<197 ms, $F(1,18)=65.8$, $p<.0001$; T3 vs. T2:, 130<196 ms, $F(1,18)=80.1$, $p<.0001$). No difference was found between T1 and T2 onsets $F(1,18)<1$. *Place* was significant, $F(1,18)=71.4$, $p<.0001$: *Duration* was longer for dental than labial fricatives (191>160 ms). The *Tone* × *Place* interaction was not significant, $F(2,36)=1.3$, $p=.29$. Neither *Gender* nor *Age* was significant, Fs(1,18)<1. Finally, *Tone* did not interact with either *Gender* or *Age*, Fs(2,36)<1.

For fricatives in checked syllables, *Tone* was significant globally, $F(1,18)=27.8$, $p<.0005$: *Duration* was shorter for T5 than T4 fricatives (162<208 ms). *Place* was significant, $F(1,18)=63.4$, $p<.0001$: *Duration* was, again, longer for dental than labial fricatives (214>161 ms). The *Tone* × *Place* interaction was not significant, $F(1,18)=3.4$, $p=.08$. Neither *Gender* nor *Age* was significant, Fs(1,18)<1. Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(1,18)<1$; *Tone* × *Age*: $F(1,18)=1.7$, $p=.21$).

For nasals in unchecked syllables, neither *Tone* nor *Place* was significant, Fs(1,18)<1. Neither *Gender* nor *Age* was significant, Fs(1,18)<1. *Tone* did not interact with *Gender*, $F(1,18)<1$, but the *Tone* × *Age* interaction was significant, $F(1,18)=5.8$, $p<.05$: *Tone* was not significant for young speakers, $F(1,10)<1$, but was significant for elderly speakers, $F(1,8)=11.2$, $p<.05$, with longer nasal onsets for T3 than T2 (115>82 ms), contrary to the pattern found for fricatives.

4.1.6.4.1.2 Rime duration

Figure 65 shows the average duration of /ɛ/ and /aʔ/ rimes as a function of tone in the monosyllabic context, pooled across all onset types and speaker groups. Figure 66 separately shows (A) the young speakers' and (B) the elderly speakers' data.

To summarize the main trends of the statistical analyses run on the data, the /ɛ/ rime (unchecked syllables) was the shortest for the falling *yin* tone T1 and the longest for the *yang* tone T3. The /aʔ/ rime (checked syllables) was significantly longer for T5 (*yang*) than T4 (*yin*). The difference between T2 and T3 /ɛ/ rimes, however, was less

clear-cut among young speakers. The /ɛ/ rimes were about twice as long as the /aʔ/ rimes. Furthermore, female speakers produced longer rimes than male speakers overall. Details are listed in Appendix 2.

We ran two by-subject ANOVAs separately for unchecked (T1-3) syllable and checked (T4-5) syllable data. In both analyses, *Duration* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) and *Onset* (unchecked: zero, /p/, /t/, /f/, /s/, /m/; checked: zero, /p/, /t/, /f/, /s/) were within-subject factors. (/m/ or /n/ in checked syllables and the /n/ in unchecked syllables did not appear as levels of the *Onset* factor, due to the onset-rime co-occurrence restrictions in our data.)

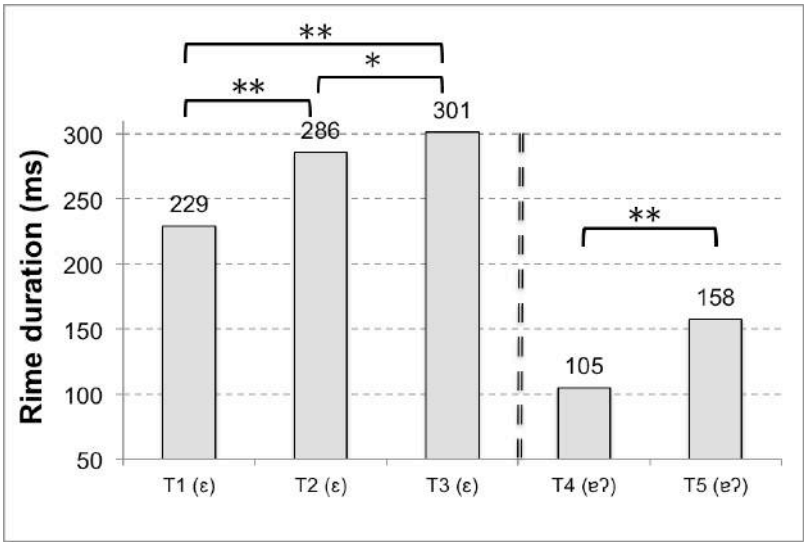


Figure 65. Average /ɛ/ and /aʔ/ rime durations in monosyllables as a function of tone, pooled across onset types and speaker groups. Significance levels: * for $p<.01$; ** for $p<.001$.

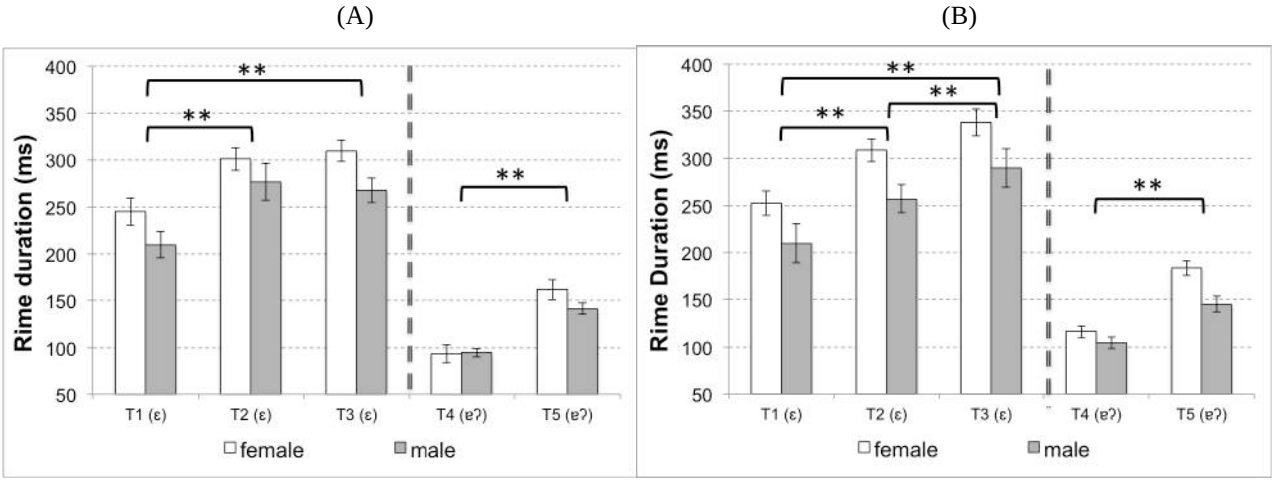


Figure 66. Average /ɛ/ and /aʔ/ rime durations in monosyllables as a function of tone, pooled across onset types. (A) young speakers and (B) elderly speakers.
Significance levels: * for $p<.05$; ** for $p<.01$.

For unchecked syllables, *Tone* was highly significant globally, $F(2,36)=71.9$, $p<.0001$: T1 rimes were significantly shorter than T2 rimes (229<286 ms, $F(1,18)=61.7$, $p<.0001$), which were significantly shorter than T3 rimes (286<301 ms, $F(1,18)=12.9$, $p<.005$). The *Onset* factor was significant, $F(5,90)=23.4$, $p<.0001$, due to longer rimes after zero onset than other onsets (zero: 301 ms; /p/: 272 ms; /t/: 276 ms; /f/: 269 ms; /s/: 263 ms; /m/: 265 ms). The *Tone* × *Onset* interaction, within the T2 and T3 data subset, was significant, $F(5,90)=3.6$, $p<.01$. T3 rimes were significantly longer than T2 rimes for the zero onset (Δ=35 ms), $F(1,18)=14.8$, $p<.005$, for the /p/ (Δ=24 ms) and /s/ onsets (Δ=12 ms), ps<.05, but not for the /t/ (Δ=–4 ms), /f/ (Δ=15 ms), and /m/ (Δ=8 ms) onsets. *Gender* was significant, $F(1,18)=6.3$, $p<.05$: female speakers produced longer rime durations than male speakers (292>252 ms). *Age* was not significant (young vs. elderly: 268≈276 ms), $F(1,18)<1$, and did not interact with *Gender*, $F(1,18)<1$. Finally, within the T2 and T3 data subset, *Tone* did not interact with *Gender*, $F(1,18)=1.5$, $p=.24$, but interacted with *Age*, $F(1,18)=9.0$, $p<.01$: T2 and T3 rimes did not differ for young speakers (289=289 ms), $F(1,10)<1$, but differed significantly for elderly speakers, (T3=314>T2=283 ms), $F(1,8)=32.4$, $p=.0005$.

For checked syllables, *Tone* was highly significant overall, $F(1,18)=492.1$, $p<.0001$: T4 rimes were significantly shorter than T5 rimes (105<158 ms). *Tone* was highly significant as well for each of the five onsets, with T5–T4 differences ranging from 45 to 68 ms, ps<.0001. *Onset* was significant, $F(4,72)=24.5$, $p<.0001$: rimes were

longer with the zero onset than the other onsets, and rimes were slightly longer with fricative than stop onsets (zero: 155 ms; /p/: 126 ms; /t/: 125 ms; /f/: 136 ms; /s/: 133 ms). The *Tone* × *Onset* interaction was significant, $F(4,72) = 5.0$, $p<.005$, indicating slightly variable —although all highly significant— T5–T4 differences across onsets (/p/: Δ=55 ms; /t/: Δ=67 ms; zero: Δ=65 ms; /f/: Δ=45 ms; /s/: Δ=68 ms). *Gender* was significant, $F(1,18)=10.4$, $p<.005$: female speakers produced longer rime durations than male speakers (141>122 ms). *Age* was marginal, $F(1,18)=4.3$, $p=.054$: elderly speakers produced longer rime durations than young speakers (137>126 ms). *Tone* did not interact with *Age*, $F(1,18)<1$, but interacted with *Gender*, $F(1,18)=16.3$, $p<.001$, reflecting larger T5 vs. T4 differentials for female (Δ=63 ms, $F(1,8)=312.5$, $p<.0001$) than male speakers (Δ=43 ms, $F(1,8)=172.0$, $p<.0001$).

4.1.6.4.2 First syllable in disyllables (S1)

4.1.6.4.2.1 Onset duration

Figure 68 shows the average durations of fricative onsets in the S1 context (first syllable of disyllabic words) as a function of S1's underlying tone, across all speaker groups. Nasal-onset syllables were not recorded in the S1 context. Figure 69 separately shows (A) the young speakers' and (B) the elderly speakers' data.

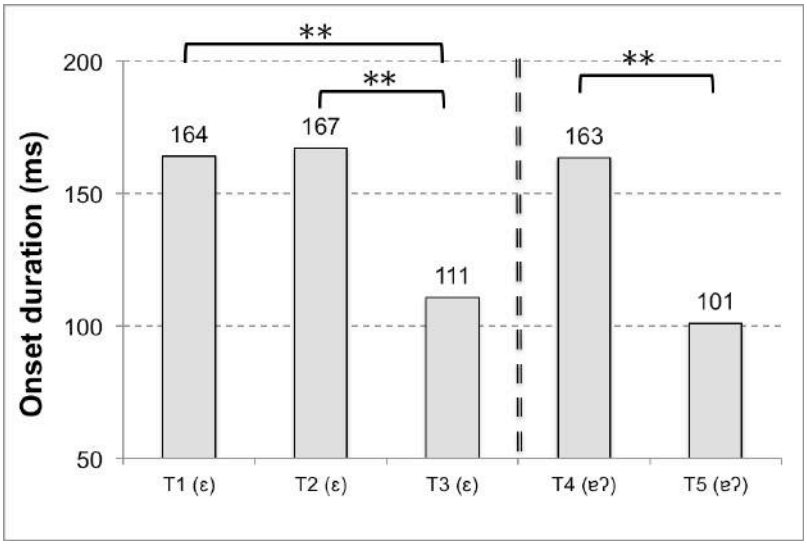


Figure 67. Average durations of fricative onsets in the S1 context according to S1's tone. Significance levels: * for $p<.01$; ** for $p<.001$.

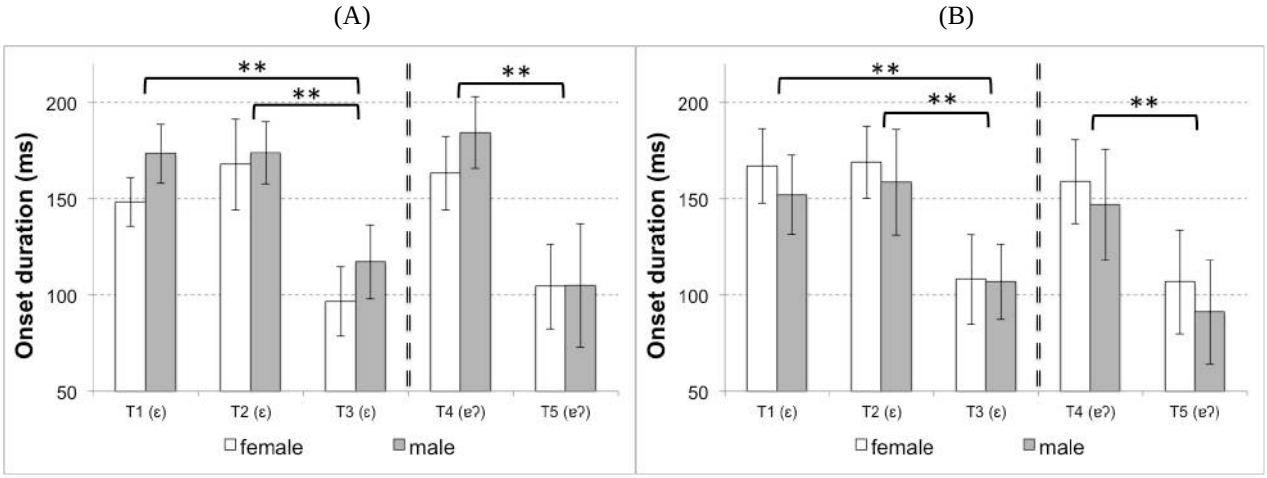


Figure 68. Average durations of fricative onsets in the S1 context according to S1's tone and speaker group. (A) young speakers; (B) elderly speakers. Significance levels: ** for $p<.01$.

To summarize the main trends of the statistical analyses run on the data, for both age and gender groups, T3 (*yang*) syllable onsets were significantly shorter than T1 or T2 (*yin*) syllable onsets, and T5 (*yang*) syllable onsets were significantly shorter than T4 (*yin*) syllable onsets. Moreover, the average onset duration was shorter in the S1 than the monosyllabic context. Details are reported in Appendix 2.

As for monosyllables, in order to estimate the contribution of phonetic voicing to the *yin–yang* duration difference, we further show, in Figure 69, the durations of the fricatives whose v-ratio was lower than 0.2 (which we consider as voiceless). Statistics were not conducted on this data subset because it was unbalanced with respect to the number of items in the different categories. The rimes in this data subset were longer and more variable in duration than for the entire data set, but they were still shorter for *yang* than *yin* syllables.

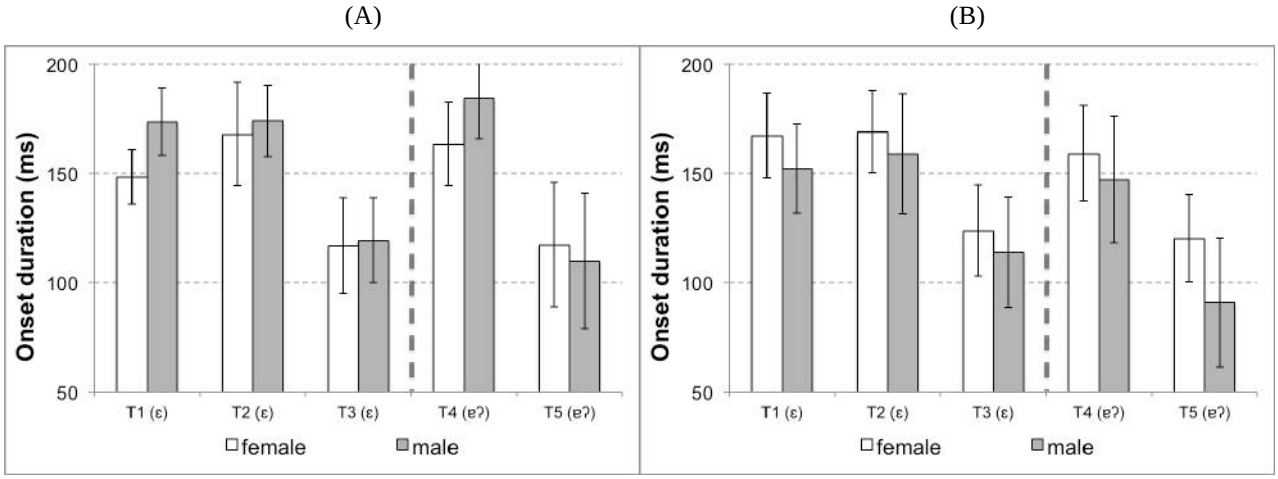


Figure 69. Average durations of fricative onsets (S1 context), phonetically voiced fricative onsets excluded, according to S1's tone. (A) young speakers; (B) elderly speakers.

We ran two by-subject ANOVAs separately for the unchecked (T1-3) and the checked (T4-5) syllable data. In both analyses, *Duration* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. Underlying *Tone* of S1 (unchecked: T1, T2, T3; checked: T4 vs. T5) and *Place* of articulation (labial vs. dental) were within-subject factors.

For unchecked syllables, *Tone* was highly significant overall, $F(2,36)=71.0$, $p<.0001$: The (fricative) onsets were significantly shorter for T3 than T1 (111<164 ms), $F(1,18)=95.6$, $p<.0001$, or T2 (111<167 ms), $F(1,18)=81.8$, $p<.0001$. The difference between T1 and T2 was not significant, $F(1,18)=2.9$, $p=.11$. *Place* was significant, $F(1,18)=80.9$, $p<.0001$: *Duration* was longer for dental than labial fricatives (159>132 ms). The *Tone* × *Place* interaction was significant, $F(2,36)=9.7$, $p<.0005$, reflecting larger T1-2 vs. T3 differentials for labial (Δ=71 ms), $F(2,36)=58.2$, $p<.0001$, than dental fricatives (Δ=43 ms), $F(2,36)=39.2$, $p<.0001$. Neither *Gender* nor *Age* was significant, Fs(1,18)=<1. *Tone* did not interact with either *Gender* or *Age*, Fs(2,36)<1.

For checked syllables, *Tone* was again highly significant, $F(1,18)=227.0$, $p<.0001$: fricative onsets were significantly shorter for T5 than T4 (101<163 ms). *Place* was significant, $F(1,18)=86.7$, $p<.0001$: *Duration* was longer for dental than labial fricatives (153>109 ms). The *Tone* × *Place* interaction was significant, $F(1,18)=8.5$, $p<.01$, reflecting larger T1-2 vs. T3 differentials for labial (Δ=75 ms), $F(1,18)=106.8$, $p<.0001$, than dental fricatives (Δ=38 ms), $F(1,18)=26.3$, $p=.0001$. Neither *Gender* nor

Age was significant, *F*s(1,18)<1. *Tone* did not interact with either *Gender* or *Age*, *F*s(1,18)<1.

4.1.6.4.2.2 Rime duration

Figure 70 shows the duration of /ɛ/ and /aʔ/ rimes of S1 according to S1's underlying tone in disyllabic words, averaged across onset types and speaker groups. Figure 71 separately shows (A) the young speakers' and (B) the elderly speakers' data.

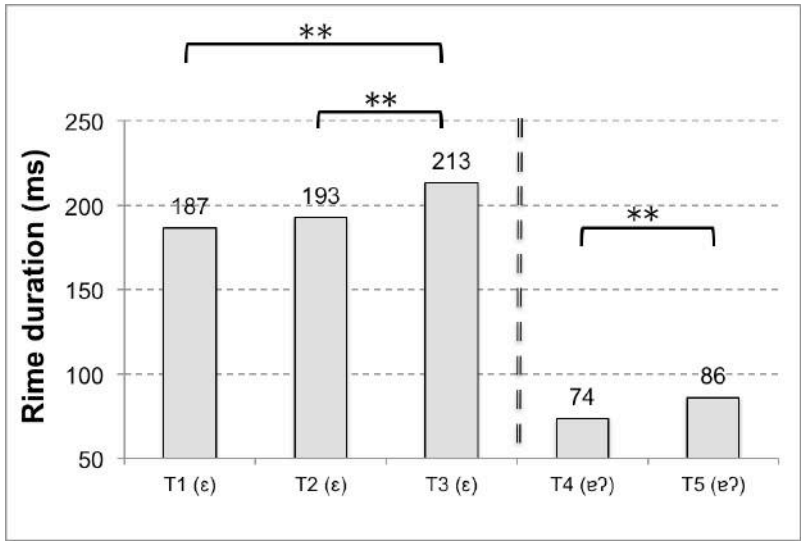


Figure 70. Average /ɛ/ and /aʔ/ rime durations in S1, as a function of S1's underlying tone, pooled across onset types and speaker groups. Significance levels: * for p<.01; ** for p<.001.

To summarize the statistical results, /ɛ/ rimes were significantly longer overall for T3 (*yang*) than T1 (*yin*) or T2 (*yin*); likewise, /aʔ/ rimes were significantly longer for T5 (*yang*) than T4 (*yin*); /ɛ/ rimes were significantly shorter for T1 than T2 or T3. In the S1 context, as in the monosyllabic context, /ɛ/ rimes were about twice as long as /aʔ/ rimes. Unsurprisingly, S1 syllables were shorter than monosyllables. Details are shown in Appendix 2.

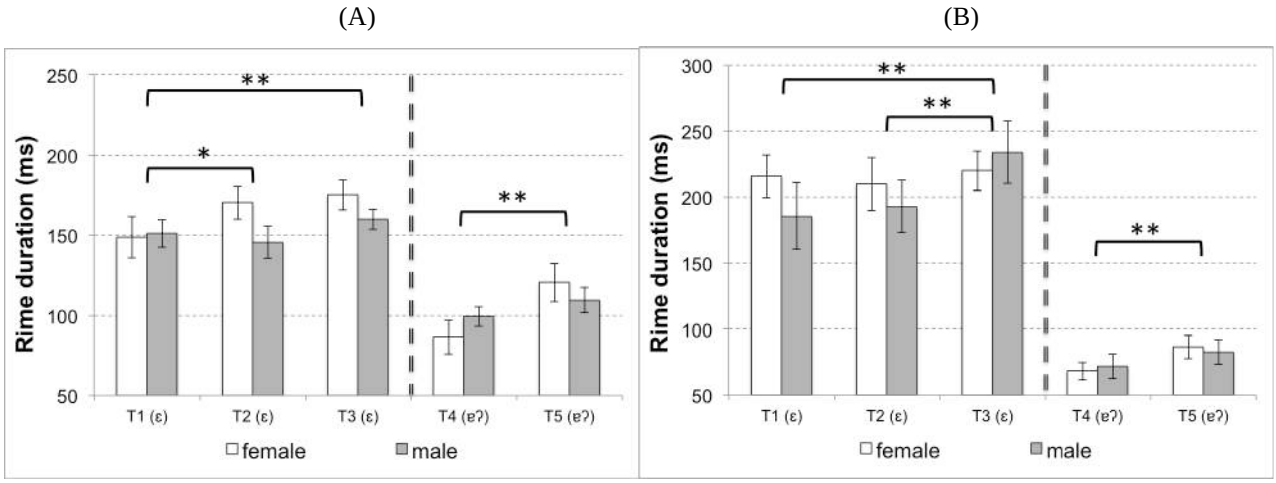


Figure 71. Average /ɛ/ and /aʔ/ rime durations in S1, according to S1's underlying tone and speaker group, pooled across onset types. (A) young speakers and (B) elderly speakers. Significance levels: * for $p<.05$; ** for $p<.01$.

We ran two by-subject ANOVAs separately for unchecked (T1-3) and checked (T4-5) S1 syllable data. In both analyses, *Duration* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. Underlying *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) and *Manner* of articulation (stop vs. fricative) and *Place* of articulation (labial vs. dental) were within-subject factors.

For unchecked syllables, *Tone* was highly significant overall, $F(2,36)=23.3$, $p<.0001$. The /ɛ/ rimes were significantly longer for T3 than T1 (213>187 ms), $F(1,18)=27.9$, $p<.0005$, or T2 syllables (213>193 ms), $F(1,18)=44.1$, $p<.0001$. The /ɛ/ rimes were somewhat shorter for T1 than T2 (187 vs. 193 ms), but significantly so only for young speakers (173<186 ms), $F(1,10)=6.4$, $p<.05$. *Manner* was significant, $F(1,18)=10.0$, $p<.01$: rimes were significantly longer after fricative than stop onsets (202>193 ms). *Place* was marginal, $F(1,18)=4.1$, $p=.059$. The *Tone* × *Manner* interaction was not significant, $F(2,36)<1$. The *Tone* × *Place* interaction was significant overall, $F(2,36)=3.8$, $p<.05$, but not when restricted to the T1-2 vs. T3 comparison, $F(1,18)<1$. Neither *Gender* nor *Age* was significant (Gender: $F(1,18)=1.9$, $p=.19$; Age: $F(1,18)=2.3$, $p=.15$). Finally, *Tone* did not interact with either *Gender* or *Age* (*Tone* × *Gender*: $F(2,36)=2.4$, $p=.10$; *Tone* × *Age*: $F(2,36)=1.7$, $p=.20$).

For checked syllables, *Tone* was highly significant overall, $F(1,18)=54.3$, $p<.0001$: /a?/ rimes were significantly longer for T5 than T4 (86>74 ms). This difference was significant at least at the $p<.05$ level in each speaker group. *Manner* was not significant, $F(1,18)=2.9$, $p<.10$, but *Place* was significant, $F(1,18)=17.5$, $p<.001$: /a?/ rimes were longer after dental than labial onsets (85>73 ms). The *Tone* × *Manner* interaction was significant, $F(1,18)=62.7$, $p<.0001$: *Tone* was significant only for stop (Δ=19 ms), $F(1,18)=150.4$, $p<.0001$, but not for fricative onsets (Δ=3 ms), $F(1,18)=2.4$, $p=.14$. The *Tone* × *Place* interaction was also significant, $F(1,18)=38.9$, $p<.0001$: *Tone* was significant only for dental (Δ=22 ms), $F(1,18)=135.1$, $p<.0001$, but not for labial onsets (Δ=1 ms), $F(1,18)<1$. Neither *Gender* nor *Age* was significant (*Gender*: $F(1,18)=1.4$, $p=.25$; *Age*: $F(1,18)<1$). The *Tone* × *Gender* interaction was marginal, $F(1,18)=3.8$, $p=.067$, reflecting a trend for larger T5 vs. T4 differentials for female (Δ=14 ms) than male speakers (Δ=11 ms). The *Tone* × *Age* interaction was significant, $F(1,18)=5.8$, $p<.05$, reflecting larger T5 vs. T4 differentials for elderly (Δ=15 ms) ,than young speakers (Δ=10 ms).

4.1.6.4.3 Second syllable in disyllables (S2)

4.1.6.4.3.1 Onset duration

Figure 72 shows the average onset durations in the S2 context (second syllable of disyllabic words), pooled across onset types and speaker groups. Figure 73 and Figure 74 separately show the stop and fricative data for (A) for young speakers and (B) elderly speakers. Stop duration consists of two parts: release duration and closure duration (which could be measured in the S2 context).

To summarize our statistical results, significant onset duration differences were found between T3 (*yang*) and the two *yin* tones T1 and T2, as well as between T5 (*yang*) and T4 (*yin*), regardless of the sandhi pattern, that is, regardless of the first syllable's underlying tone T1 or T2. The data were thus pooled across the two possible tones T1 or T2 of the first syllable. The closure duration of stop onsets was shorter for *yang* than *yin* tones in both unchecked and checked syllables, whereas their release

duration was shorter for *yang* than *yin* tones only in unchecked syllables (T1-3). Details are reported in Appendix 2.

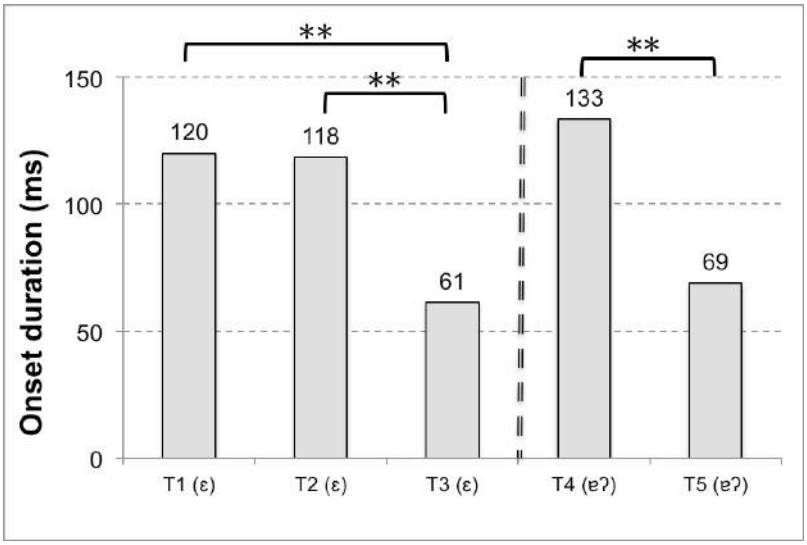


Figure 72. Average onset durations in S2, according to S2's underlying tone, pooled across onset types and speaker groups. Significance level: ** for p<.001.

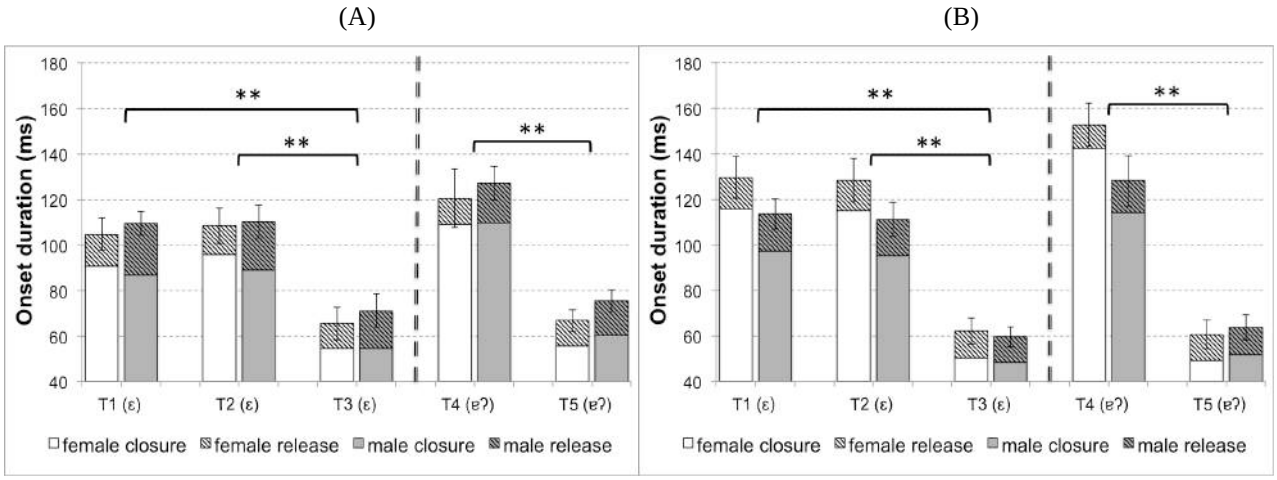


Figure 73. Average stop onset durations (closure and release) in S2, according to S2's underlying tone and speaker group. (A) young speakers; (B) elderly speakers. Significance levels: ** for p<.01.

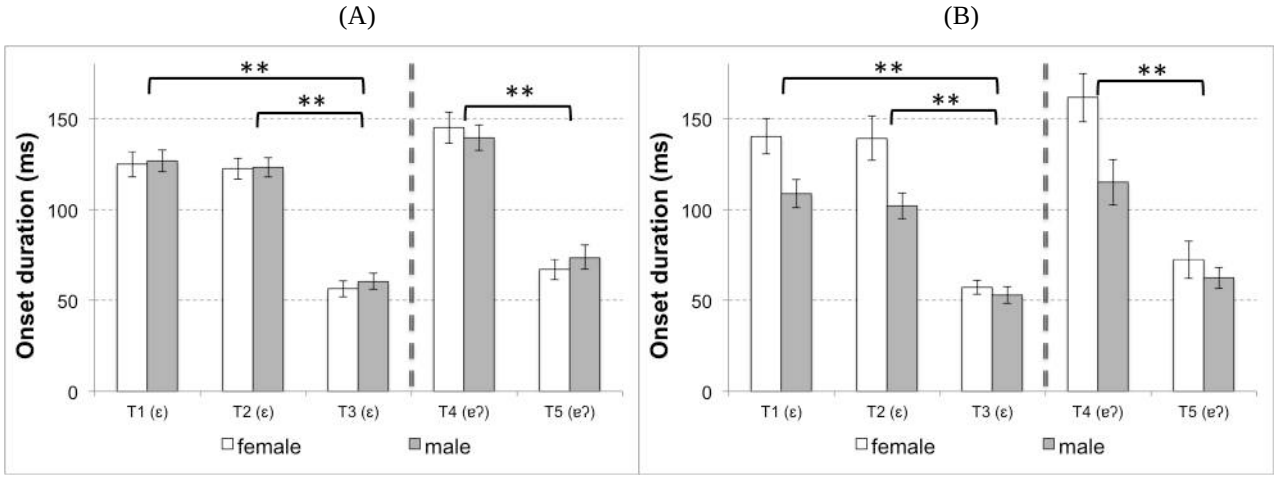


Figure 74. Average fricative onset durations in S2, according to S2's underlying tone and speaker group. (A) young speakers; (B) elderly speakers. Significance levels: ** for $p<.01$.

We ran two by-subject ANOVAs separately for the unchecked (T1-3), and the checked syllable (T4-5) data. In both analyses, *Duration* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. Underlying *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5), *Manner* of articulation (stop vs. fricative), *Place* of articulation (labial vs. dental) and *Sandhi* pattern (post-T1 vs. post-T2) were within-subject factors.

For unchecked syllables, *Tone* was highly significant overall, $F(2,36)=504.2$, $p<.0001$. Onsets were significantly shorter for T3 than T1 (61<120 ms), $F(1,18)=576.4$, $p<.0001$, or T2 (61<118 ms), $F(1,18)=524.7$, $p<.0001$. There was no significant difference between T1 and T2, $F(1,18)=1.5$, $p=.24$. *Place* was not significant, $F(1,18)=2.4$, $p=.14$, nor was *Sandhi*, $F(1,18)<1$. *Manner* was significant, $F(1,18)=5.6$, $p<.05$: *Duration* was longer for fricative than stop onsets (102>99 ms). The *Tone* × *Manner* interaction was significant, $F(2,36)=51.1$, $p<.0001$, reflecting larger T1-2 vs. T3 differentials for fricative (Δ=68 ms), $F(2,36)=587.2$, $p<.0001$, than stop onsets (Δ=49 ms), $F(2,36)=7.3$, $p<.005$. The *Tone* × *Place* interaction was significant, $F(2,36)=7.3$, $p<.005$, reflecting larger T1-2 vs. T3 differentials for dental (Δ=61 ms), $F(2,36)=502.6$, $p<.0001$, than labial onsets (Δ=56 ms), $F(2,36)=353.3$, $p<.0001$. The *Tone* × *Sandhi* interaction was not significant, $F(2,36)=1$. Neither *Gender* nor *Age* was significant (*Gender*: $F(1,18)=1.3$, $p=.27$; *Age*: $F(1,18)<1$). Finally, both the *Tone* × *Gender* and the *Tone* × *Age* interactions were significant (*Tone* × *Gender*: $F(2,36)=5.4$, $p<.01$; *Tone* ×

Age: $F(2,36)=7.3$, $p<.005$), reflecting larger T1-2 vs. T3 differentials for female than male speakers, and for elderly than young speakers, with the largest differentials for elderly female speakers (Δ=74 ms, against 51 to 54 ms for the other speaker groups).

For checked syllables, *Tone* was highly significant, $F(1,18)=455.5$, $p<.0001$: T5 onsets were significantly shorter than T4 ones (69<133 ms). *Place* was not significant, $F(1,18)=2.2$, $p=.16$. *Manner* was significant, $F(1,18)=4.6$, $p<.05$: *Duration* was longer for fricative than stop onsets (105>99 ms). *Sandhi* was significant, $F(1,18)=8.9$, $p<.01$: *Duration* was longer for post-T1 than post-T2 onsets (102>99 ms). The *Tone* × *Manner* interaction was significant, $F(1,18)=11.7$, $p<.005$, reflecting larger T4 vs. T5 differentials for fricative (Δ=72 ms), $F(1,18)=479.3$, $p<.0001$, than stop onsets (Δ=67 ms), $F(1,18)=250.9$, $p<.0001$. Neither the *Tone* × *Place* nor the *Tone* × *Sandhi* interaction was significant, Fs(1,18)<1. Neither *Gender* nor *Age* was significant (*Gender*: $F(1,18)=2.0$, $p=.18$; *Age*: $F(1,18)<1$). Finally, *Tone* did not interact with *Age*, $F(1,18)=3.6$, $p=.074$, but did interact with *Gender*, $F(1,18)=9.9$, $p<.01$, reflecting larger T4 vs. T5 differentials for female (Δ=75 ms), $F(1,10)=280.6$, $p<.0001$, than male speakers (Δ=54 ms), $F(1,8)=165.4$, $p<.0001$.

Two other by-subject ANOVAs were conducted on the release duration of S2 stop onsets, with *Duration* as the dependent variable, *Subject* as the random factor, and underlying *Tone* of S2 as the within-subject factor.

For unchecked syllables, *Tone* was highly significant, $F(2,36)=26.6$, $p<.0001$. Stop release duration was shorter for T3 than T1 (13.1<16.7 ms), $F(1,18)=38.5$, $p<.0001$, or onionmarginal (16.7 vs. 15.9 ms), $F(1,18)=4.1$, $p=.058$).

For checked syllables, *Tone* was not significant, $F(1,18)=2.8$, $p=.11$.

4.1.6.4.3.2 Rime duration

Figure 75 shows the average duration of /ɛ/ and /aʔ/ rimes in the S2 context (second syllable of disyllabic words), according to the underlying tone of S2, pooled across onset types and speaker groups. Figure 76 shows the same data detailed by speaker group and according to the tone of the first syllable (post-T1 vs. post-T2).

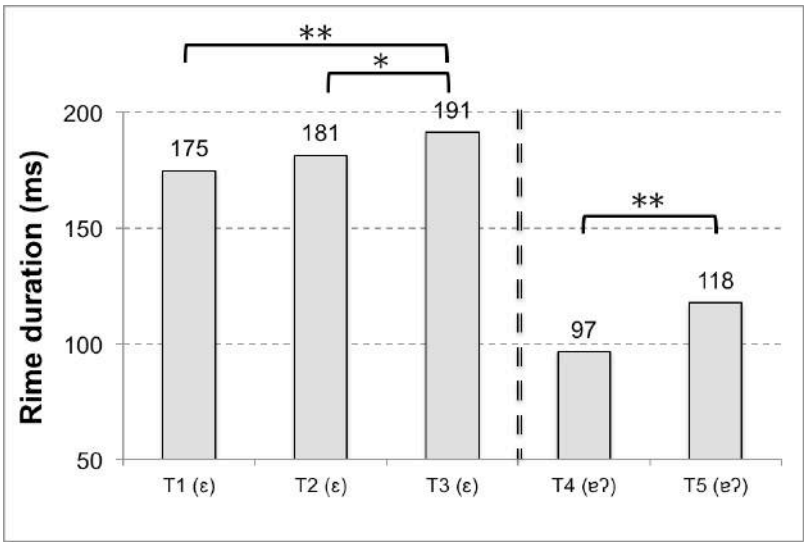
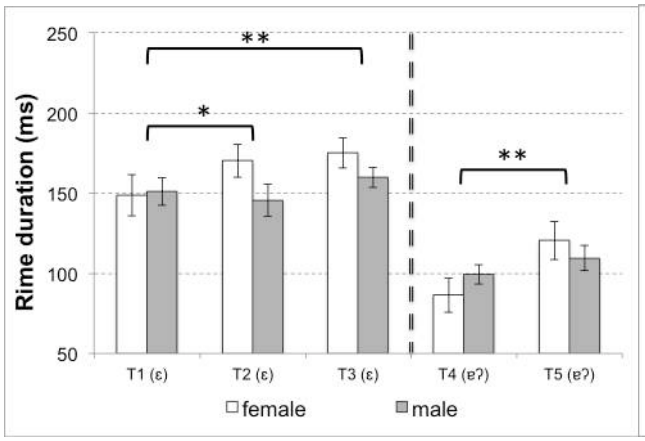


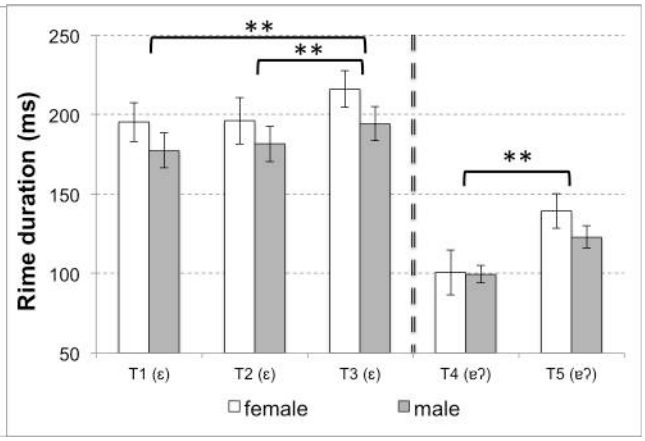
Figure 75. Average /ɛ/ and /aʔ/ rime durations in S2 according to S2's underlying tone, pooled across onset types and speaker groups. Significance levels: * for $p<.01$; ** for $p<.001$.

To summarize the general trends shown by the statistical analysis, for both age and gender groups, T3 (*yang*) rimes were significantly longer than T1 or T2 (*yin*) rimes and T5 (*yang*) rimes significantly longer than T4 (*yin*) rimes. Rimes were longer after a T2 than a T1 first syllable. Female speakers produced longer (unchecked) /ɛ/ rimes than male speakers.

(A-1)



(A-2)



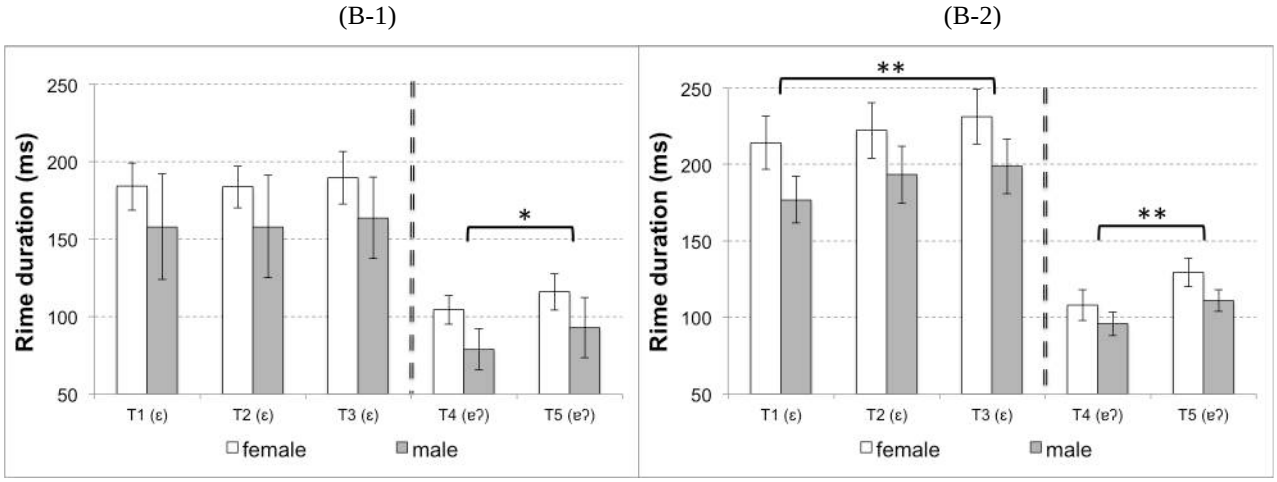


Figure 76. Average /ɛ/ and /aʔ/ rime durations in S2 according to S2's underlying tone, pooled across onset types. (A) young speakers and (B) elderly speakers; (1) Post-T1 and (2) Post-T2. Significance levels: * for $p<.05$; ** for $p<.01$.

We ran two by-subject ANOVAs separately for the unchecked (T1-3) and checked (T4-5) syllable data. In both analyses, rime *Duration* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* group (young vs. elderly) were between-subject factors. Underlying *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5), *Manner* of articulation (stop vs. fricative), *Place* of articulation (labial vs. dental), and *Sandhi* pattern were within-subject factors.

For unchecked syllables, *Tone* was highly significant overall, $F(2,36)=24.4$, $p<.0001$: S2 rimes were significantly longer for T3 than T1 (191>175 ms), $F(1,18)=27.9$, $p<.0005$, or T2 (191>181 ms), $F(1,18)=44.1$, $p<.0001$. They were also slightly longer for T2 than T1 (181>175 ms), $F(1,18)=5.6$, $p<.05$. *Place* was significant, $F(1,18)=5.0$, $p<.05$: rimes were longer with dental than labial onsets (184>180 ms). *Sandhi* was highly significant, $F(1,18)=54.5$, $p<.0001$: post-T1 rimes were shorter than post-T2 rimes (165<200 ms). *Manner* was not significant, $F(1,18)=2.2$, $p=.16$. *Tone* did not interact with any of the other within-subject variable, Fs$(2,36)<1$. *Age* was not significant, $F(1,18)=2.3$, $p=.14$, but *Gender* was, $F(1,18)=4.4$, $p<.05$: female speakers produced longer rimes than male speakers (194>171 ms). Finally, *Tone* did not interact with *Gender* ($F(2,36)<1$) or *Age* ($F(2,36)=1.4$, $p=.25$).

For checked syllables, *Tone* was highly significant, $F(1,18)=111.2$, $p<.0001$. S2 rimes were significantly longer for T5 than T4 (118>97 ms). Manner was significant, $F(1,18)=13.6$, $p<.005$: rimes were longer with fricative than stop onsets (109>105 ms).

Sandhi was significant, $F(1,18)=18.7$, $p<.0005$: post-T1 rimes were shorter than post-T2 rimes (101<113 ms). *Place* was not significant, $F(1,18)=1.6$, $p=.22$. *Tone* did not interact with Manner, $F(1,18)=1.3$, $p=.26$, but did interact with *Place*, $F(1,18)=28.9$, $p<.0001$, reflecting larger T5 vs. T4 differentials for rimes with dental (Δ=28 ms), $F(1,18)=159.7$, $p<.0001$, than labial onsets (Δ=28 ms), $F(1,18)=31.7$, $p<.0001$. The *Tone* × *Sandhi* interaction was also significant, reflecting larger T5 vs. T4 differentials for post-T2 (Δ=25 ms), $F(1,18)=137.6$, $p<.0001$, than post-T1 rimes (Δ=18 ms), $F(1,18)=31.7$, $p<.0001$. Neither *Gender* nor *Age* was significant (Gender: $F(1,18)=1.8$, $p=.20$; Age: $F(1,18)<1$). Finally, both the *Tone* × *Gender* and the *Tone* × *Age* interactions were significant, reflecting larger T5 vs. T4 differentials for female than male speakers, and for young than elderly speakers, with the largest differential for young female speakers (Δ=36 ms against 14 to 17 ms for the other speaker groups).

4.1.6.4.4 Discussion

In this section, we investigated how duration patterns correlated with tone categories in monosyllabic and disyllabic contexts.

In all contexts (monosyllables, first and second syllable in disyllables), *yin* obstruents were consistently longer than their *yang* counterparts. This included both fricatives and stops in word-medial position but only fricatives in word-initial position since closure duration is not measurable from the acoustic signal for phonetically voiceless stops in word-initial position (see §4.3 for articulatory measurements). This result is in line with the findings of prior studies (Liu, 1925: 62; Shen et al., 1987; Gao & Hallé, 2012). Nasal onsets, however, did not follow this pattern, and even tended to follow the opposite pattern for elderly speakers. The data from Gao et al. (2011) on a single speaker who produced longer *yin* than *yang* nasals thus cannot be generalized.

The phonetic voicing of *yang* fricatives might contribute partly to this difference of onset duration, as voiced fricatives are shorter than their voiceless counterparts due to aerodynamic constraint in the production of voicing (Ohala, 1997, see §3.3). But probably it is not the only contributor. Since *yang* labial fricatives are more often voiced than *yang* dental fricatives, if phonetic voicing was the main contributor, we would expect notably greater *yin* vs. *yang* duration differentials for labial than dental

fricatives. In the S1 context, we indeed found significant interaction between tone and place of articulation, which indicates that phonetic voicing does play a role. But we did not find such interaction for the monosyllabic context, which suggests that phonological voicing contributes to the duration difference as well. Furthermore, elderly male speakers maintain the fricative duration difference between *yin* and *yang* but have a much lower occurrence of voiced fricatives (in monosyllables, 14.3%, compared to 32.5% for elderly female, 35.4% for young male, and 70.5% for young female speakers, see Table 17 for a detailed distribution). Therefore, apart from phonetic voicing, phonological voicing related to the *yin* vs. *yang* register must affect onset duration.

Prior studies found that vowels are longer before a voiced than a voiceless consonant, but vowels following a consonant were not found to vary as a function of the consonant's voicing (Lisker, 1957; Peterson & Lehiste, 1960).

Our data suggest that the vowel following an obstruent also varies as a function of the consonant's phonological voicing, that is, *yang* rimes were globally longer than their *yin* counterparts. This difference is consistent and robust for checked syllables (T5 rimes longer than T4) in all contexts and for all speaker groups. For unchecked syllables, T3 rimes are globally longer than T1 and T2 rimes. But the difference varies according to speakers' age group and context. For elderly speakers, in monosyllables, T1 rimes are shorter than T2 rimes, which are shorter than T3 rimes; in the S1 context, T1 and T2 rimes have a similar duration, and both are shorter than T3 rimes. For young speakers, in both the monosyllabic and the S1 contexts, T1 rimes are shorter than T2 or T3 rimes, but no difference is found between T2 and T3 rimes. In the S2 context, where the syllable is unstressed and the tone contour neutralized, the difference between *yin* and *yang* rimes is smaller than in the other contexts.

Gao & Hallé (2012) explained the rime duration difference by a tendency for the syllable to be a stable rhythmic unit, so that a longer consonant would be compensated by a shorter vowel. However, consonant-vowel compensation might not be crucial, since *yin* and *yang* nasal onsets do not differ in duration, but rimes following zero or nasal onsets also exhibit the *yang*>*yin* duration difference.

Other proposals have been put forward to explain rime duration difference. It has been proposed as related to F0 height or F0 contour (Gandour, 1977), to F0 range

(Hombert, 1977), or to phonation type (e.g., Fischer-Jørgensen, 1967) (see §3.3). Given the fact that the duration difference between T2 and T3 is found only among elderly speakers, and breathy voice is also more salient among elderly speakers, we might reasonably conclude that for T2 and T3 syllables, the rime duration mainly depends on phonation type. In Shanghai Chinese, breathy vowels would thus be longer than modal phonation ones. (Phonation difference is less notable for syllables with nasal onsets, but is still present.)

Checked syllables have probably another phonetic mechanism that contributes to the rime duration difference. In all contexts, during the production of the onset, the glottis is slacker for *yang* than *yin*, while the vocalic part should be produced with a constricted glottis for the final glottal stop of the checked vowels. The required laryngeal maneuver thus seems more demanding for *yang* than *yin* rimes, hence might take a longer time, yielding longer *yang* than *yin* rimes.

The fact that T1 rimes are the shortest in monosyllables is in line with Gandour's (1977) observation according to which vowels carrying a falling tone are shorter than those carrying a rising tone.

As for the duration of the intervocalic stops' release, it is shorter for T3 than T1 or T2 stops. This is contrary to the pattern of word-initial stops' release, estimated with the VOT values. Word-initial VOTs are longer for T3 than T1 or T2 syllables (see §4.1.6.2). As found in prior studies (Maddieson & Ladefoged, 1985; Andruski & Ratliff, 2000, see §4.1.6.2.4), we attribute longer VOTs of *yang* stops in word-initial position to the breathy voice during the stop release, which implies a slack configuration of the vocal folds and delays their vibration. In intervocalic position, *yang* stops are not breathy and are fully voiced. Their release is consequently shorter than their *yin* voiceless counterparts. Release duration is indeed generally shorter for voiced stops than voiceless stops (e.g., Meynadier & Gaydina, 2012).

Finally, female speakers produced longer rimes than male speakers, while male speakers produced longer VOTs or stop releases than female speakers. This is contrary to the findings of previous studies, according to which female speakers have higher VOTs than males (Whiteside & Marshall, 2001; Whiteside, Hanson & Cowell, 2004). It seems that, in Shanghai Chinese, female and male speakers have different strategies for the production of tone contrasts: female speakers rely more on the

prosodic information in the vocalic part, while male speakers lengthen the consonantal part that contains essential information on the phonation type.

4.2 Experiment 1 (EGG data): production of phonetic correlates of Shanghai tones [15]

Electroglottography (EGG) is a widely used and noninvasive method to measure normal and pathological glottal state and vibratory characteristics during phonation (e.g., Childers, Hicks, Moore, Eskenazi, & Lalwani, 1990). The degree of contact of the vocal folds can be estimated from the variations in vocal fold contact area (VFCA), which modulate the EGG signal (Rothenberg & Mahshie, 1988), as shown in Figure 77. The period of each closing-opening glottal cycle corresponds to a period in the EGG signal, with visually separable fully closed and opening glottis phases. Consequently, it is possible to measure the proportion of time when the vocal folds are abducted during each glottal cycle (open quotient, or OQ), or the proportion of time when they are adducted during each glottal cycle (closed quotient, or CQ). The OQ or CQ measure can be used to describe phonation types. Theoretically, in modal voice, the open phase and the closed phase of the glottal cycle have approximately equal durations, that is, both OQ and CQ should be around 0.5. OQs higher than 0.5 (hence, CQs lower than 50%) indicate more abducted phonation, and thus breathier voice. On the contrary, OQs lower than 0.5 (hence, CQs higher than 0.5) indicate a more adducted or constricted phonation, which may be called "laryngealized." OQ values are assumed to correlate positively with H1–H2 values (e.g., Holmberg, Hillman, Perkell, Guiod, & Goldman, 1995), although some authors find this correlation is rather weak (Kreiman, Iseli, Neubauer, Shue, Gerratt, & Alwan, 2008). In fact, when phonation types are contrastive, as in contrastive tones in some languages, relative OQ values rather than absolute values distinguish phonation types from one another. In this experiment, we looked at OQ variation according to tone, with the working hypothesis that *yang* tones should be breathier than *yin* tones, hence be characterized by larger OQ values.

---

[15] Part of the results of this section have been reported at the 2nd International Congress on Phonetics of the Languages of China (ICPLC), december 2013, Hong Kong, and published in the proceedings of the conference.

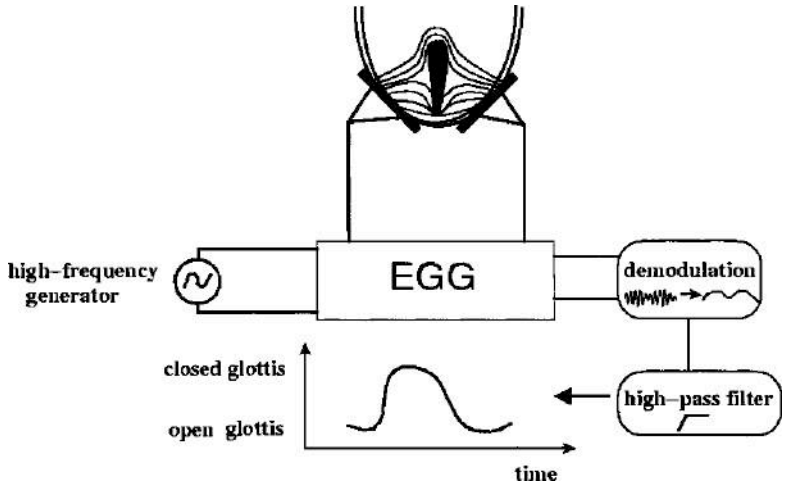


Figure 77. The EGG signal represented as a function of vocal-fold contact, from Henrich et al. (2004).

4.2.1 Method

4.2.1.1 Participants

We report the data from 10 native speakers of Shanghai Chinese, including 6 young speakers (3 males and 3 females, mean age 24.3, range 24–25) and 4 elderly speakers (3 males and 1 female, mean age 67.3, range 64–72). Ten more speakers were recorded but we retained only the speakers of whom the EGG signal was the least noisy. All the speakers participated in the audio recordings reported in §4.1. All the young speakers learned Standard Chinese before the age of eight and learned English at school; one young speaker spoke German and two spoke French; one of them spoke Chuansha 川沙 dialect (Songjiang subgroup), and the others did not speak any other dialect than Shanghai Chinese and Standard Chinese. As for the elderly speakers, all had learnt Standard Chinese at primary school or at adult age. None of them spoke any foreign language, but three of them spoke another Wu dialect of Jiangsu or Zhejiang than Shanghai Chinese. All the speakers were born in Shanghai urban or suburban areas. All of them had spent most of their lifetime in Shanghai urban area. One elderly male speaker had been living in the 崇明 Chongming county, where the Chongming subgroup is spoken, for nearly 30 years.

4.2.1.2 Speech materials

The same thirty-two monosyllabic words as mentioned in §4.1.3 were produced in the carrier sentence /_ gəʔ əʔ zɻ̩ ŋo nin təʔ əʔ/ (__ 这个字/词我认得的。'__ this word/character, I know it'). Each target syllable carried one of the five citation tones. The checked syllables (tones T1, T2 or T3) shared the /ɛ/ rime and the unchecked syllables (tones T4 or T5) shared the /aʔ/ rime.

In the monosyllabic context, the onset could be empty, or was a labial or dental stop or fricative or nasal: /Ø, p, (b), t, (d), f, (v), s, (z), m, n/. There is no T2 /nɛ/ syllable, nor T4 /maʔ/ or /naʔ/ syllable. This made a total of 32 monosyllables (7 onsets * 5 tones – 3).

Each speaker read the list of sentences with target monosyllables twice, except one young female speaker who read it only once. For each target syllable, the repetition for which the EGG signal was less noisy was chosen for analysis. In those cases in which the two signals were of comparable quality, the first repetition was retained.

4.2.1.3 Data recordings and analyses

Speakers were recorded individually in a quiet room. The EGG signals was recorded simultaneously with the acoustic signal, using a Voicevista EGG recording device (http://www.eggsforsingers.eu/), with an internal 3 Hz high-pass filter. This device was connected to a laptop computer in which the acoustic and EGG signals were digitized and stored, at 44,100 Hz sampling rate and 16 bit precision, in WAV format files. The EGG signal was then high-pass filtered at 30 Hz using a Butterworth filter to eliminate the noise due to gross larynx movements. The analyses were conducted on each target syllable from the first to the last glottal closure in the vocalic portion of the syllable.

For F0 and OQ measurements, the opening and the closing instants of the glottis need to be determined. A classic method consists of defining a fixed threshold between the maximum and the minimum amplitude of the EGG signal. But the various

thresholds proposed have made this method controversial and have been found to lack accuracy and reliability, especially for non-pathological voices.

More recently, a method based on the derivative of the EGG signal (or dEGG) has been compared to the threshold-based methods and found more reliable for F0 and OQ measurements (Henrich, d'Alessandro, Doval, & Castellengo, 2004; Michaud, 2005). Based on numerous studies conducted by Childers and colleagues (e.g., Childers, Naik, Larar, Krishnamurthy, & Moore, 1983) using simultaneous EGG, dEGG and high-speed cinematographic data, the positive and negative peaks of the dEGG signal are found to correspond respectively to the instants when the vocal fold contact area increases and decreases with the greatest velocity, and are interpreted as the closing and opening instants of the glottis.

We adopted the dEGG method, using a semi-automatic Matlab program "peakdet", written by Alexis Michaud[16]. It first plots the EGG and the (smoothed) dEGG signals, as shown in Figure 78a-b, and then computes, for each glottal cycle, an F0 value based on the duration between two consecutive positive peaks on the dEGG signal, and finally the OQ value based on the duration between the negative and the positive peak divided by the duration of the entire glottal cycle. The F0 and the OQ trajectories are then plotted on the same figure, as shown in Figure 78c. In this illustration, both positive and negative peaks can be clearly defined, except for the last part of the rime in which some glottalization occurs.

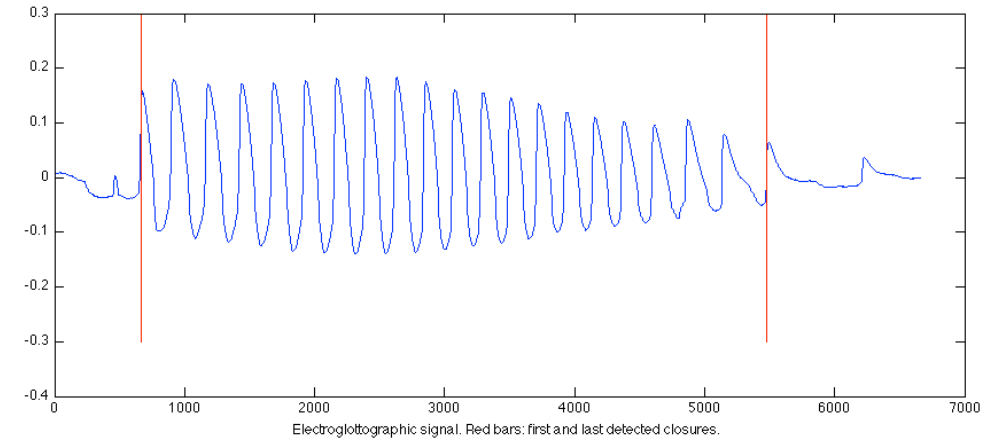
However, in many cases, the peaks cannot be detected that easily. The peaks may be too weak, or double (or sometimes triple) peaks may appear. Weak peaks always are the negative peaks signaling glottis opening and are presumably due to a gradual opening of the glottis. As for double peaks, negative or positive, the most plausible explanation is that they correspond to the consecutive opening or closing of two parts of the glottis. Such movements are indeed confirmed by high-speed movies showing the glottal activity (Anastaplo & Karnell, 1988; Hess & Ludwigs, 2000; Henrich, 2001: 109ff). For example, during slack phonation, the opening of the

[16] Downloadable from http://voiceresearch.free.fr/egg (Henrich, Gendrot, & Michaud, 2004), and from COVAREP (A COoperative Voice Analysis REPository for speech technologies): https://github.com/covarep/covarep/blob/f726176223c1cc808ece6ae1bfa5ee9ddd73cb48/glottalsource/egg/peakdet/peakdet.m

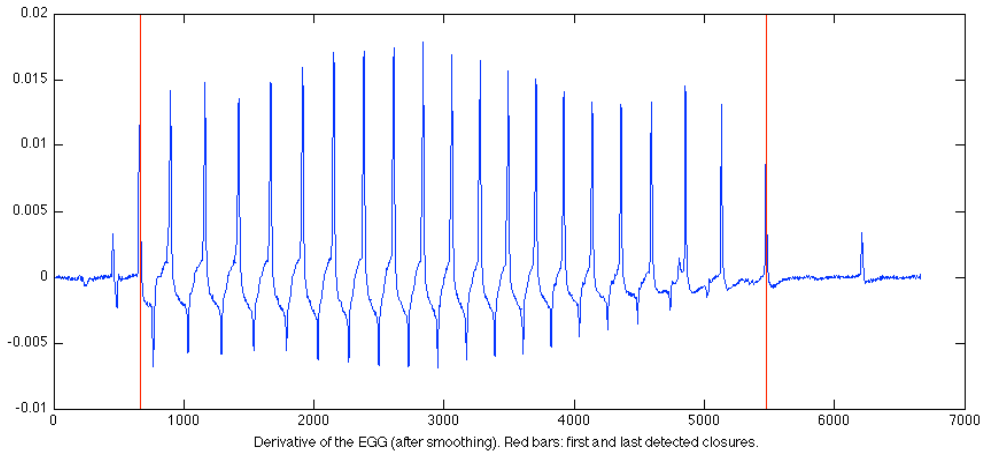
posterior part of the glottis, corresponding to the first peak, precedes that of the anterior part of the glottis, corresponding to the second peak (Hess & Ludwigs, 2000). In our dEGG signals representing breathy vowels, double negative peaks often occurred. In these cases, the strongest peak was retained for analysis.

The "peakdet" program offers four methods to calculate OQ: (1) the minima method (detection of the most strongly negative peak) on the unsmoothed dEGG signal; (2) the minima method on the smoothed dEGG signal; (3) the barycenter method on the unsmoothed dEGG signal; and (4) the barycenter method on the smoothed dEGG signal. The OQ values measured with the four methods are represented in four different colors. When a simple peak is detected, the results of the minima and barycenter methods converge, as shown in Figure 78. In case of double or imprecise peaks, if the OQ values computed by all four methods are discontinuous during a given vowel, or if the OQ values computed by the minima and barycenter methods are very different from one another, the data for this vowel are excluded from analyses, as shown in Figure 79. Otherwise, for the analyzable vowels, we selected the minima method on smoothed dEGG signal, as shown in Figure 80. In all three figures, green asterisks mark results of method (1), blue stars and line for method (2), red circles for method (3), and black squares for method (4).

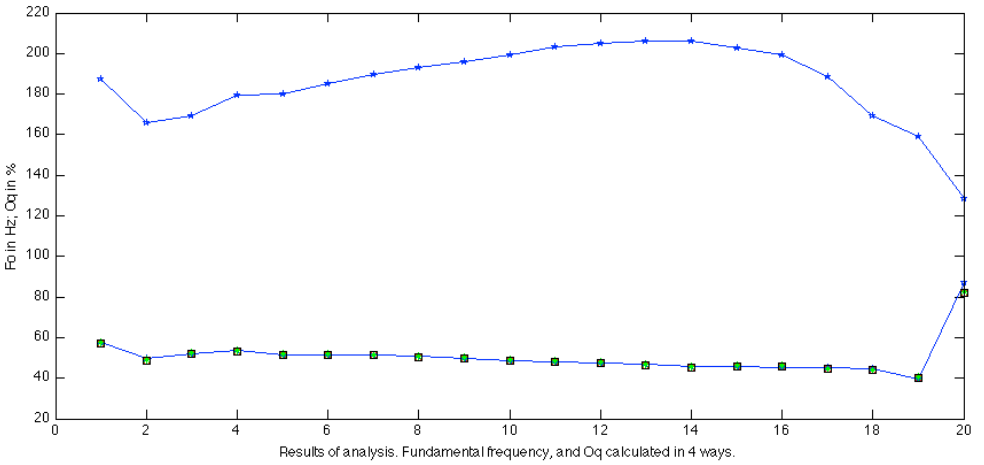
Checked syllables do not present special difficulties for analysis. They are not creaky during the vocalic part, but only during the final part of the rime, which is glottalized (on more or less long portion according to speaker). This final part can be defined as creaky and may cause irregularities in the EGG signal. In case of very short non-creaky vocalic duration, the signal was not usable and we discarded such occurrences (sound example for audio and EGG: CD 4.2.1.3_p183). In passing, unchecked syllables also are often glottalized in their final portion, with a shorter duration in general. These glottalized parts were not taken into account for F0 or OQ analyses.



(a)

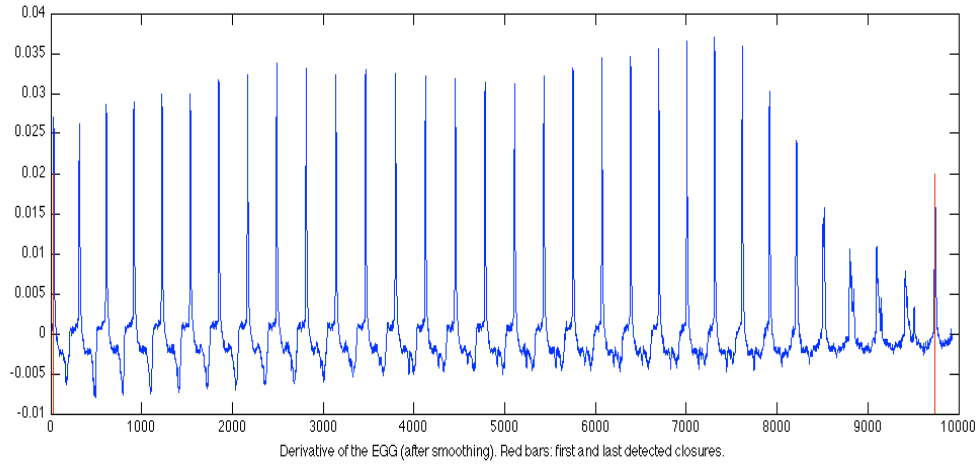


(b)

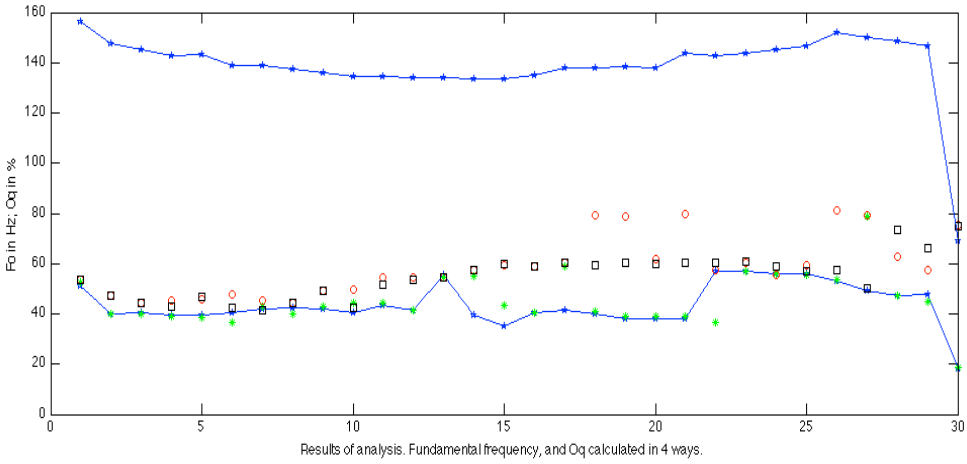


(c)

Figure 78. (a) EGG waveform, (b) smoothed dEGG, and (c) F0 and OQ computed on each glottal cycle for an /aʔ/ (T4) syllable produced by a 72-year-old male speaker: illustration of clearly defined peaks in the dEGG signal.

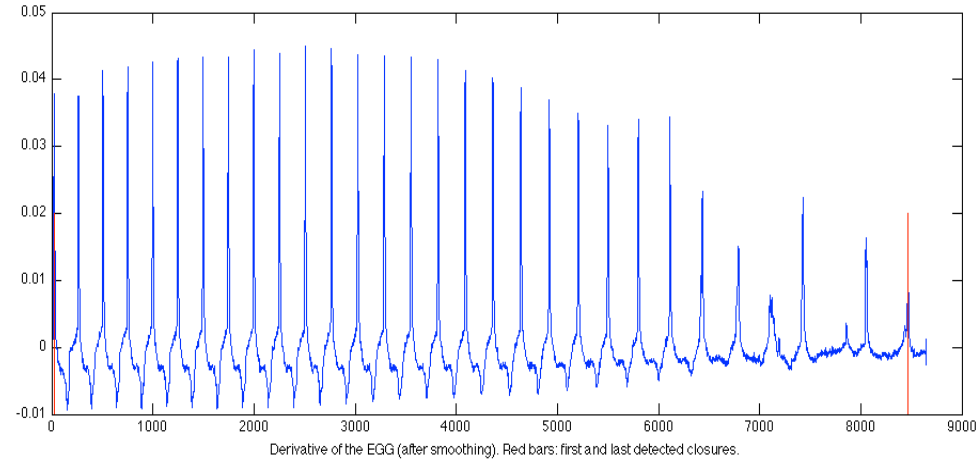


(a)

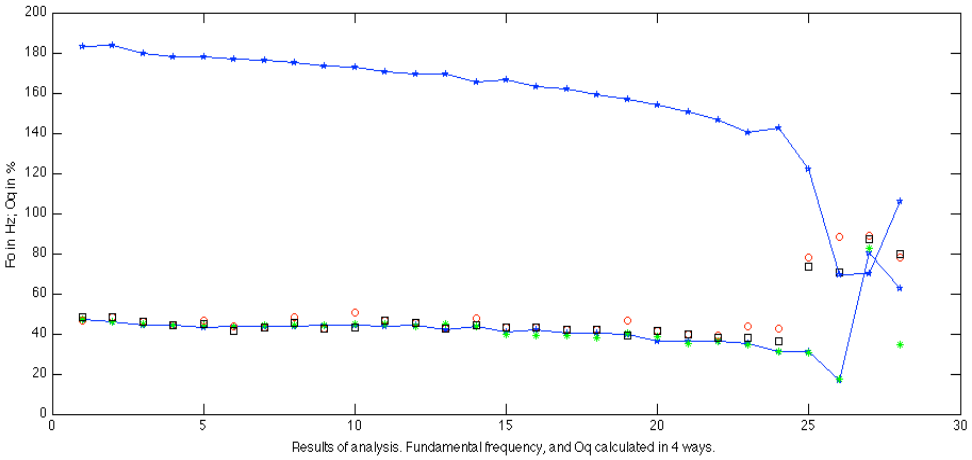


(b)

Figure 79. (a) dEGG signal, and (b) F0 and OQ computed on each glottal cycle for a /tɛ/ (T2) syllable produced by a 64-year-old male speaker: illustration of a great disagreement for OQ between the minima and barycenter methods.



(a)



(b)

Figure 80. (a) dEGG signal, and (b) F0 and OQ computed on each glottal cycle for a /tɛ/ (T1) syllable produced by a 64-year-old male speaker: illustration of minor disagreement for OQ between the minima and barycenter methods, except on the final part of the rime.

4.2.2 Results

4.2.2.1 Individual F0 illustrations

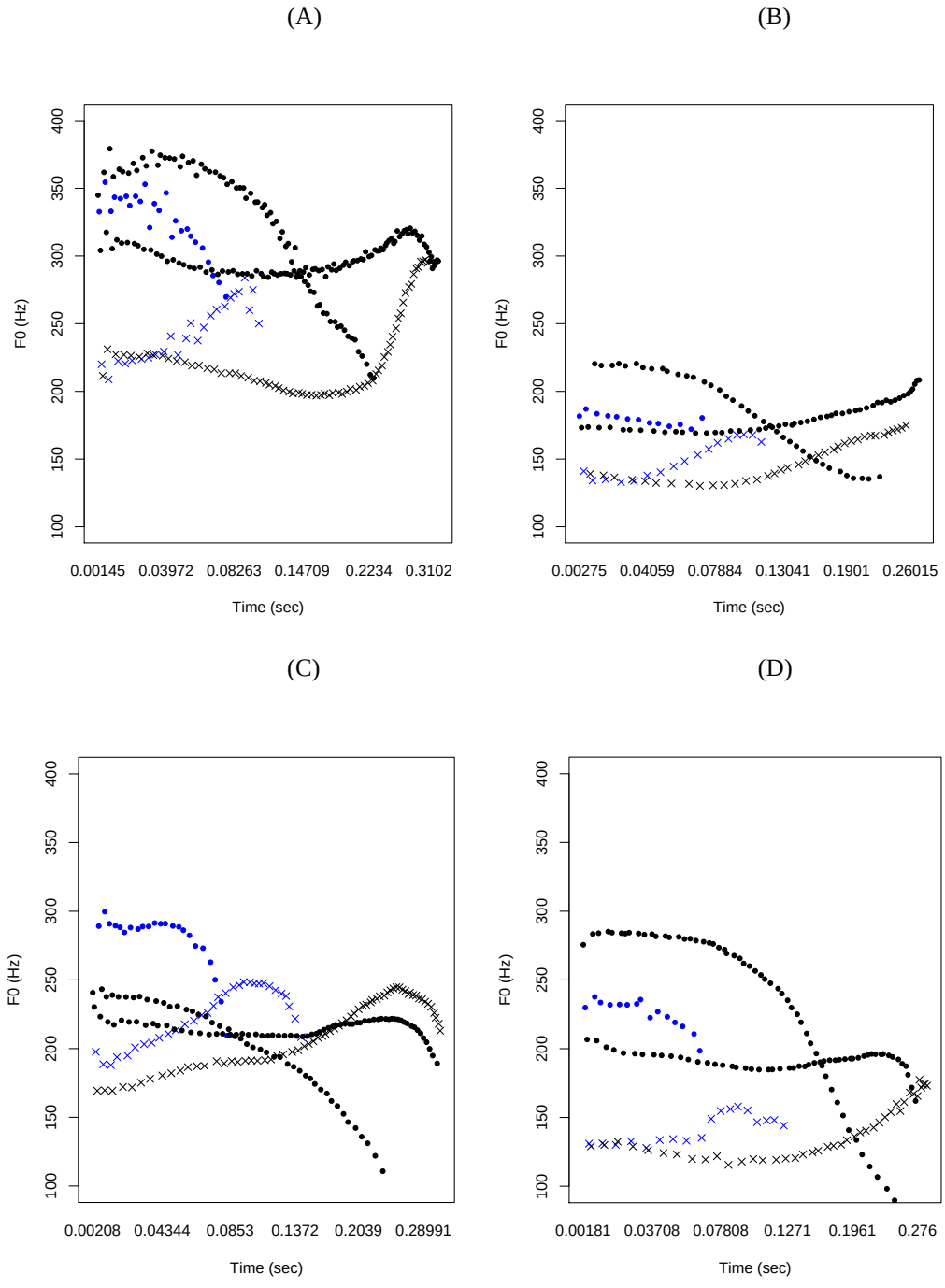


Figure 81. F0 trajectories of syllables with /p, b/ onset for the five citation tones produced by four individual speakers: (A) young female aged 24; (B) young male aged 25; (C) elderly female aged 67; (D) elderly male aged 66. Black for unchecked tones, blue for checked tones; dotted lines for *yin* tones, crossed-lines for *yang* tones.

As an illustration, Figure 81 shows the F0 trajectories of the five Shanghai citation tones produced by four speakers, one from each gender/age group. They produced /p-bɛ/ (T1-3) and /p-baʔ/ (T4-5) syllables. The F0 value of each glottal cycle was calculated from the derivative of the EGG (dEGG) signal. For each tone, these speakers followed the usual, canonical F0 contours described for Shanghai Chinese (see §2.1.3.1), as illustrated in §4.1.6.1. (Note that the elderly male speaker chosen for illustration used an especially high pitch compared to the other male speakers.) At the same time, some variation across the speakers can be observed. First, the two young speakers slightly raised F0 in the second half of T2 syllables, whereas the two elderly speakers produced this tone as rather flat. Second, the F0 contour for T3 was quite different from one speaker to the other. F0 rise was moderate for the two male speakers, with the first half of the contour quite flat and F0 rise achieved in the second half. For the young female speaker, the contour began with a smooth F0 fall until near the end of the rime, where F0 was raised abruptly. In contrast, the elderly female speaker produced a monotonically rising F0 contour along the entire syllable, with a bell-shaped F0 fall at the very end of the rime. This speaker was the only one to produce higher F0 contour offsets for T3 than T2. Third, although the *yang* tone contours of T5 and T3 began with almost the same F0 (except, perhaps, for the elderly female speaker), the relation between the *yin* tone contours of T4 for checked syllables and T1 and T2 for unchecked syllables varied a lot according to speaker: the F0 onset of T4 was closer to that of T1 than T2 for the young female speaker, closer to that of T2 than T1 for the young male speaker, between that of T1 and T2 for the elderly male speaker, and much higher than that of T1 or T2 for the elderly female speaker.

4.2.2.2 Open Quotient (OQ)

Figure 82 shows the time course of OQ along the /ɛ/ rime of unchecked syllables) according to tone (T1-3), and Figure 83 the time course of OQ along the /aʔ/ rime of checked syllables according to tone (T4-5). To summarize the statistical results, OQ values were higher for *yang* than *yin* vowels for the elderly male speakers only, which indicates a breathier phonation during *yang* vowels but only for this speaker group.

Surprisingly, this pattern was reversed for female speakers. No *yin-yang* difference was found overall for young male speakers but we observed between-speaker variations, which we describe in §4.2.2.3.

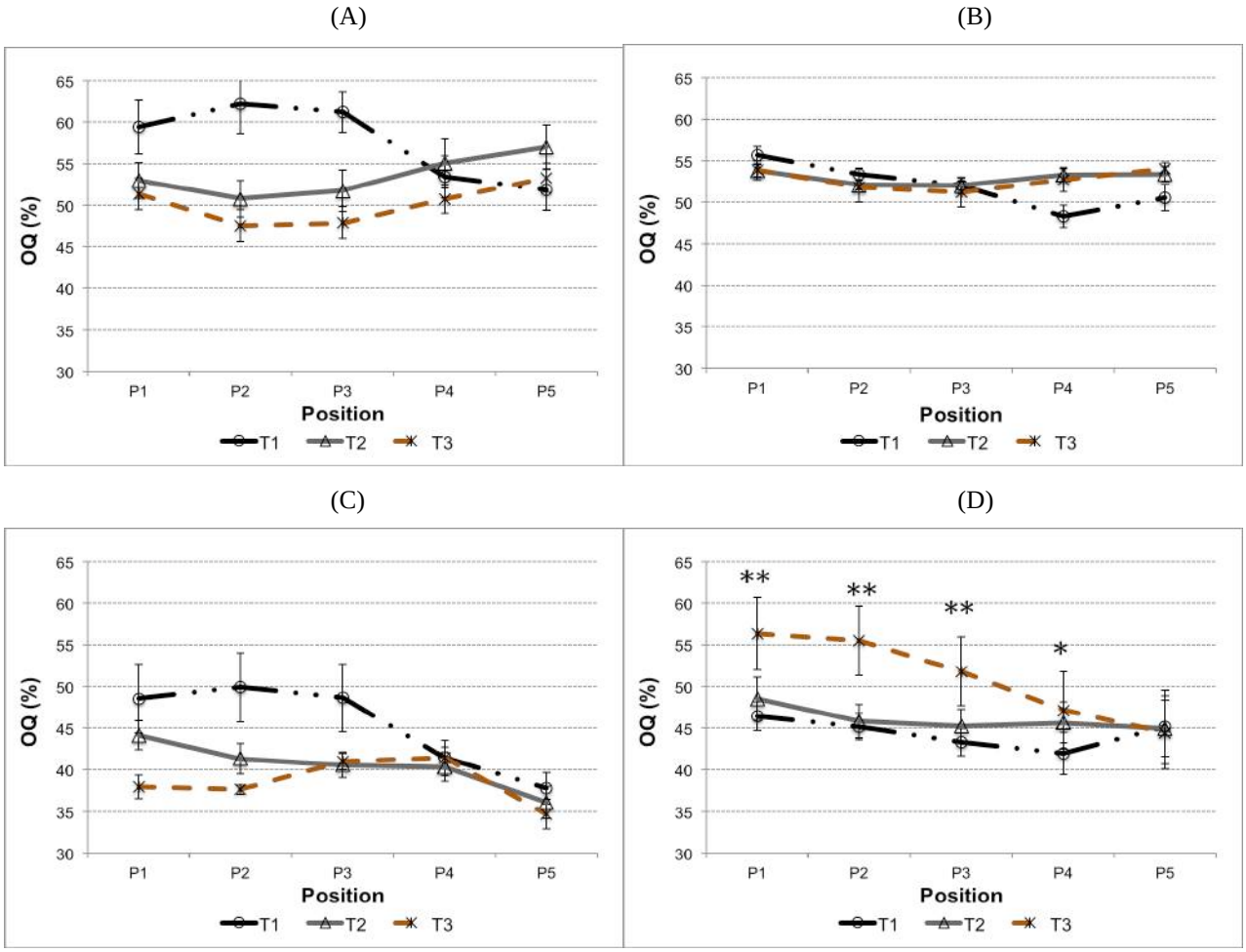
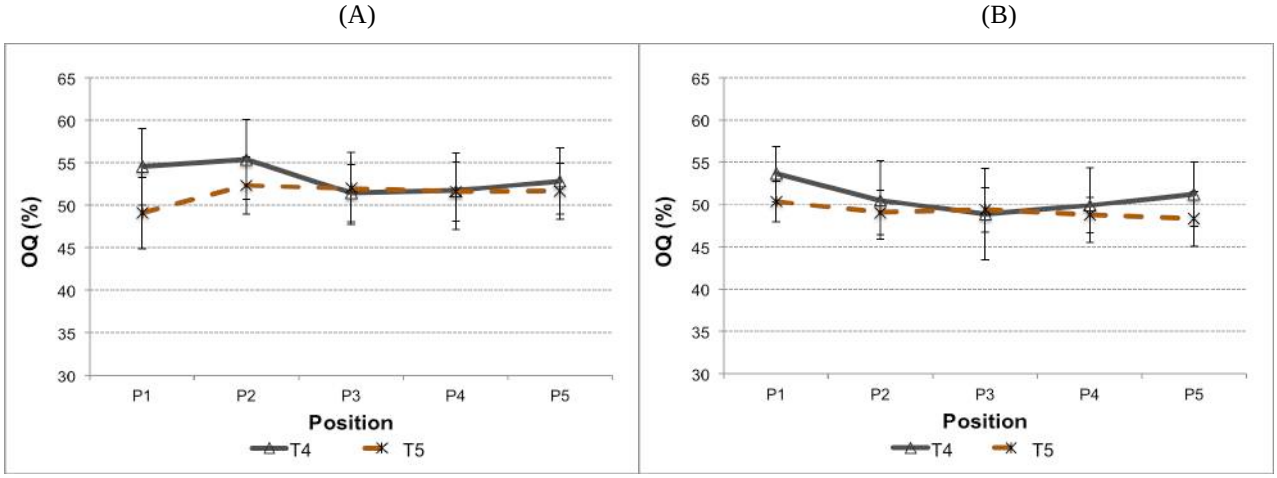


Figure 82. Average OQ according to tone and position in unchecked monosyllables (T1-3) for (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers. Significance levels: * for $p<.05$; ** for $p<.01$ (higher OQ for T3 than T1-2).



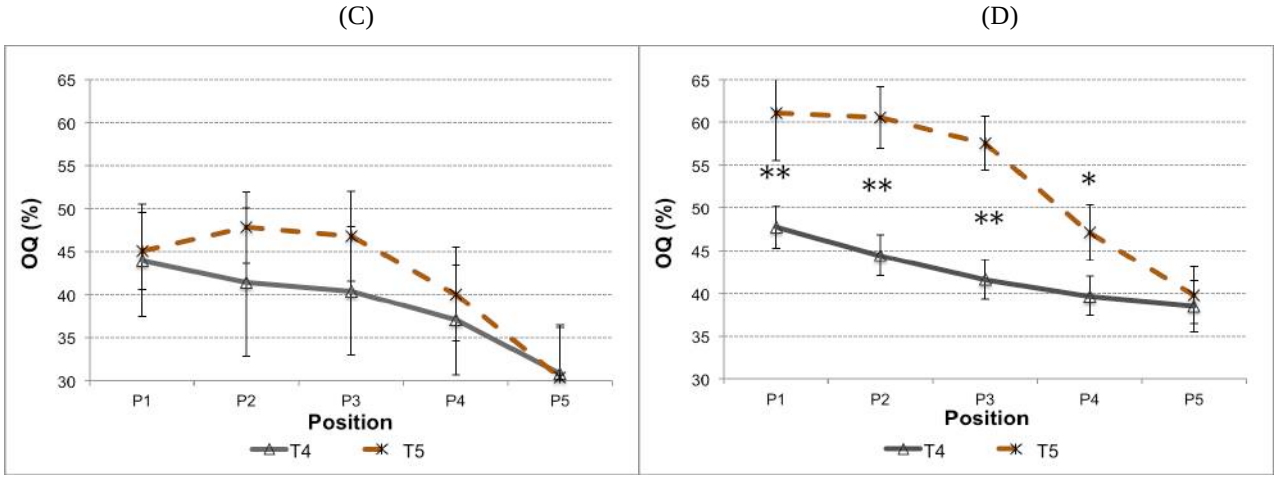


Figure 83. Average OQ according to tone and position in checked monosyllables (T4-5) for (A) young female (B) young male (C) elderly female and (D) elderly male speakers. Significance levels: * for $p<.05$; ** for $p<.01$ (higher OQ for T5 than T4).

One male speaker aged 25 produced very short checked syllables ending with irregular glottal periods, so that his data on checked syllables were not usable.

Since the number of speakers of each speaker group was rather limited and a large variation was observed between speaker groups, we ran by-item instead of by-subject ANOVAs for each speaker group, and separately for unchecked and checked syllables. In all analyses, *OQ* was the dependent variable, and *Item* (unchecked syllables: /Ø, p, (b), t, (d), f, (v), s, (z), m, n/-onset syllables; checked: /Ø, p, (b), t, (d), f, (v), s, (z)/-onsets syllables[17]) was the random factor. *Tone* (unchecked syllables: T1, T2, T3; checked syllables: T4 vs. T5), *Position* (P1, P2, P3, P4, P5) and *Speaker* (three speakers for the young female and elderly male speaker analyses; two speakers for the young male speaker analysis; there was no Speaker factor in the elderly female speaker analysis since there was only one elderly female speaker) were within-item factors.

We only report the results for the main factor, *Tone*, and, when *yang* syllables have a significantly higher *OQ* than *yin* syllables, those of *the Tone* × *Position* interaction. The *Tone* × *Speaker* interaction will also be reported in case of observed

[17] The zero onset was not included in the analyses for the elderly speakers, because the dEGG signal of the T5 /aʔ/ syllable was not useable for two out of the three speakers.

variations between speakers. Details on individual variations will be explained in the next section.

Let us first examine the unchecked syllables.

Tone was significant for female speakers, both young ($F(2,10)=11.8$, $p<.005$) and elderly ($F(2,10)=31.6$, $p<.0001$). Contrary to the predicted results, *OQ* was higher for *yin* than *yang* vowels; more precisely, *OQ* was higher when the F0 onset of the tone contour was higher. For the elderly female speaker, *OQ* was significantly higher for T1 than T2 (0.45>0.41, $F(1,5)=27.1$, $p<.005$) and marginally higher for T2 than T3 syllables (0.41>0.39, $F(1,5)=5.6$, $p=.064$). For the young female speakers, *OQ* was marginally higher for T1 than T2 (0.57>0.53, $F(1,5)=5.8$, $p<.061$) and numerically but not significantly higher for T2 than T3 syllables (0.53 vs. 0.51, $F(1,5)=3.7$, $p=.11$); the difference between T1 and T3 syllables (0.57 vs. 0.51) was significant, $F(1,5)=42.6$, $p<.005$. For the young male speakers, *Tone* was not significant overall, $F(2,10)<1$. The *Tone* × *Speaker* interaction was significant, $F(4,20)=34.5$, $p<.0001$, showing substantial individual variability in this group (see §4.2.2.3). For the elderly male speakers, *Tone* was significant overall, $F(2,10)=25.6$, $p=.0001$. *OQ* was significantly higher for T3 than T1 syllables (0.51>0;44),, $F(1,5)=34.4$, $p<.005$, or than T2 syllables (0.51>0.46), $F(1,5)=20.3$, $p<.01$. The *Tone* × *Position* interaction was significant, $F(8,40)=10.6$, $p<.0001$: the T3 vs. T1-2 differentials were greater in the first than the second half of the vowel (P1: $\Delta=0.088$; P2: $\Delta=0.100$; P3: $\Delta=0.075$; P4: $\Delta=0.033$; P5: $\Delta=-0.006$); the significance of the T3 vs. T1-2 differential is indicated for each time point in Figure 82D.

Now we turn to the results of the checked syllables.

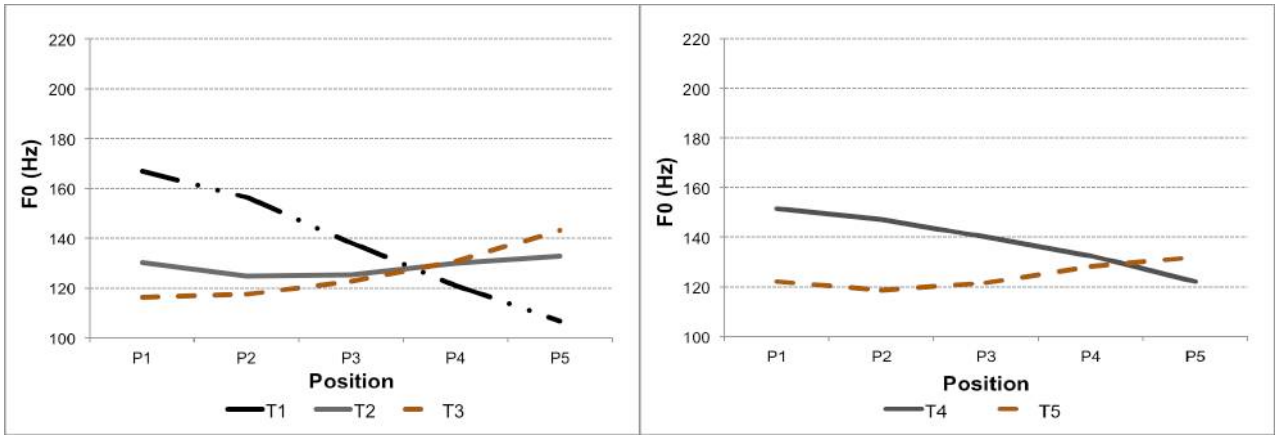
Tone was significant for young female speakers, $F(1,4)=8.1$, $p<.05$. Again, *OQ* was higher when the F0 onset of the tone contour was higher (T4 vs. T5: 0.53>0.51). For the elderly female speaker and the young male speakers, *Tone* was not significant (elderly female: $F(1,4)=3.7$, $p=.13$; young males: $F(1,4)<1$). The *Tone* × *Speaker* interaction was, again, significant for young male speakers, $F(1,4)=18.5$, $p<.05$ (see §4.2.2.3). For the elderly male speakers, *Tone* was significant overall, $F(1,4)=13.4$, $p<.05$: *OQ* was significantly higher for T5 than T4 syllables (0.53>0.42). The T5 vs. T4 differentials were larger in the first than the second half of the vowel (P1: $\Delta=0.133$;

P2: Δ=0.161; P3: Δ=0.159; P4: Δ=0.074; P5: Δ=0.013); the significance level is indicated for each time point in Figure 83D.

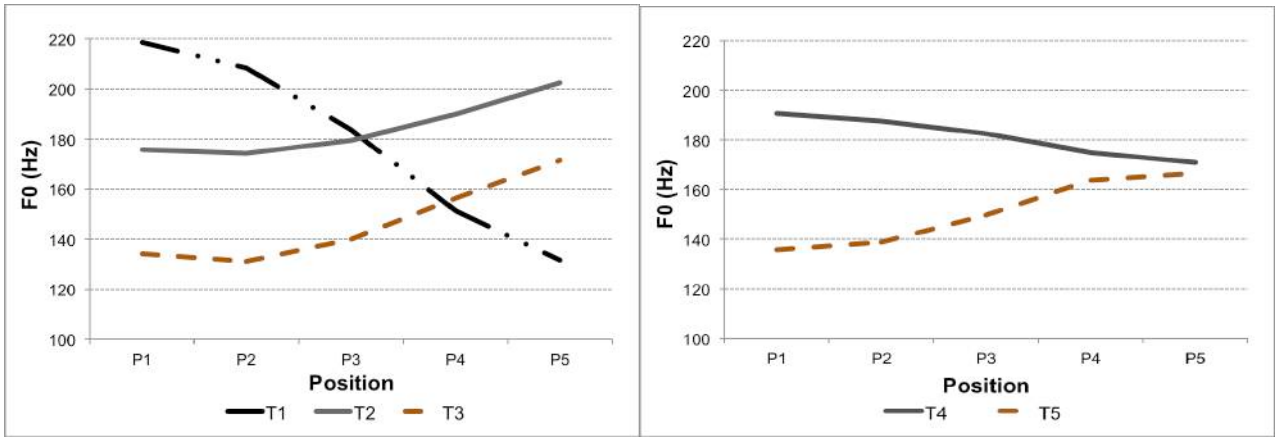
4.2.2.3 F0–OQ tradeoff

In the previous section, we saw that *Tone* had no overall effect on OQ for the three young male speakers (two aged 24, and one aged 25). However, if we take a closer look at their individual data, we observe large inter-speaker variations. In particular, one of the three speakers (hereafter Speaker 1) showed the same pattern as the elderly male speakers, that is significantly higher OQ for *yang* than *yin* vowels (0.56> 0.50), $F(1,5)=22.5$, $p=.005$; another one (hereafter Speaker 2) showed a similar pattern to that of the young female speakers, that is, significantly higher OQ for *yin* than *yang* vowels (0.54>0.48), $F(1,5)=15.6$, $p<.05$; the third one (hereafter Speaker 3) showed no significant OQ difference between *yin* and *yang* vowels (0.53≈0.54), $F(1,5)<1$. If we compare the OQ patterns with the F0 ranges, a trend for a tradeoff between F0 range and OQ emerges. Speaker 1 had the smallest F0 range and higher OQ for *yang* than *yin* vowels: we may speculate that he produced a rather weakly contrastive F0 contour information which was compensated by a clear phonation type difference between the *yin* and *yang* tone registers. Speaker 2 had the largest F0 range and a reversed OQ pattern compared to elderly male speakers: we may speculate that he did not contrasted *yang* from *yin* rimes in producing breathiness for the *yang* type; rather, the OQ values they produced correlated with tone contour onset F0, just like for the female speakers. Between these two extremes, Speaker 3 had an intermediate F0 range and did not exhibit any OQ differences between *yin* and *yang* tones. F0 and OQ trajectories are shown in Figure 84 and Figure 85 for (a) for Speaker 1, (b) Speaker 2, and (c) Speaker 3 for unchecked syllables only. (Speaker 3's dEGG data on checked syllables were hardly usable.)

(a) Speaker 1



(b) Speaker 2



(c) Speaker 3

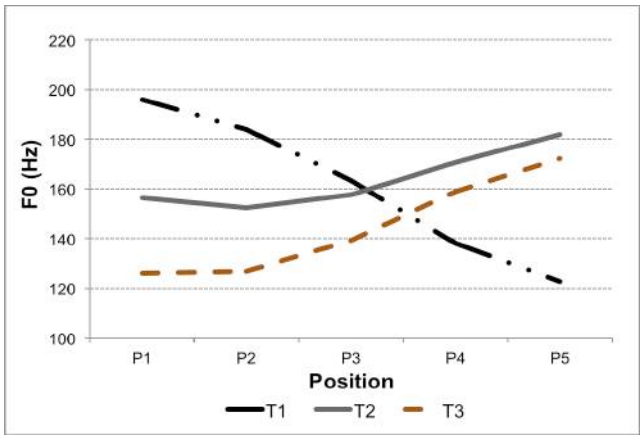
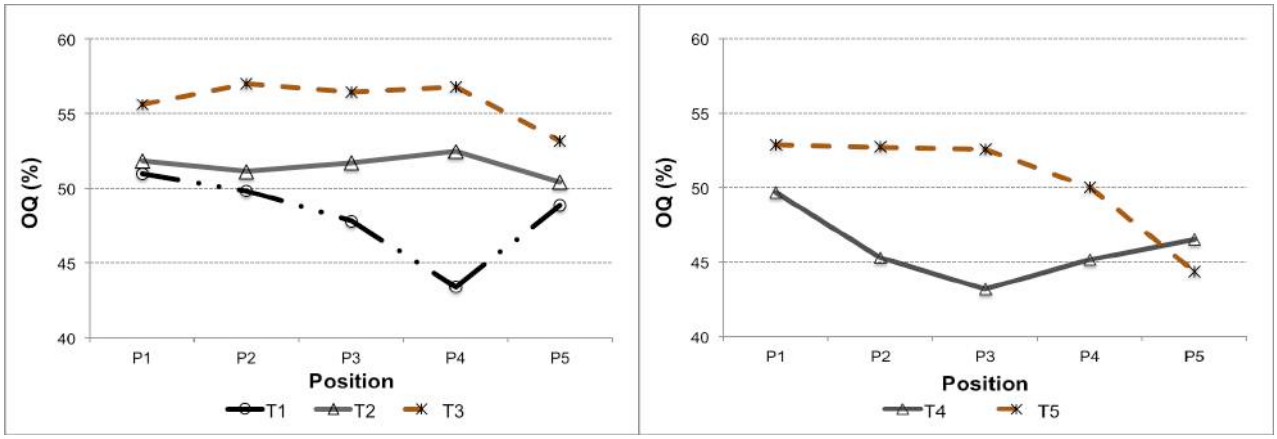
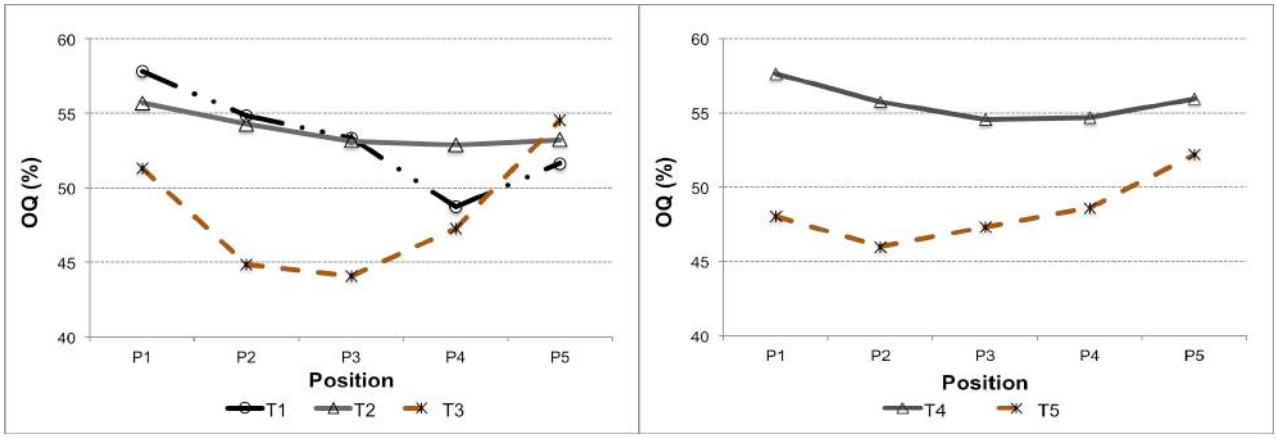


Figure 84. Time normalized F0 trajectories for each of the three young male speakers. Left: unchecked syllables; Right: checked syllables.

(a) Speaker 1



(b) Speaker 2



(c) Speaker 3

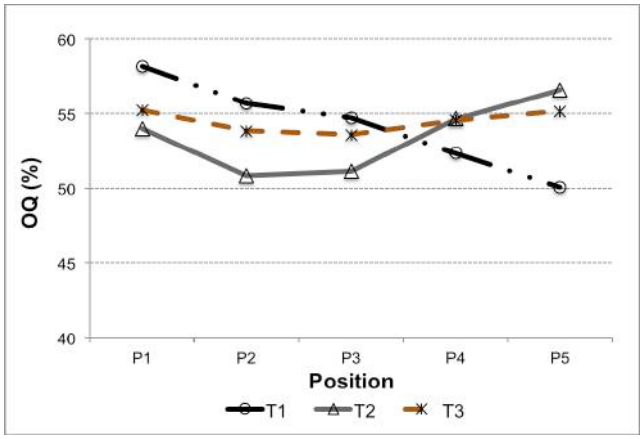


Figure 85. Time normalized OQ trajectories for each of the three young male speakers. Left: unchecked syllables; Right: checked syllables.

4.2.2.4 Variable realizations of breathy voice

Not only is there inter-speaker variation in that some speakers produced breathy voice with *yang* tones and others did not, but the production of breathy voice itself was variable as well in that different strategies were used to produce the "breathy voice." Furthermore, intra-individual variation was also observed: the same speaker may use different strategies for different items or different occurrences.

Two of the elderly male speakers (aged 66 and 72) sometimes produced the *yang* tone syllables, especially the zero-onset ones, with a voice quality that can be qualified as "harsh whispery" voice.[18] This voice quality was not observed in the other speakers. Figure 86a-b compares the waveforms, the spectrograms, and the EGG signals between a tone T2 modal voice /ɛ/ syllable and a tone T3 "harsh whispery" /ɛ/ syllable produced by the 66-year-old male speaker. Figure 86c shows a short term spectrum within the "harsh whispery" syllable produced by the 72-year-old male speaker. Both harsh and whisper imply a constricted epilaryngeal tube. Harsh voice adds an aperiodic component resulting in frequency and intensity perturbation, and whispery voice adds noise to the speech signal, as can be observed in Figure 86b. Sub-harmonics in the lower frequencies are observed in the spectrum of the T3 syllable in Figure 86c: these sub-harmonics are the main contributors to the perception of harshness (or roughness) (Omori, Kojima, Kakani, Slavit, & Blaugrund, 1997) and probably correspond to the vibration of the ventricular folds and/or the aryepiglottic folds (Esling, personal communication). Sub-harmonics can also occur due to asymmetric vocal fold vibrations, which are called "bifurcations" (Titze, 1994: 298). But the EGG signal of the syllables shown in Figure 86b are more likely due to harsh voice: indeed, a part of the EGG signal is very weak, suggesting great glottal impedance, perhaps due to other vibration sources than the glottis.

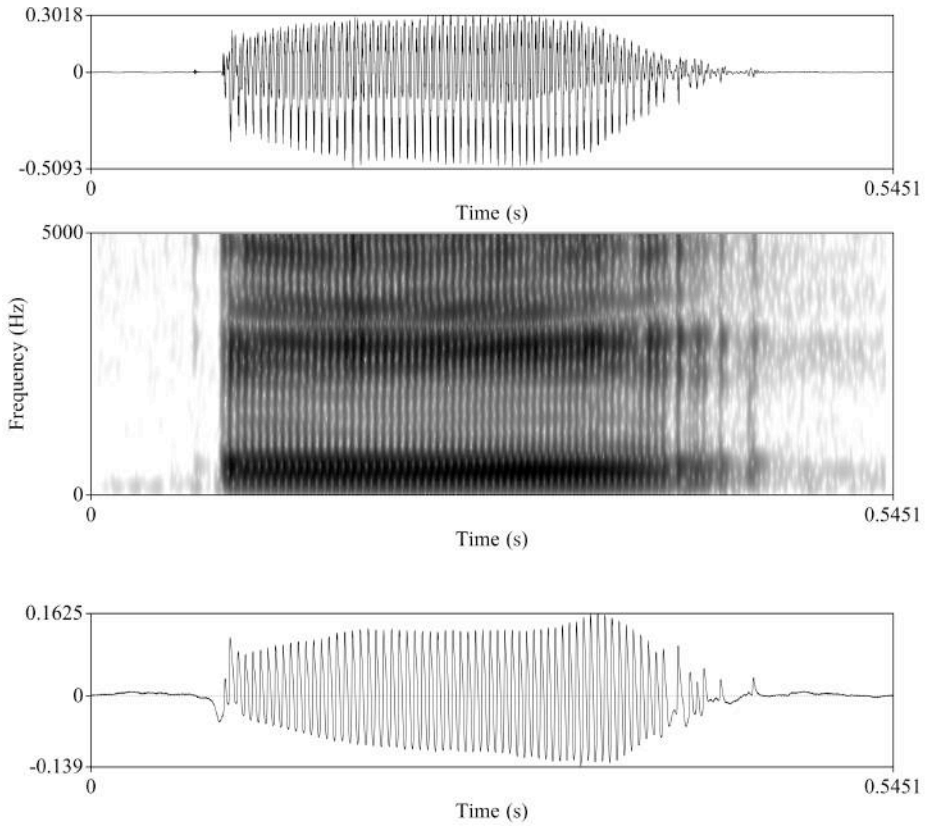
The 72-year-old speaker was born in Shanghai urban area, but had stayed in the 崇明 Chongming island for nearly 30 years. His parents were from 海门 Haimen (Jiangsu province), which is located at the opposite side of the Yangtze River from the

[18] I am grateful to John Esling and Scott Moisik for our discussions on the EGG and acoustic signals of these two male speakers.

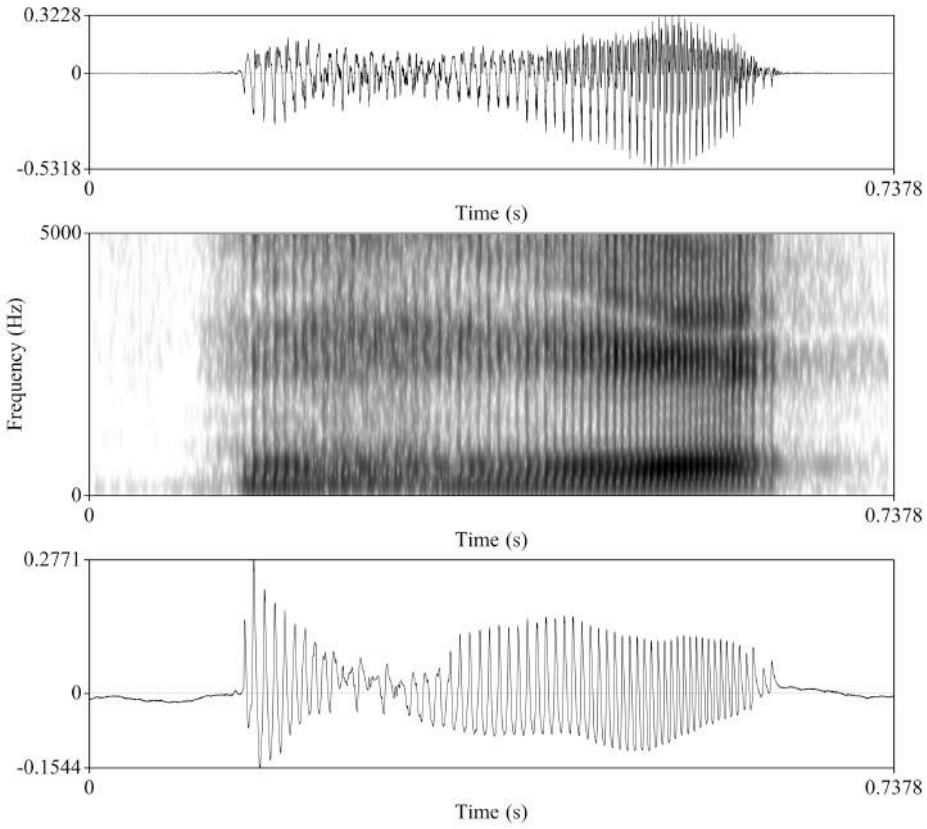
Chongming island. Besides Shanghai Chinese, he also spoke a variety of Haimen Wu dialect spoken in Chongming which is very similar to the Chongming dialect. As far as we know, no study has reported this "harsh whispery" voice quality in dialects spoken in Chongming or Haimen. In any case, given that the other speaker using this voice quality had never been out of Shanghai for more than six months and only had limited knowledge of some Zhejiang dialects, this voice quality is probably not due to some interference with the Haimen dialect.

It is probably not a voice quality related to age, either. The 66-year-old speaker was relatively young and was an amateur Shanghai opera singer. His voice was perfectly clear and modal when he produced *yin* tone syllables, and this "harsh whispery" voice only occurred with the *yang* tone syllables, especially in citation form.

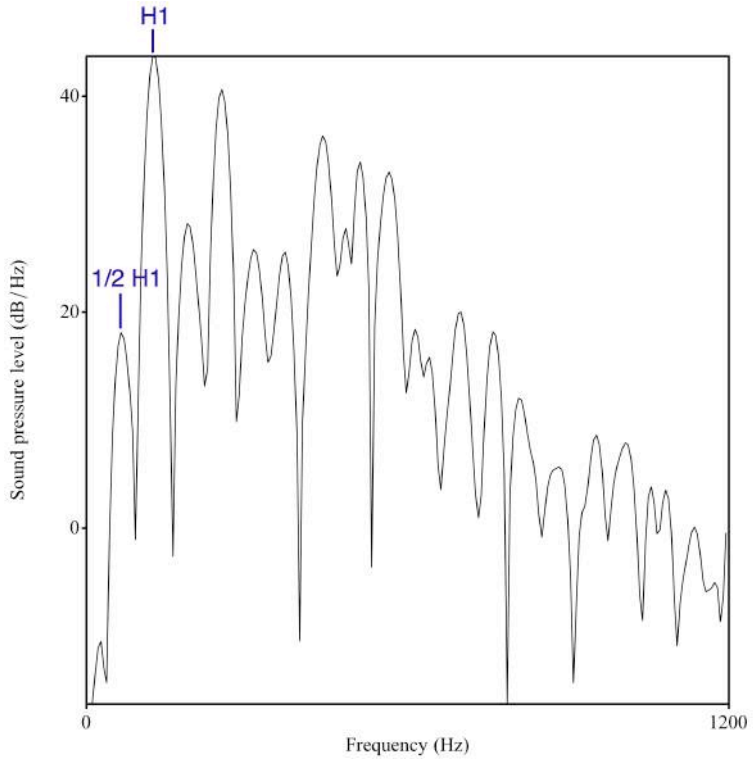
The "harsh whispery" voice is apparently quite similar to a voice quality in 镇海 Zhenhai dialect, a Northern Wu dialect spoken in the Zhenhai County in 宁波 Ningbo (Zhejiang), which Phil Rose called "growl" voice (Rose, 1982b, 1989). In this dialect, *yang* tone syllables are realized with whispery voice, and more specifically, *yang* tone syllables with open oral vowels or nasalized vowels are accompanied with this "growl" voice. Phil Rose describes this voice quality as "diplophonic," giving "an auditory impression of a series of strong pulses" and of harshness. A preliminary fiberoptic investigation of his own imitation of the Zhenhai "growl" voice shows extreme epiglottalization: violent vibration of the whole larynx structure and the epiglottis, which could be the source of the perceptible diplophonic pulses.


0.3018
0
-0.5093
0
0.5451
Time (s)
5000
Frequency (Hz)
0
0
0.5451
Time (s)
0.1625
0
-0.139
0
0.5451
Time (s)

(a)


0.3228
0
-0.5318
0
0.7378
Time (s)
5000
Frequency (Hz)
0
0
0.7378
Time (s)
0.2771
0
-0.1544
0
0.7378
Time (s)

(b)



(c)

Figure 86. (a) – (b): Comparison of waveforms, spectrograms, and EGG signals (from top to bottom) between tone T2 and T3 /ɛ/ syllables produced by a 66-year-old male speaker; (c): short-term spectrum (150 ms analysis winow) of a "harsh whispery" portion in a tone T3 /ɛ/ syllable produced by a 72-year-old male speaker. (CD: 4.2.2.4_fig87)

NB: In the EGG signal, positive values correspond to glottal closure and negative values to glottal opening.

Whereas Phil Rose suggests a systematic production of whispery voice or "growl" voice quality depending on the vowel in Zhenhai *yang* syllables, the production of "harsh whispery" voice is more idiosyncratic in Shanghai Chinese. Phil Rose cites Ladefoged (1983) who wrote "one person's voice disorder is another person's phoneme" to comment on the "pathological–like" Zhenhai growl. Our data suggest this observation applies not only to different phonological systems, but also within one language across different speakers, or even within one speaker across different utterances, as shown by the inter- and intra-speaker variation in Shanghai speakers.

4.2.3 Discussion

In this articulatory investigation of Experiment 1, we attempted to complement our acoustic data with evidence from vocal fold behavior. We used dEGG, the derivative of the EGG signal (the EGG signal monitors vocal fold contact area) for an accurate measurement of the successive cycles (and hence F0) and a good estimation of the glottal open quotient (OQ).

Compared to the five time points' acoustic analysis of F0, a more precise description of F0 trajectories shows inter-speaker variations of tone realization. The difference mainly lies in the timing and degree of F0 rise in the production of the two unchecked rising tones T2 and T3, as well as in the variation in (checked) *yin* T4 contour onset F0 relative to the contour onset F0 for the unchecked *yin* tones T1 and T2.

The OQ results measured from the dEGG signal show cross-gender and cross-age variations in the production of the phonation types that may index the *yin–yang* distinction. In agreement with the acoustic measures of breathiness, OQ was higher for *yang* than *yin* vowels most significantly for elderly male speakers, indicating a breathier phonation for *yang* than *yin* vowels in this speaker group. The *yin–yang* difference was also consistently found larger at the beginning than the end of the vowel. For young male speakers, between-speaker variations were so large that some speakers produced breathier voice on *yang* than *yin* vowels and others not or even showed the opposite pattern. It turns out that those speakers with higher OQs for *yang* than *yin* syllables had a small F0 dynamics, while speakers with the opposite pattern, or those who exhibited no *yin–yang* OQ differences, had a larger F0 dynamics. A similar F0–OQ tradeoff pattern is observed in Risiangku Tamang (Tibeto-Burman), in which tone contour and phonation type co-exist: speakers who produce clear F0 contrasts tend to produce weak phonation differences, and vice versa (Mazaudon, 2012). For the elderly female speaker of our study, no *yin-yang* difference in OQ was found. This is probably due to the limited number of speakers (n=1). Inter-speaker variation certainly could also be found among elderly female speakers, provided that sufficient data is collected. In line with the F0-OQ tradeoff hypothesis, our elderly female speaker used a quite wide F0 range, so that she did not have to

resort to the use of breathy phonation for *yang* syllables. Young female speakers did not produce *yang* syllables with higher OQs than for *yin* syllables, that is, with a breathier voice. On the contrary, their OQs were higher for higher F0 syllables, that is, *yin* syllables, a pattern also found in our elderly female speaker but for unchecked syllables only. This pattern of covariation is in line with the finding of a weak positive OQ–F0 correlation by Holmberg et al. (1989) and Koreman (1996, [cited in Iseli et al., 2006]).

Finally, inter- and intra- speaker variations were also observed in the way *yang* syllables were signaled by voice quality. The EGG data, in line with the acoustic data, showed that two elderly male speakers occasionally produced "harsh whispery" voice on *yang* vowels.

Vocal fold contact area is only part of the thoroughly complex picture of the laryngeal configurations, described as a "system of valves" by Edmondson and Esling (2006). The description of variable realizations in terms of glottal articulation requires more invasive techniques, such as fiberoptic endoscopy. Moreover, in order to obtain complementary data (acoustic, articulatory, aerodynamic), the way forward lies in multi-sensor recordings, as suggested in Vaissière, Honda, Amelot, Maeda, and Crevier-Bushman (2010), combining different data at the same time, such as EGG, fiberoptic images, airflow, motion capture system and video data (see §4.3).

4.3 Experiment 2: articulatory investigation of closure duration

As explained in the preceding section on the acoustic analyses conducted for Experiment 1, it is not possible to measure the closure duration of word-initial phonetically voiceless stops on the sole basis of the acoustic signal. Articulatory measurements are required to estimate such closure durations. For example, electropalatography (EPG) could be used to measure closure duration in absolute initial position, though only for coronal stops, as was done, for example, to compare the closure durations of singleton vs. geminate utterance-initial stops in Tashlhiyt Berber (Ridouane, 2007). One drawback of this method, however, is the time-consuming (and costly) manufacturing of custom artificial palates for each speaker.

In this section, we report a pilot articulatory study, which used a motion capture system (Qualisys) able to capture lips' movements. This methodology has been used to study lip rounding and lip aperture during vowel articulation (Georgeton & Fougeron, 2014). In our study, we used the Qualisys motion capture system to measure and record the movements of the upper and lower lips during the production of stop consonants. We propose a method to determine stop closure duration from the lips movement data. Of course, the limit of this method is that only labial stops can be analyzed.

4.3.1 Method

4.3.1.1 Qualisys motion capture system and recording procedure

Qualisys (http://www.qualisys.com/) is an optical motion capture system that captures three-dimensional movements by means of high-speed infra-red cameras.

As shown in Figure 87, four reflecting markers with a diameter of 2.5 mm were fixed on the external contour of the speaker's lips, at the midpoint of the upper and lower lips, and at the left and right corners of the lips. Five reference markers with a diameter of 4 mm were fixed on a helmet worn by the speaker to compensate for head movements, one marker at the top of the head and the other four markers around the forehead. Four infra-red cameras were arranged in a semi-circle, with the speaker seated at its center. These cameras recorded the movements of the reflecting markers at a 100 Hz frame rate. The wand calibration method was used for the camera system, using as a reference structure a stationary L-shaped object with four markers attached to it, together with a moving wand (with two markers at its extreme ends). The reference structure determined the coordinate system, with the longer arm as the x-axis, the shorter arm as the y-axis and the direction perpendicular to the two arms as the z-axis, with the x-y plane parallel to the floor. The wand was moved by the experimenter in the three directions to assure that the axes were correctly calibrated.

After the calibration, followed by a training session, the audio recording system was launched just before the Qualisys system, which generates triggers every 10 milliseconds in order to synchronize the audio signal with and the motion capture

data. The audio recordings were made with an AKG C520L headband microphone through an external Digidesign sound card connected to a computer, using the Protools software. The audio signal was recorded on one channel, and the triggers signal on a second channel. Both signals were digitized at a 44,100 sampling rate, with 16 bit precision and stored in WAV format files. The recordings took place in the sound-treated studio of the Institut de Linguistique et Phonétique Générales et Appliquées (ILPGA) at Université Sorbonne Nouvelle – Paris 3.

The lip movements were recorded and subsequently analyzed by the Qualisys Track Manager (QTM) software. We focused on the measure of lip aperture as an indication of closure duration, which means that the two markers on the upper and lower lips were sufficient to achieve the analyses relevant to our purpose. The data from the other markers might be of use for future analyses.

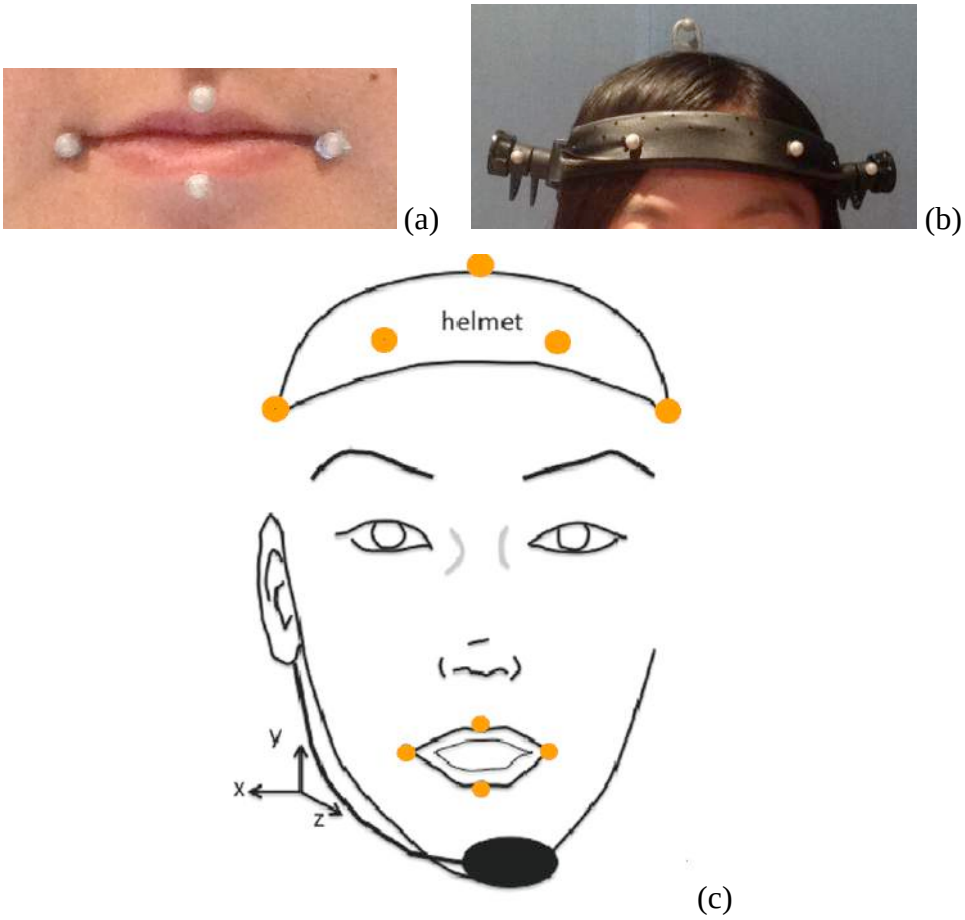


Figure 87. The reflecting markers of the motion capture system: (a) lips markers, (b) helmet reference markers, and (c) schematic view of the markers' position (adapted from Georgeton & Fougeron, 2014).

4.3.1.2 Participants

Two female native speakers of Shanghai Chinese, including the author, participated in the recordings.

The author (Speaker 1) was a 27-year-old female at the time of the recording, born in Shanghai suburbs and raised in Shanghai urban area. She stayed in Shanghai until age 21, and lived in France since then. She began to learn English at primary school and French at university. She frequently used Standard and Shanghai Chinese in Shanghai. Since she lived in France, she used more frequently French than any other language but still used Chinese (Standard or Shanghai) on a daily basis.

Speaker 2 was a 40-year-old female at the time of the recording, born and raised in Shanghai urban area. She stayed in Shanghai until age 35 and in France since then. She used Shanghai Chinese as frequently as Standard Chinese before moving to France. Since she lived in France, she used much more frequently Standard than Shanghai Chinese. She self-evaluated quite positively her competence in Standard and Shanghai Chinese (5 for both, on a 1-5 scale), as well as in French and English (4 for both, on the same scale).

4.3.1.3 Speech materials

36 monosyllabic words were recorded in a carrier sentence /i dʊʔ _ gəʔ əʔ zʐ/ (伊读 ___ 这个字。'He reads ___ this character'). The target syllables, listed in Table 31, were composed of a labial stop onset /p/ or /b/ both phonetically realized as voiceless [p] and 13 rimes compatible with the onset [i, ɛ, a, ø, u, o, ɔ, ɪʔ, əʔ, ʊʔ, ɑŋ, əŋ, ɪŋ], at the five tones (T1-3 for the 10 unchecked syllables and T4-5 for the six checked syllables). There were thus 36=3*10+2*3 syllable items. Each syllable was repeated four times, yielding a total of 36*4=144 syllables. The meaning of each monosyllabic word can be found in Appendix 1.

The target syllable was preceded by other syllables in order to avoid that the lip closure during the initial silence be taken as stop closure. The speakers were instructed to read the sentences slowly and to lightly stress (without exaggeration)

the target syllable, so as to avoid potential phonetic voicing of *yang* stops. Indeed, the data showed that they realized these stops as phonetically voiceless.

Table 31. Speech materials of Experiment 2: syllables with labial stop onsets.

Tone	phonemic	phonetic	i	ɛ	a	ø	u	o	ɔ	ɑŋ	əŋ	ɪŋ	ɪʔ	ʊʔ	aʔ
T1	/p/	[p]	边	班	巴	搬	波	疤	包	帮	奔	冰			
T2	/p/	[p]	比	板	摆	半	布	把	宝	榜	本	饼			
T3	/b/	[p]	皮	办	排	盘	步	爬	暴	碰	笨	病			
T4	/p/	[p]											笔	剥	百
T5	/b/	[p]											鼻	薄	白

Each recording session was divided into four blocks, in which the 36 sentences were presented in random order. Speaker 1 read the sentences in one order and Speaker 2 in the reversed order. Both speakers took a break after two blocks. Speaker 2 repeated the first two blocks a second time, because the percentage of markers' identification was not satisfactory during the first repetition. In fact, it turned out that only her second repetition of the first two blocks could be analyzed.

4.3.1.4 Data analyses

A first pass of pre-processing was completed by the Qualisys Tract Manager (QTM) software. It consisted of identifying the different markers and labeling them. We labeled the five markers on the helmet "casque1", "casque2" … "casque5", the two markers at the midpoint of the upper and lower lips "levreH" and "levreB", respectively, and the two markers at the left and right corners of the lips "levreG" and "levreD", respectively. When the cameras correctly detected a marker 100% of the time, this marker simply needed labeling. When the cameras occasionally missed a marker, its movement was divided into several time periods. Visual inspection of the several resulting movement curves was then needed in order to put together the time periods corresponding to the movement of the concerned marker. The marker could then be labeled.

The raw 3D data from this pre-processing step were exported in ".mat" files, readable in Matlab. A given such file consisted of a matrix in which the successive rows corresponded to the successive time frames; for each marker, three consecutive columns were used to code its 3D position on the X, Y, and Z axes (in millimeters) and a fourth column contained the average of the residuals of these 3D coordinates, as an indicator of the detection precision of the marker's position.

A first Matlab program was used to synchronize the audio data with the markers' movement data. Its output was the vertical distance between the lower and upper lips, that is, roughly, lip aperture, as a function of time. A second Matlab program resampled the lip-aperture data at 44,100 Hz for ease of alignment with the audio signal, and computed the first derivative of the lip-aperture signal. The same reasoning as for the dEGG signal applies here: a negative peak of the lip-aperture derivative corresponds to an instant of maximum velocity in lip-aperture decrease, and roughly signals the beginning of a labial stop closure, just like the dEGG negative peaks are taken to signal glottis closure. Conversely, a positive peak of the lip-aperture derivative corresponds to an instant of maximum velocity in lip-aperture increase and roughly signals the onset of a labial stop release, just like the dEGG positive peaks are taken to signal glottis opening. Here, two positive peaks are generally observed in the vicinity of labial stop release (see Figure 88). The first peak presumably truly corresponds to the onset of stop release (i.e., the end of the closure), and the second peak might correspond to a second stage of lip aperture for the production of the following vowel. The Matlab program looked for a succession of a negative peak followed by a positive peak within the manually segmented interval between the target syllable and the preceding vowel (in the sentence /i dʊʔ _ gəʔ əʔ zẓ/) and computed closure duration as the distance between those two peaks. In some utterances, the first positive peak was too weak and the second peak was detected instead, thus requiring a manual correction, based on the acoustic signal (Figure 89). The closure duration estimated this way is, however, probably longer than the real duration.

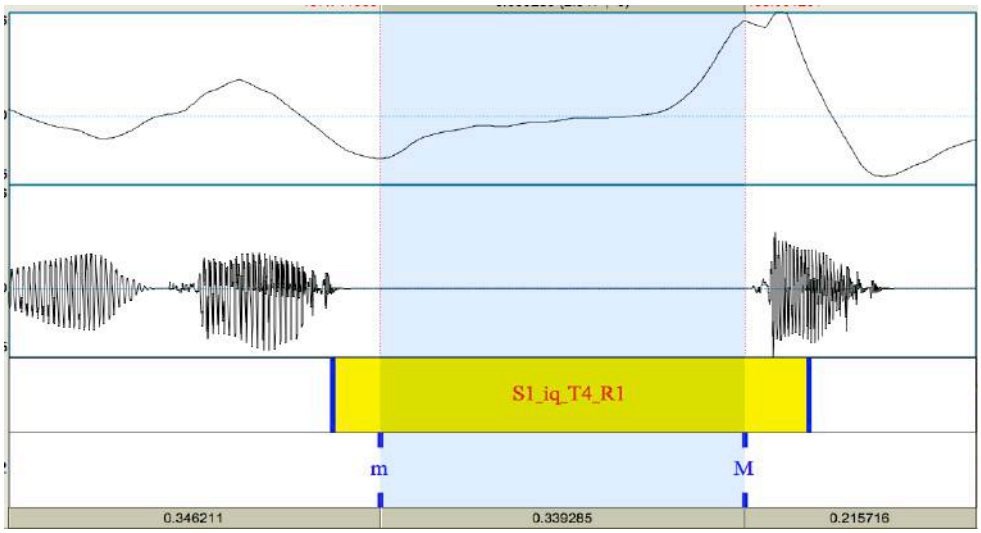


Figure 88. Closure duration of the labial stop onset of [pɪʔ] (T4). Lip-aperture derivative (top), speech wave, and segmentation tiers (first tier manually segmented; second tier computed: 'm' and 'M' correspond to the instants of closure and release, respectively.

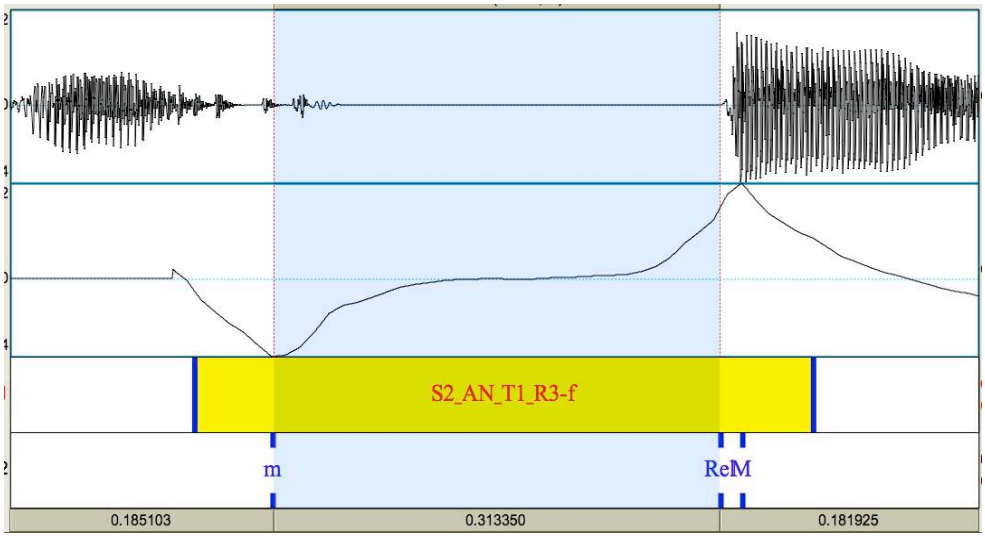


Figure 89. Same as **Figure 88**, for a syllable [pɑŋ] (T1). The computed 'M' missed the stop release, which has been manually corrected to 'Rel', based on the speech waveform.

4.3.2 Results

We report here the data of the four repetitions produced by Speaker 1 and of the two repetitions produced by Speaker 2. Figure 90 shows the target syllable's stop onset closure duration data for each speaker, as a function of syllable tone. Overall, these

closure durations are numerically longer in *yin* than *yang* unchecked syllables for both speakers. However, for checked syllables, this pattern is found only for Speaker 1.

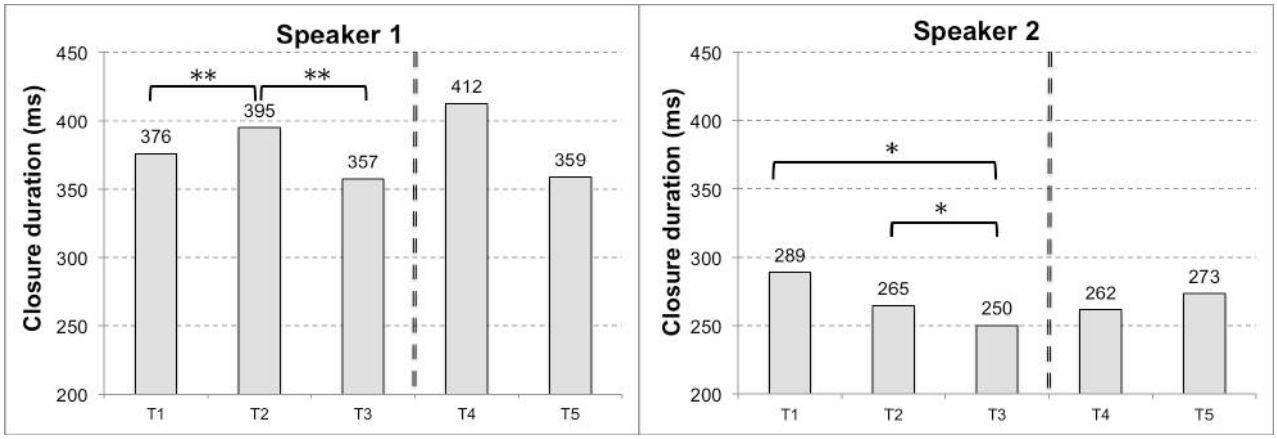


Figure 90. Labial stop closure durations in target syllables according to tone and speaker, pooled across vowel contexts and repetitions. Significance levels: * for $p<.05$; ** for $p<.01$.

To substantiate these observations, we ran two by-item ANOVAs for each speaker, one for the unchecked and the other for the checked syllables. *Duration* was the dependent variable, and *Item* the random factor (whose levels corresponded to the different rimes: 10 or 3 for unchecked and checked syllables, respectively). *Tone* (unchecked: T1, T2, T3; checked: T4 vs. T5) and *Repetition* (Speaker 1: repetitions 1-4; Speaker 2: repetitions 1-2) were within-item variables.

For unchecked syllables, *Tone* was significant overall for both speakers (Speaker 1: $F(2,18)=8.6$, $p<.005$; Speaker 2: $F(2,18)=7.0$, $p<.01$). For Speaker 1, *Duration* was the longest for T2: T2 vs. T1: 395>376 ms, $F(1,9)=11.8$, $p<.01$; T2 vs. T3: 395>357 ms, $F(1,9)=16.7$, $p<.005$. Yet, the difference between T1 and T3 was not significant, $F(1,9)=2.6$, $p=.14$. T3 was however significantly shorter than T1 and T2 pooled together, $F(1,9)=7.9$, $p<.05$. The *Tone* × *Repetition* interaction was significant. *Tone* was significant for the first three repetitions (R1: $\Delta=52$ ms, $F(2,18)=6.9$, $p<.01$; R2: $\Delta=60$ ms, $F(2,18)=11.8$, $p=.0005$; R3: $\Delta=33$ ms, $F(2,18)=3.7$, $p<.05$) but not for the fourth one ($\Delta=-34$ ms, $F(2,18)=1.1$, $p=.35$). For Speaker 2, *Duration* was the shortest for T3: T3 vs. T1: 250<289 ms, $F(1,9)=8.3$, $p<.05$; T3 vs. T2: 250<265 ms, $F(1,9)=7.5$, $p<.05$. The difference between T1 and T2 was marginal, $F(1,9)=5.0$, $p=.052$. Finally, the *Tone* × *Repetition* interaction was not significant, $F(1,9)<1$.

For checked syllables, *Tone* was not significant for either Speaker 1 ($F(1,2)=3.7$, $p=.19$) or Speaker 2 ($F(1,2)<1$). For Speaker 1, the lack of significance was possibly

due to the small number of items (n=3). A pairwise *t*-test between all items of all repetitions show significantly higher closure duration for T4 than T5 stops for Speaker 1, $t(11)=8.3$, $p<.05$.

4.3.3 Discussion

In this section, we used the Qualisys motion capture system to investigate the possible differences in (labial) stop closure duration between *yin* and *yang* syllables in word-initial position. We measured the temporal evolution of the medial distance between the upper and lower lips, that is, roughly, lip aperture. We based our estimation of the instants of stop closure onset and release on the negative and positive peaks, respectively, of the derivative of the lip-aperture function of time.

For unchecked syllables, the overall tendency of longer closure durations for *yin* than *yang syllables* was observed for both speakers. The difference was significant between each *yin–yang* pair and for all repetitions for Speaker 2, but it was less consistently significant for Speaker 1. For checked syllables, stops had longer closure durations for *yin* than *yang* syllables for Speaker 1 only. The difference was not significant on a by-item ANOVA with only three levels for the random factor. It was significant on a *t*-test between all the T4 items and all the T5 items. For Speaker 2, the *yin*>*yang* pattern was mysteriously reversed, though not significantly so.

We thus cannot conclude at this stage that obstruents are longer for *yin* than *yang* syllables in all positions. Such a trend was observed but more data from more speakers need to be collected to confirm our results, which should be viewed as preliminary data. The methods also need to be improved, for example, by combining Qualisys data with video data, and by using only unrounded vowels since lips' vertical distance is affected by rounding and protrusion.

Finally, the only way to avoid the lack in statistical power for the checked syllable category would be to test many more subjects and run by-subject analyses of variance. Indeed, we used all the possible rimes occurring with the /p/ onset for checked syllables in Shanghai Chinese. And there are only three of them.

5 EXPERIMENTAL INVESTIGATIONS OF THE PERCEPTION OF "YIN" VS "YANG" TONE SYLLABLES: MAIN AND SECONDARY CUES

Summary

This chapter reports three perception experiments whose goal is to evaluate the relative contributions of the main and secondary correlates of the *yin* vs. *yang* distinction in Shanghai Chinese monosyllables.

In Experiment 3 (§5.1), we looked whether other phonetic properties than F0 contour and height influence syllable identification in minimal triplets of unchecked monosyllables differing by tone only (T1, T2, or T3). We manipulated the F0 contour of such monosyllables while maintaining all the other phonetic properties. We found that syllable identification was mainly determined by the imposed F0 contour, except for syllables with a voiced labial fricative onset. Yet, response times tended to increase and syllables tended to be judged less natural when the imposed contour differed from the original one.

In Experiment 4 (§5.2), we manipulated the consonant/vowel duration pattern and created *yin–yang* tone contour continua from T2 to T3 or from T4 to T5. The duration pattern manipulation influenced syllable identification in that high C/V duration ratios induced more frequent and faster *yin* responses. We conclude that the C/V duration pattern contributes to the perception of the *yin* vs. *yang* tonal identity.

In Experiment 5 (§5.3), we manipulated the voice quality of synthesized or natural unchecked monosyllables and created *yin–yang* tone contour continua from T2 to T3. The voice quality manipulation influenced syllable identification in that breathy voice induced more frequent and faster *yang* responses, except for nasal onset syllables. We conclude that breathy voice still is an important cue for the perception of low tone syllables.

In this chapter, we investigate the perception of Shanghai tones and of their acoustic correlates. We ask whether the acoustic cues correlated with Shanghai tones contribute to their identification and to which extent.

Perceptual studies on Shanghai Chinese are relatively rare, among which Cao (1987) and Ren (1992). Cao (1987) studied the perception of voicing by phoneticians and phonologists familiar with Wu dialects and by a Chinese student in phonetics from Beijing with no knowledge of Chinese phonology in general and Wu dialects in particular. Cao found that the listeners familiar with Wu dialects tend to perceive voiced onsets when the syllable onset has a low F0 (as in *yang* tones) even though all the (monosyllabic) stimuli had voiceless onsets. The listener with no knowledge of Wu dialects perceived correctly all the stimuli as beginning with a voiceless onset. Cao concluded that the perception of a [+voice] feature by her linguist subjects was motivated by its *yang* tone identity. The study of Ren (1992) bore on the perception of breathy voice. His study will be summarized in §5.3, where we present the results of a study we ran, which asked similar questions.

Using monosyllables as stimuli in our experiments, we investigated the role of durations and of voice quality in tone perception. In the first experiment (Experiment 3), we look whether acoustic characteristics other than pitch contour affect tone perception. In the following experiments (Experiments 4 and 5), we further explore the role of segmental durations and of voice quality in tone perception.

5.1 Experiment 3: Is F0 contour the unique cue for tone perception?[19]

To test for the possible role in tone perception of acoustic characteristics other than F0 contour (which we call here "non-tonal" characteristics for sake of simplicity), we constructed syllables combining an F0 contour and "non-tonal" characteristics that were either "congruent" or "incongruent." The rationale was that if tone perception is

[19] Part of the results presented in this section have been reported at the 14th Annual Conference of Interspeech in 2013 in Lyon and published in the proceedings of the conference.

solely based on F0 contour, it should not be influenced by the congruency of the non-tonal characteristics. Here, we focus on the *yin* vs. *yang* tone register distinction in Shanghai Chinese, with typically high vs. low F0 tone contours, respectively. In "congruent" syllables, the naturally produced non-tonal acoustic characteristics of a syllable produced with a given pitch register were combined with an F0 contour matching that pitch register. In "incongruent" syllables, the non-tonal acoustic characteristics of a syllable produced with a given pitch register were combined with an F0 contour matching the opposite tone register. An example of incongruent syllable is /dɛ/ (*yang* tone T3) on which the F0 contour of /tɛ/ (*yin* tone T1 or T2) is imposed, maintaining all the other acoustic characteristics of /dɛ/. An example of congruent syllable is /dɛ/ (*yang* tone T3) on which a stylized T3 F0 contour (see below) is imposed. In this experiment, the stimuli we constructed were based on minimal triplets of unchecked monosyllabic words, which only differed in tone (T1, T2 and T3). The syllables of these minimal triplets did not differ in the phonetic voicing of their onset, except the T3 syllable /vɛ/, in which the onset was naturally produced with phonetic voicing, as is often observed (see §4.1.5.2). The first test was a forced choice identification task. On each trial, participants were presented with one of the stimuli (congruent or incongruent) and asked to identify its tone by choosing between two Chinese characters that made a *yin-yang* minimal pair in Shanghai Chinese. In the second test, participants were presented on each trial with one of the same auditory stimuli, together with a Chinese character matched with it in terms of tone contour and segments: in other words, they heard a "pronunciation" of the character they saw; the participants were asked to rate on a 1-5 scale how natural was the pronunciation they heard of the displayed character.

The goal of the identification test was to determine whether non-tonal information plays a role in *yin* vs. *yang* tone perception, and if yes, to what extent listeners rely on this information to identify a syllable. The naturalness-rating asked a similar question in a different way: is incongruent non-tonal information perceived as such and thus influences subjective judgments on a syllable's pronunciation?

We predict that the F0 contour is the main cue to tone identity and is the dominant factor in the listeners' distinction between the *yin* and *yang* registers.

Accordingly, the response accuracy in tone identification should depend on the F0 information much more than on the other "non-tonal" accompaniments, such as duration patterns or voice quality.

Two hypotheses can be made concerning the role of non-tonal cues in tone identification.

H1. Listeners identify a *yin* or *yang* syllable solely based on F0 contour, and do not use the other accompanying cues to recognize syllable identity.

H2. Listeners also rely on cues other than F0 to identify a syllable as *yin* vs. *yang*.

If H1 is correct, listeners' identification responses (in terms of "accuracy" and response time), as well as their naturalness ratings should be similar for congruent and incongruent stimuli. If H2 is correct, listeners' identification should be less accurate and slower, and their naturalness ratings lower for incongruent than congruent stimuli. (We arbitrarily define as "accurate" the responses solely determined by the F0 contour.)

5.1.1 Method

5.1.1.1 Speech materials and design

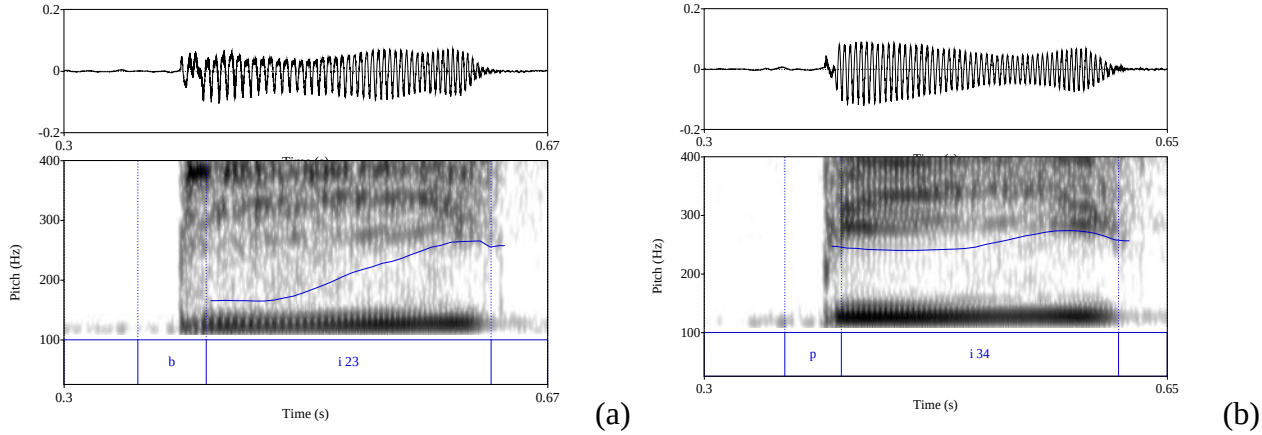
Thirty-six natural CV monosyllabic words were recorded in the carrier sentence /_ gəʔ əʔ zɻ̩ ŋo nin təʔ əʔ/ (__ 这个字/词我认得的。__ This word/character, I know it) by a 26-year-old female native speaker of Shanghai Chinese who was not aware of the purpose of the experiment. The onset of the target syllable (C) was a labial or dental stop or fricative. V was a cardinal vowel among /i, ɛ, u/. Each target syllable carried one of the three unchecked tones, making a total of *4 onset-types* * *3 vowels* * *3 tones* = *36 syllables*, as listed in Table 32. Among the 36 syllables, 24 were *yin* syllables, 12 carrying T1 and 12 carrying T2, and the other 12 were *yang* syllables. Each syllable can be written with different Chinese characters. The Chinese characters for identification responses were selected based on the subjective frequencies (on a 1-5 scale) collected from 17 native speakers of Shanghai Chinese about three months

before the experiment. They rated the lexical frequency for 107 characters (mixed with 10 distracters). For each minimal tone triplet, we selected one character per tone from the pool of homophones so that the three characters had as close ratings as possible. This procedure was followed in all the experiments presented in this chapter. The meaning of each monosyllabic word can be found in Appendix 1.

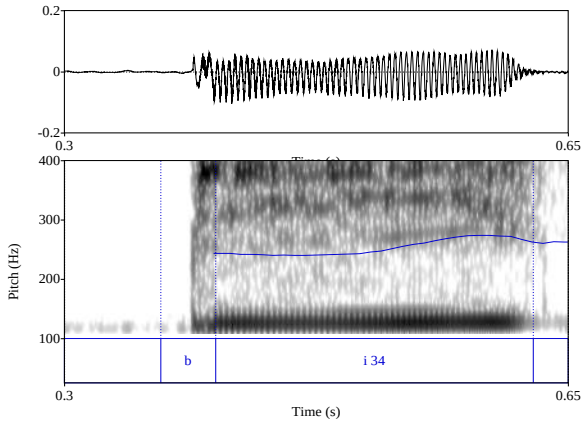
Table 32. List of materials in Experiment 3.

C	p, b			t, d			f, v			s, z		
V	i	ɛ	u	i	ɛ	u	i	ɛ	u	i [ʐ]	ɛ	u
T1	边	杯	波	低	堆	多	妃	翻	夫	丝	三	苏
T2	比	板	布	底	胆	堵	费	反	付	试	伞	所
T3	皮	办	步	提	台	涂	维	烦	扶	是	馋	坐

We constructed 84 syllables from the 36 naturally produced target syllables. All the natural target syllables were imposed a stylized F0 contour of their underlying tone, thus yielding 36 congruent syllables. A stylized T3 (*yang*) contour was imposed on the originally T1 and T2 (*yin*) syllables, producing 24 incongruent T3-contour syllables. A stylized T1 or T2 contour was imposed on the originally T3 syllables, producing 12 T1- and 12 T2-contour incongruent syllables. A total of 36 congruent and 48 incongruent syllables were thus created. Note that both congruent and incongruent syllables were constructed in altering a natural syllable with the same F0 contour manipulation.



(a) (b)



(c)

Figure 91. Waveforms and spectrograms of (a) the original syllable /bi/ (T3), (b) the original syllable /pi/ (T2), and (c) the incongruent syllable [pi] with the T2 contour of /pi/ imposed on /bi/.

The F0 stylization and transformation was made using the PSOLA technique (Valbret, Moulines & Tubach, 1992), as implemented in the Praat software (Boersma & Weenink, 1992-2015). The segmental durations of the original syllables were maintained. An example of the construction of an incongruent syllable is shown in Figure 91. The stimuli were stored in individual speech files, with homogeneous silent leading and trailing edges: 400 ms before vowel onset and 50 ms after vowel offset.

5.1.1.2 Participants

Fifteen young native speakers of Shanghai Chinese (8 male, 7 female) participated in both the identification test and the natural rating test. They were aged from 21 to 29 (mean age 25).

Ten elderly native speakers of Shanghai Chinese (6 male, 4 female) participated in the tests. Six of them (3 male, 3 female) participated in both tests. Two other male subjects participated only in the identification test, and the two other elderly subjects (1 male, 1 female) participated only in the naturalness-rating test. In all, eight elderly subjects thus participated in the identification test and eight in the naturalness-rating test. Those who participated in only one test found it difficulty to maintain their attention for a long time on tasks that require using a computer. The elderly subjects were aged from 64 to 79 and their mean age was 72.

All the participants were native speakers of Shanghai Chinese. The young participants were born and raised in Shanghai and had been living most of their life in the urban areas of Shanghai. This was also true for most of the elderly participants. One elderly participant was born in 宁波 Ningbo city in the neighboring Zhejiang province but moved to Shanghai during childhood, and his parents spoke Shanghai Chinese. Two other elderly participants had stayed for 28 years in 崇明 Chongming, a Shanghai county where the Chongming subgroup is spoken. One more elderly participant had stayed for more than 10 years in the province of 黑龙江 Heilongjiang, in Northeast China, where Northeast mandarin is spoken. Yet the three latter participants were born and raised in Shanghai. No participant reported any hearing or reading disorder. All were naïve as to the purpose of the experiment.

5.1.1.3 Procedure

Participants were tested individually in a quiete room, in front of a laptop; they were presented with the stimuli through professional quality headphones. Participants in both tests took the identification test before the naturalness-rating test, so that their tone identification was not affected by a prior evaluation of the stimuli. These participants took a break of about ten minutes between the two tests.

The identification test was conducted using the E-Prime 2.0 software. During the test phase, each trial consisted of the following events, as illustrated in Figure 92: at trial onset, a fixation cross was displayed at the center of the screen; 500 ms after trial onset, one of the auditory stimuli was presented; at stimulus offset, the fixation cross disappeared and was replaced with two Chinese characters on the left and right side of the screen, representing the two possible responses to the trial. The character whose reading matched the auditory stimulus at the tonal level was counted as the correct response and appeared equiprobably at either side of the screen. The correct response side was counterbalanced across the two or four repetitions of the same stimulus. Hence, the 12 T3 congruent syllables were presented four times and all the other stimuli were presented twice, so that an equal number of congruent and incongruent stimuli (96=2*24+4*12=2*48) were presented. There were thus 192 trials in total in the test phase. The stimuli were presented in a different random order for

each subject, with a break after 96 trials. The test phase was preceded by a training phase of five trials, in which participants received a feedback on their accuracy and response time. For each stimulus, they were asked to indicate the Chinese character whose reading matched best the stimulus by pressing one the two labeled keys on the left and right of the keyboard, as quickly and accurately as possible. The response time-out, counted from the display of the two response choices, was set to 2.5 seconds.

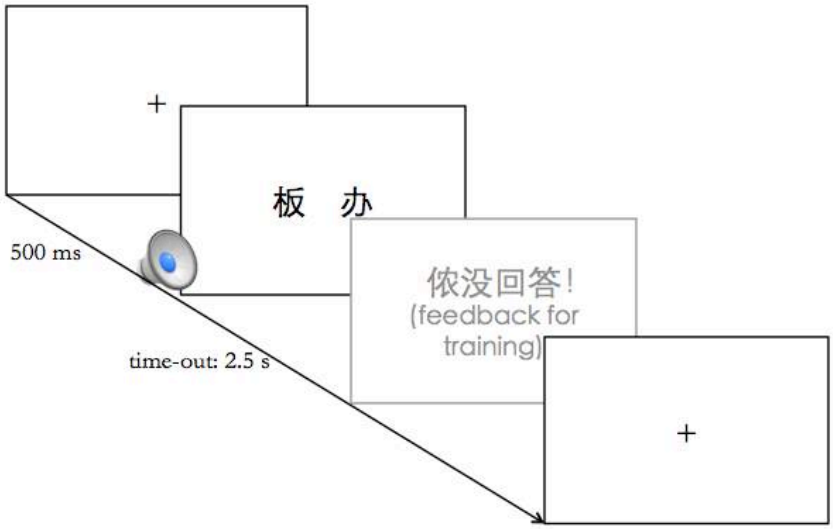


Figure 92. Time-line of the identification test in Experiment 3.

The naturalness-rating test was conducted using the Praat software (Boersma & Weenink, 1992-2015). On each trial, an auditory stimulus (from the set of 36 congruent and 48 incongruent stimuli) was presented and a Chinese character was displayed, during the entire trial, at the upper center of the screen, as shown in Figure 93. The participants were informed that what they heard was a pronunciation of the character they saw. The reading of the displayed character matched the auditory stimulus with respect to tone. The participants were asked to click on a number between 1 and 5 to rate the naturalness of the pronunciation they were presented with (from 1 for the least natural to 5 for the most natural). There was no response time-out: subjects simply clicked on the "ok" button to proceed with the next trial. The 84 stimuli (36 congruent syllables and 48 incongruent syllables) were presented each once, in a different random order for each participant. The test phase was preceded by a 5-trial training phase of five trials without feedback.

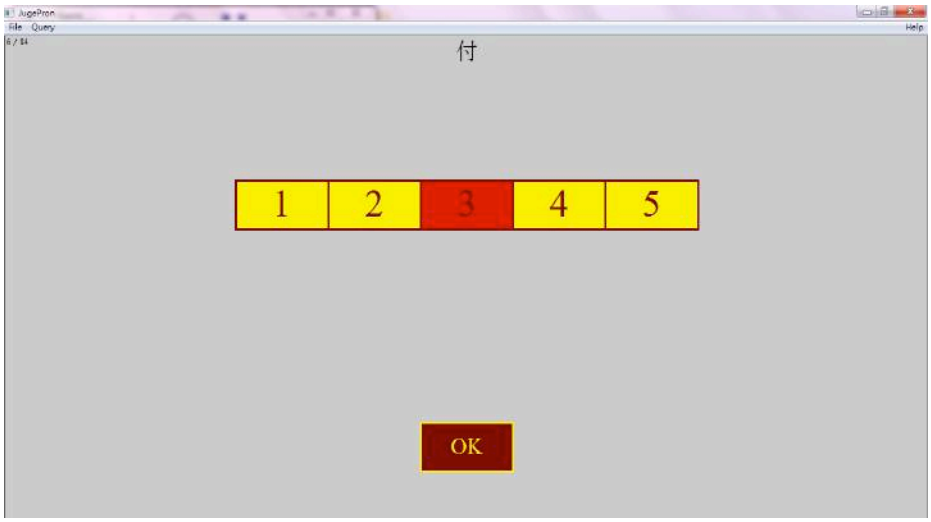


Figure 93. Set up of the naturalness-rating test in Experiment 3.

5.1.2 Results

5.1.2.1 Identification test

Figure 94 shows the accuracy data for congruent versus incongruent syllables as a function of the imposed tone register (*yin* vs. *yang*), and of the syllable onset type (according to the place and manner of articulation). For each category, accuracy was computed as the percentage of correct responses out of all the given responses (that is, missing responses were ignored) Overall, there was a slight advantage of congruent over incongruent syllables (96.1% vs. 89.4%). But under closer scrutiny, the advantage was substantial only for fricative onset syllables (96.7% vs. 72.3%), not for the other onsets (95.1% vs. 94.4%). We return to this point in the Discussion.

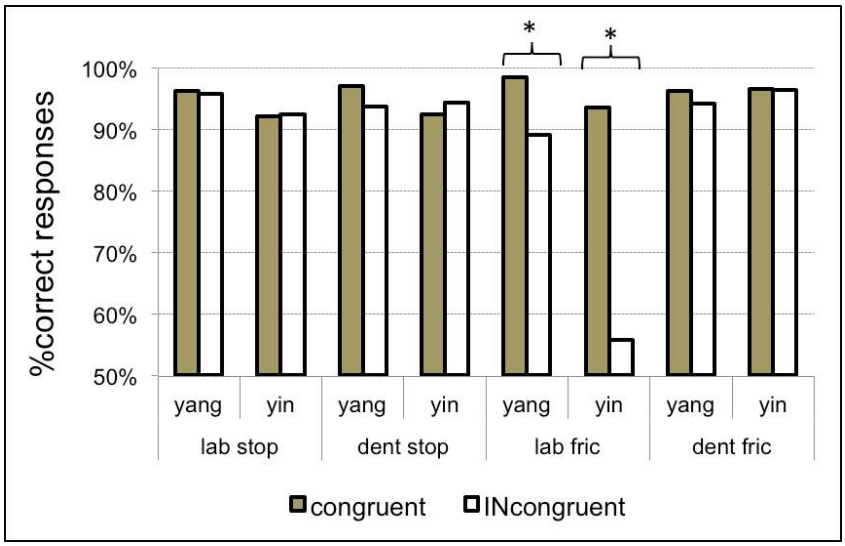


Figure 94. Accuracy data (% correct) for congruent vs. incongruent syllables, according to imposed tone register (*yin* vs. *yang*) and onset type. * for statistically significant difference between correct and incorrect.

To substantiate these observations, a series of generalized linear models (GLM) were fit to the binomial correct/incorrect response data, using the *lme4* package (Bates, Maechler & Dai, 2008) in the R software (R Core team, 2014). Both *Subject* and *Vowel* were random factors. In the full model, *Congruence* (congruent vs. incongruent), *Place* (labial vs. dental), *Manner* (stop vs. fricative), *Vowel* (/i, ɛ, u/), and *Imposed tone* (*yin* vs. *yang*) were fixed factors. The random factor structure included by-subject intercept. Likelihood ratio tests were performed to compare the full model, which included all the fixed effects, to models with one fixed effect missing, using the *anova* function. Two other models including *Congruence* × *Place* or *Congruence* × *Manner* interaction effects were also compared with the model without interaction effects.

Congruence was significant, $\chi^2(1)=16.1$, $p<.005$. The *Congruence* × *Place* interaction was significant, $\chi^2(1)=21.2$, $p<.0001$. The *Congruence* × *Manner* interaction was also significant, $\chi^2(1)=20.0$, $p<.0001$. These two interaction effects were due to the labial fricative onsets which differed from the other onset-types. We thus ran the same models separately on the data subset restricted to the labial fricative onsets and the data subset restricted to the other onsets. For these other onsets, *Congruence* was no longer significant, $\chi^2(1)=1.4$, $p=.23$. In contrast, for the labial fricative onsets, *Congruence* was significant, $\chi^2(1)=160.6$, $p<.0001$, reflecting significantly higher

response accuracy for congruent than incongruent labial fricative onset syllables (96.7%>72.3%). The detailed R outputs including the other predictors are shown in Appendix 3.

The accuracy data thus suggested that the congruence between tonal and non-tonal characteristics was especially important in the identification of labial fricative onset syllables. Now we turn to the response time data. Figure 95 shows the average response time (RT) data (for correct responses) for congruent and incongruent syllables, separately for young and elderly speakers, according to onset manner.

Table 33 shows the detailed RT data according to Onset type and Imposed contour category.

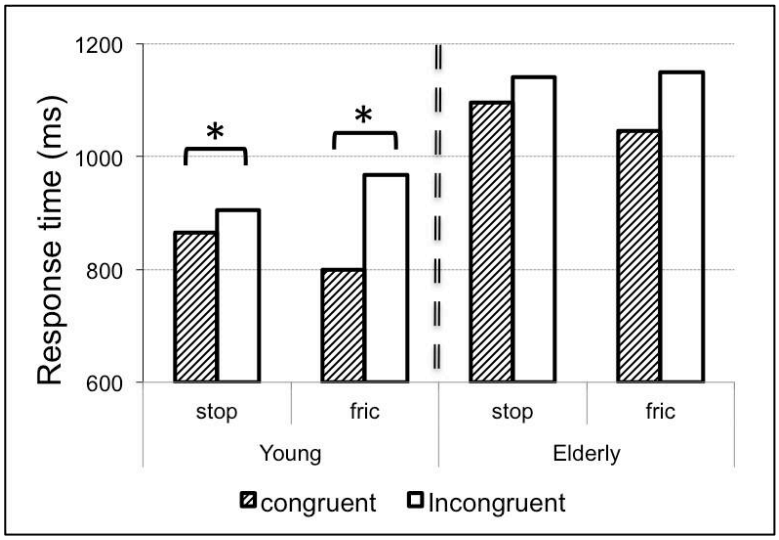


Figure 95. Correct response time data (ms) for congruent vs. incongruent syllables, according to onset manner (stop vs. fricative) and age group. Significance level: * for $p<.05$. Although this figure shows the labial fricative data, these data were excluded from statistical analyses due to the small number of observations compared to the other syllables: there were much less correct responses for labial fricative than other items.

Table 33. Correct response time data (ms) for congruent vs. incongruent syllables, according to onset type, imposed tone, and age group.

Onset Manner		**Stop**				**Fricative**				
Onset Place		**Labial**		**Dental**		**Labial**		**Dental**		**Mean (ms)**
Imposed tone		***Yang***	***Yin***	***Yang***	***Yin***	***Yang***	***Yin***	***Yang***	***Yin***	
Young	congruent	888	833	896	846	835	900	692	772	**833**
	incongruent	977	880	924	837	1006	1322	772	767	**936**
Elderly	congruent	1098	1110	1072	1105	1076	1201	886	1022	**1071**
	incongruent	1170	1161	1139	1090	1136	1545	961	952	**1144**

RTs were generally longer for elderly than young speakers. For all onset types, incongruent syllables required numerically longer RT than congruent syllables. To substantiate these observations, a by-subject ANOVA was run on the data. *RT* (of correct responses) was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* (young vs. elderly) were between-subject factors. *Congruence* (congruent vs. incongruent), *Onset type* (labial stop, dental stop, dental fricative), *Vowel* (ɛ, i, u) and *Imposed tone* (*yin* vs. *yang*) were within-subject factors. A large amount of RT data were missing for labial fricative onset syllables. For this reason, these data were excluded from the statistical analyses because the number of observations was very unbalanced compared to the other onset-types. In any case, the accuracy data for these syllables clearly showed that listeners did not solely rely on F0 contours to identify them (see the following discussion).

Congruence was significant, $F(1,19)=13.3$, $p<.005$, reflecting longer RTs for incongruent than congruent syllables (969>935 ms). The *Congruence* × *Onset type* interaction was significant, $F(2,38)=3.3$, $p<.05$, reflecting a larger Congruent vs. Incongruent differential for labial stop (Δ=65 ms), $F(1,19)=14.6$, $p<.005$, than dental fricative (Δ=20 ms), $F(1,19)=6.4$, $p<.05$, or dental stop (Δ=18 ms), $F(1,19)<1$. The *Congruence* × *Vowel* interaction was not significant, $F(2,38)<1$. The *Congruence* × *Imposed tone* interaction was significant, $F(1,19)=6.0$, $p<.05$, reflecting a significant *Congruence* effect only for imposed *yang* contours (Δ=69 ms), $F(1,19)=17.0$, $p<.0001$, but not for imposed *yin* contours (Δ=0 ms), $F(1,19)<1$.

Gender was not significant, $F(1,19)<1$, but *Age* had a significant effect on *RT* $F(1,19)=17.3$, $p=.0005$: RTs were longer overall for elderly than young speakers (1064>840 ms). Finally, *Congruence* did not interact with *Gender* or *Age*, Fs$(1,19)<1$.

5.1.2.2 Naturalness rating test

Figure 96 shows the average ratings for congruent and incongruent syllables, pooled across listener groups, vowels and imposed tone categories. Overall, congruent syllables obtained higher scores than incongruent syllables, and the difference is especially great for labial fricative onset syllables. Table 34 shows the detailed rating data according to *Onset type* and *Imposed tone*.

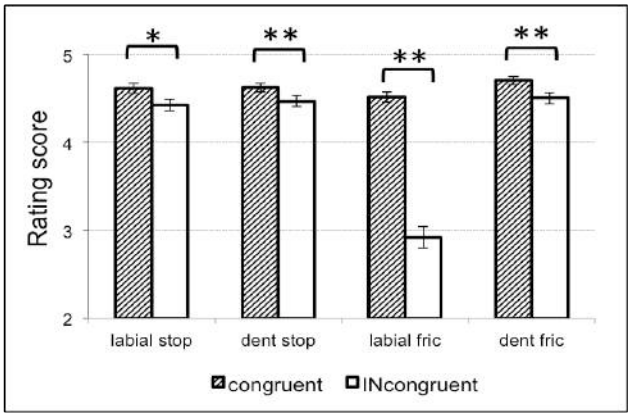


Figure 96. Rating scores for congruent vs. incongruent syllables, according to onset type. Significance level: * for $p<.05$.

Table 34. Naturalness ratings for congruent vs. incongruent syllables, according to onset type, imposed tone and age group.

Onset Manner		**Stop**				**Fricative**				**Mean**
Onset Place		**Labial**		**Dental**		**Labial**		**Dental**		
Imposed tone		***Yang***	***Yin***	***Yang***	***Yin***	***Yang***	***Yin***	***Yang***	***Yin***	
Young	congruent	4.53	4.63	4.58	4.70	4.53	4.66	4.82	4.66	**4.61**
	incongruent	4.18	4.46	4.38	4.47	3.28	1.66	4.33	4.60	**3.92**
Elderly	congruent	4.79	4.56	4.75	4.46	4.54	4.58	4.63	4.67	**4.62**
	incongruent	4.71	4.56	4.65	4.46	4.15	3.42	4.50	4.67	**4.39**

To substantiate these observations, a by-subject ANOVA was run on the data. *Rating* was the dependent variable, and *Subject* was the random factor. *Gender* (female vs. male) and *Age* (young vs. elderly) were between-subject factors. *Congruence* (congruent vs. incongruent), *Place* of articulation (labial vs. dental), *Manner* of articulation (stop vs. fricative), *Vowel* (ɛ, i, u) and *Imposed tone* (*yin* vs. *yang*) were within-subject factors.

Congruence was highly significant, showing higher score for congruent than incongruent syllables (4.62>4.08), $F(1,19)=102.5$, $p<.0001$. The *Congruence* × *Manner* and *Congruence* × *Place* interactions were both significant, $F(1,19)>62.8$, $p<.0001$, and the *Congruence* × *Manner* × *Place* interaction as well, $F(1,19)=56.8$, $p<.0001$, reflecting higher Congruent vs. Incongruent differentials for labial fricative onsets ($\Delta=1.59$, $F(1,19)=111.5$, $p<.0001$), than the other three onset types (labial stop: $\Delta=0.19$, dental stop: $\Delta=0.16$, dental fricative: $\Delta=0.20$, ps<.05). The *Congruence* × *Vowel* interaction was also significant, $F(2,38)=4.2$, $p<.05$, reflecting higher Congruent vs. Incongruent differentials for the vowel /i/ and /u/ (/i/: $\Delta=0.65$, $F(1,19)=74.5$, $p<.0001$; /u/: $\Delta=0.51$, $F(1,19)=89.5$, $p<.0001$), than for the vowel /ɛ/ ($\Delta=0.45$), $F(1,19)=40.0$, $p<.0001$. The *Congruence* × *Imposed tone* interaction was significant as well, $F(1,19)=4.6$, $p<.05$, but only for labial fricative onset syllables. For these syllables, the Congruent vs. Incongruent differentials were higher for imposed *yin* contour ($\Delta=2.23$), $F(1,19)=106.7$, $p<.0001$, than for imposed *yang* contour ($\Delta=0.96$), $F(1,19)=43.0$, $p<.0001$.

Neither *Gender* nor *Age* was significant (*Gender*: $F(1,19)=1.2$, $p=.28$; *Age*: $F(1,19)=1.7$, $p=.21$). Finally, *Congruence* did not interact with *Gender*, $F(1,19)=2.0$, $p=.18$, but interacted with *Age*, $F(1,19)=17.2$, $p=.0005$, reflecting higher Congruent vs. Incongruent differentials for young speakers ($\Delta=0.69$, $F(1,13)=102.6$, $p<.0001$) than elderly speakers ($\Delta=0.23$, $F(1,6)=8.5$, $p<.05$).

5.1.3 Discussion

Experiment 3 showed that for all speaker groups and all *yin–yang* contrasts, except for labial fricative contrasts, correct response rate was not higher for Congruent than Incongruent syllables, suggesting that F0 contour is the dominant

cue for tone identification, distinguishing between *yin* and *yang* tones. However, the congruence between F0 and non-F0 characteristics speeded up tone identification and enhanced naturalness ratings, suggesting that listeners are sensitive to other acoustic cues than F0, even though such sensitivity does not show up in the accuracy data.

The exceptional results revealed by labial fricative onset syllables are due to the fact that *yang* labial fricative onsets are realized as voiced by the young female speaker who produced the stimuli. According to §4.1.6.2 there is a general trend for young speakers to produce phonetically voiced *yang* fricatives and especially labial ones, that is [v] rather than [f] for underlying /v/. The low accuracy rate in identifying *yin* syllables with voiced [v] onset and *yang* syllables with voiceless [f] onset suggests that voicing, at least in labial fricatives, affects tone identification. As Figure 94 shows, when a *yin* tone (T1 or T2) was imposed on a *yang* syllable with a labial fricative onset (/v/ realized as [v]), the accuracy was the lowest (~56%). This presumably precluded its identification as a *yin* syllable (the expected "correct" response in this case). That is, phonetic voicing overweighted tone contour to correctly identify a labial fricative onset syllable as *yin*. For syllables with a *yin* labial fricative onset (/f/ realized as [f]), the accuracy was significantly lower in the incongruent (*yang* tone imposed) than the congruent (*yin* tone imposed) condition (~89%<98%). The congruent vs. incongruent differential was much larger in the case of syllables with a *yang* /v/ onset (realized as [v]) (~56%<93%). In other words, a low tone partly overweighs the lack of phonetic voicing, whereas a high tone does not overweigh the presence of phonetic voicing, at least in the case of labial fricative onset syllables.

What about the other acoustic cues? The next two experiments aimed at investigating the role of the consonant-vowel duration pattern, and that of voice quality in *yin* vs. *yang* syllable (or tone) identification.

5.2 Experiment 4: the role of segmental duration in tone perception[20]

This experiment was intended to test whether the C/V duration ratio affects the identification of a syllable's tone register (*yin* vs. *yang*). This possibility is suggested by our production data, which clearly show that (i) syllable-initial obstruents are shorter for yang than yin syllables and (ii) both checked and unchecked syllable rimes are longer for yang than yin syllables. We predict that high C/V duration ratios would bias the *yin* vs. *yang* identification of Shanghai Chinese syllables toward the *yin* side, and vice versa for low C/V ratios.

5.2.1 Method

We conducted a two-fold forced-choice identification test on several tone continua between a *yin* tone (T2 or T4) and a *yang* tone (T3 or T5). The endpoint stimuli were derived from monosyllabic word minimal pairs of a T2-T3 or T4-T5 contrast. The C/V duration ratio was manipulated to produce high vs. low C/V duration ratio tone continua. We tested whether this manipulation biased the identification of the stimuli in the continua toward the *yin* or *yang* tone category, respectively. The original, naturally produced monosyllabic word minimal pairs differed in tone register (*yin* vs. *yang*) but had identical segments and a similar shape tone contour: rising for T2-T3, flat for T4-T5. We only used monosyllabic words with a fricative onset.

5.2.1.1 Speech materials

The original materials were composed of eight CV(ʔ) monosyllabic words, making four tone contrasts, shown in Table 35. The rime was either /ɛ/ or /aʔ/. The onset was a

[20] Part of the results in this section have been reported at the 14th Annual Conference of Interspeech in 2013 in Lyon and published in the proceedings of the conference.

labial or dental fricative /f, v, s, z/. The tones were T2 or T3 for unchecked syllables and T4 or T5 for checked syllables. The syllables were produced in isolation by the author of this dissertation, a female native speaker of Shanghai Chinese aged 25 at the time of the recording. She deliberately produced all the syllable onsets without phonetic voicing. As can be seen in Table 35, there were two T2-T3 contrasts (based on [fɛ] and [sɛ]) and two T4-T5 contrasts (based on [faʔ] and [saʔ]). The meaning of each monosyllabic word can be found in Appendix 1.

Table 35. Materials in Experiment 4.

tone register	unchecked	checked
yin (T2 or T4)	[fɛ] 反, [sɛ] 伞 (T2)	[faʔ] 发, [saʔ] 杀 (T4)
yang (T3 or T5)	[fɛ] 烦, [sɛ] 馋 (T3)	[faʔ] 罚, [saʔ] 十 (T5)

These naturally produced syllables were time-scaled, yielding derived stimuli with two extreme duration pattern: a low C/V duration ratio and a high C/V duration ratio. The stimuli were constructed using the RELP-PSOLA (Residual-exicted linear prediction PSOLA) method of speech signal time-scaling (Macchi, Altom, Kahn, Singhal & Spiegel, 1993). We call the two duration patterns LS (long onset and short rime, hence high C/V ratio) and SL (short onset and long rime, hence low C/V ratio) patterns. For the LS pattern, the original onset duration was increased by 30% while the original rime duration was decreased by 30%. In the SL pattern, the opposite changes were applied. The duration of the glottalized part of the rime (mostly in checked syllables) was not altered for sake of naturalness. A Matlab program implementing the RELP-PSOLA method was kindly provided by Thibaut Fux, who used it for his PhD research (Fux, 2012). The waveforms in Figure 97 show the original version of a [fɛ] syllable (center panel) and the two modified LS (top) and SL (bottom) versions of this syllable obtained with the RELP-PSOLA time scaling method. It should be noted that this method preserves the spectral information (including F0), just like the more classic PSOLA method but produces more naturally sounding transformed speech, especially in the case of time-scaled (expanded or contracted) aperiodic speech segments such as voiceless fricatives (Fux, 2012).

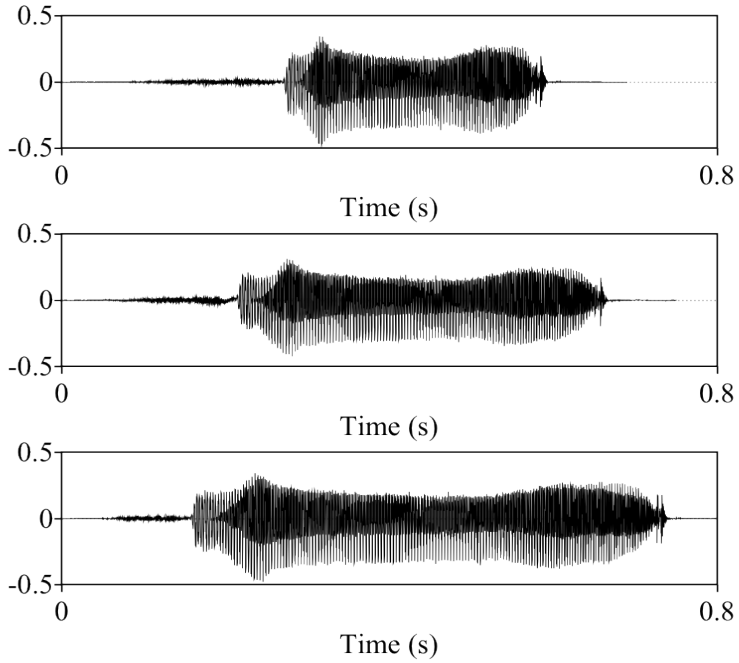


Figure 97. From top to bottom: LS, original, and SL versions of the [fɛ] syllable.

For each of the four *yin-yang* contrasts, four F0-equidistant eight-step T2-T3 or T4-T5 tone contour continua were constructed: two with the LS pattern, and two with the SL pattern. For each LS or SL pattern, one continuum was constructed from a *yin* syllable as the starting endpoint, and the other from its *yang* counterpart. The PSOLA method for pitch-scaling (Valbret et al., 1992), as implemented in the Praat software (Boersma & Weenink, 1992-2015) was used to impose the tone contours of the continua on the LS or SL time-scaled versions of the original stimuli. The endpoint contours were taken from the time-scaled versions of the original *yin* and *yang* syllables and the six intermediate contours were interpolated between these endpoints in equal F0 steps. For each continuum, all eight tone contours, including the endpoint contours, were stylized so that all the stimuli were derived from the original stimuli with a comparable speech signal transformation. Figure 98 shows the T2-T3 continuum of the /ɛ/ vowel in the SL duration pattern [fɛ] syllable. The total number of stimuli was 128 (4 contrasts * 4 continua * 8 steps).

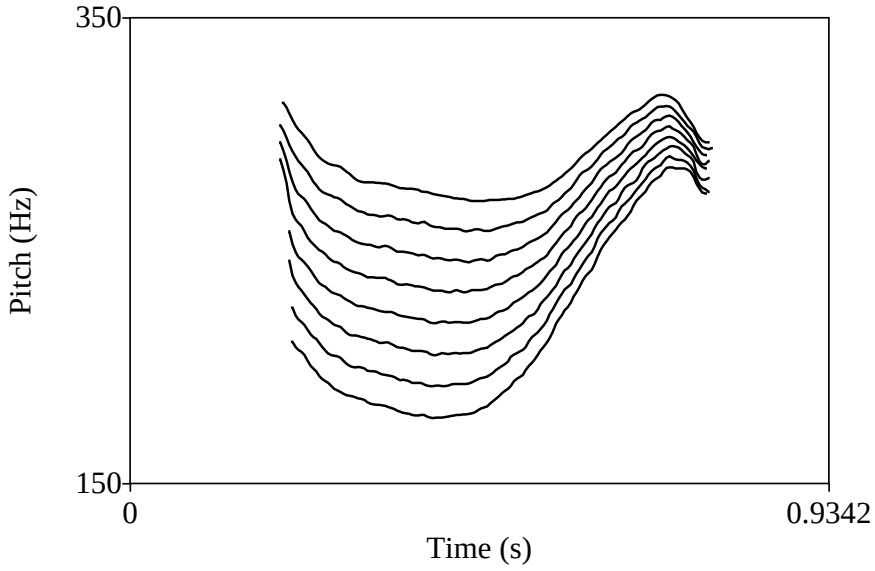


Figure 98. T2-T3 continuum of the [ɛ] vowel in the SL [fɛ] syllable.

5.2.1.2 Participants

Twenty-seven native speakers of Shanghai Chinese (9 male and 18 female), aged from 18 to 34 (mean 22.5), participated in Experiment 4.

All were born in Shanghai. All of them have spent most of their lifetime in Shanghai. All learned Standard Chinese before the age of 8. All learned English, thirteen of them learned or were learning French, two of them spoke Japanese and one of them German, at intermediate level.

As other Shanghai speakers of the young generation, they made, in general, a less frequent usage of Shanghai than Standard Chinese, with an average self-evaluation score of 3.4 (on a 1-5 scale) for Shanghai and of 4.6 for Standard Chinese. Their self-evaluation of language competence is worth to note. For each participant, self-evaluation of linguistic competence was consistently lower or equal for Shanghai than Standard Chinese.

5.2.1.3 Procedure

The identification test was conducted using the E-Prime 2.0 software. Participants were tested individually in a quiet room, in front of a laptop. Stimuli were presented to the participants through professional quality headphones. The testing procedure was the same as used in the identification test of Experiment 3: at trial onset, a fixation cross was displayed at the center of the screen; after 500 ms had elapsed, the auditory stimulus was presented; at the offset of this stimulus, the fixation cross disappeared and was replaced with two Chinese characters representing the two possible responses (a *yin–yang* minimal pair). Each stimulus was presented twice. The *yin* or *yang* response side was counterbalanced across the two repetitions of the same stimulus. Response times were measured from stimulus onset.

The 256 trials (128 stimuli * 2 repetitions) were presented to participants in a pseudo-random order with the constraint that two stimuli sharing their onset and their rime could not appear in succession. The test phase was preceded by a training phase of six trials in which a monosyllabic word carrying a canonical, unambiguous tone contour was presented, Participants received feedback during the training but not during the test phase.

5.2.2 Results

The data from one male speaker aged 19 was not retained because his miss rate during the test phase was too high: 3.5% against an average 0.29% for the other participants.

Figure 99a-b shows the identification curves averaged across the LS vs. SL pattern continua, separately for the unchecked and checked syllables. These curves show the *yin* response rate along the eight step *yin–yang* continua (with the *yin* endpoint on the left). For the unchecked /ɛ/-rime syllables, a categorical boundary shift can be seen towards the *yang* endpoint (T3) for the LS relative to the SL pattern. That is, there were more *yin* responses for the LS than SL pattern, as predicted. For the

checked /a?/-rime syllables, however, no categorical boundary shift is observed between LS and SL.

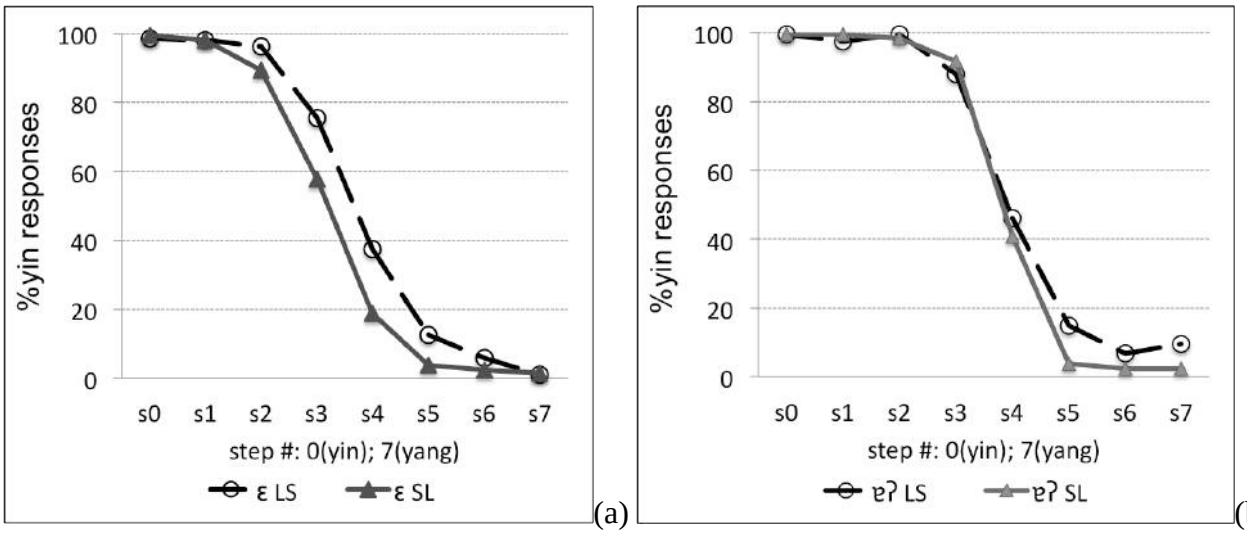
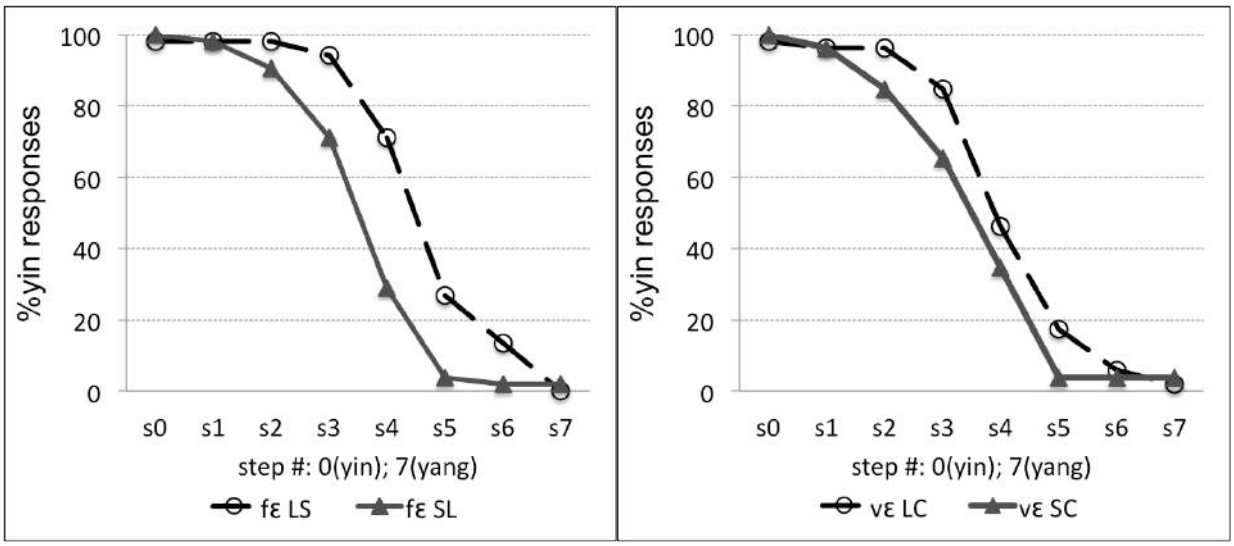


Figure 99. *yin* response rate along the *yin-yang* continua, according to the LS vs. SL duration pattern for (a) /ɛ/-rime and (b) /aʔ/-rime syllables.

Step 0 corresponds to the *yin* (T2 or T4) endpoint.



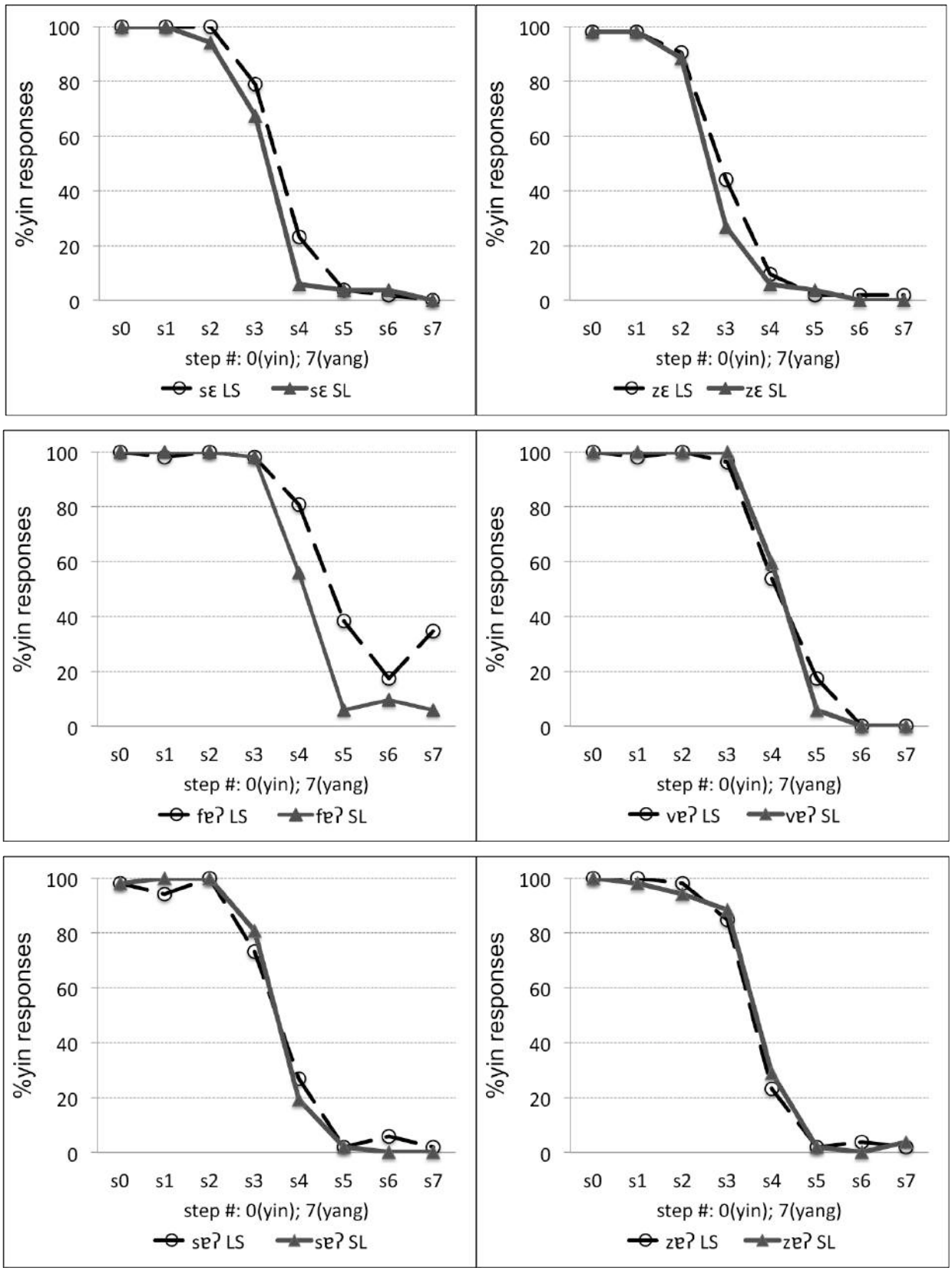


Figure 100. *yin* response rate data of Figure 99 detailed by *yin-yang* contrast and by the original syllable type on which the continuum tone contours were imposed (e.g., /sɛ/ for the continua based on *yin* /sɛ/; /zɛ/ for the continua based on *yang* /zɛ/).

Figure 100 shows the same yin response rate data as in Figure 99 but detailed by *yin-yang* contrast and by the original syllable type on which the continuum tone contours were imposed. For example, the /sɛ/ label stands for the continua based on the *yin* /sɛ/ syllable, whereas the /zɛ/ label stands for the continua based on the *yang* /zɛ/ syllable. Note that such voiced vs. voiceless labels (e.g., /zɛ/ vs. /sɛ/) are only phonological, since all the onsets were deliberately produced as *phonetically* voiceless. The figures in the left column correspond to the continua in which the tone contours were imposed on an originally *yin* syllable. Those in the right column correspond to the continua in which the tone contours were imposed on an originally *yang* syllable. Categorical boundary shifts between the LS and SL patterns can be seen more or less clearly for most unchecked syllables, whereas, among checked syllables, such a shift can only be observed for the /faʔ/ syllable.

5.2.2.1 Identification category boundary location

Category boundary locations (for the perceived *yin* and *yang* categories) can be defined as the point in the continuum at which the yin response rate reaches 0.5 (i.e., 50%). These locations, expressed in terms of step (or stimulus) number, are also called intercepts. We estimated these intercepts for each continuum and each listener, using probit analyses fitting short ogive Gaussians to the raw data, thus yielding intercepts and slopes (Best & Strange, 1992; Hallé, Best & Levitt, 1999). As in, for example, Best and Strange (1992), the raw data to which the short ogives were fitted were restricted to three data points encompassing the 50% identification crossover., These three data points generally corresponded to consecutive stimulus numbers or steps. Table 36 shows the intercept values (expressed in step numbers, from 0 to 7), averaged across participants, and according to duration pattern. The category boundary location was generally closer to the *yang* endpoint for the LS than SL pattern (3.72 vs. 3.15 for unchecked syllables, and 3.92 vs. 3.84 for checked syllables), showing that the LS duration pattern induces a bias toward *yin* responses relative to the SL pattern. This difference was larger and more systematic for the /ɛ/ than the /aʔ/ rime.

Table 36. Intercept data (stimulus number) according to the duration pattern, and the syllable-type (onset and rime) on which continua are based. Significance levels for the *t*-tests comparisons between LS and SL continua, for /ɛ/-rime (shaded cells) and /aʔ/-rime (white cells) syllable-types: * for $p<.05$, ** for $p<.01$.

Onset	LS		SL	
	/ɛ/	/aʔ/	/ɛ/	/aʔ/
/f/	4.50	4.47	** 3.48	4.06
/s/	3.48	3.56	3.19	3.52
/v/	3.90	4.06	** 3.26	4.12
/z/	3.00	3.60	* 2.67	3.67
Mean	**3.72**	**3.92**	**** 3.15**	**3.84**

We ran a by-subject ANOVA on these category boundary data. *Boundary step* number was the dependent variable and *Subject* was the random factor. *Duration pattern* (LS vs. SL), *Place* of articulation (labial vs. dental), *Original tone* (*yin* vs. *yang* syllable-type on which the continuum is based) and *Rime* (/ɛ/ vs. /aʔ/) were within-subject factors.

Duration pattern had significant effect on boundary location: the boundary was overall closer to the *yang* endpoint for the LS than SL pattern (3.82>3.50), $F(1,25)=18.2$, $p<.0005$. *Place* of articulation was highly significant: the boundary was closer to the *yang* endpoint for labial than dental onsets overall (3.98>3.34), $F(1,25)=90.8$, $p<.0001$. The *Duration pattern* × *Rime* interaction was significant, $F(1,25)=16.5$, $p<.0005$: the LS vs. SL difference was significant only for the /ɛ/ rime ($\Delta=0.57$), $F(1,25)=21.2$, $p=.0001$, and not for the /aʔ/ rime ($\Delta=0.08$), $F(1,25)<1$. The *Duration pattern* × *Place* interaction was also significant, $F(1,25)=6.1$, $p<.05$: the LS vs. SL difference was significant (across rimes) only for the labial onset ($\Delta=0.50$), $F(1,25)=20.2$, $p=.0001$), and not for the dental onset ($\Delta=0.14$), $F(1,25)=2.9$, $p=.10$. But the difference was significant for dental onset and /ɛ/ rime syllable-types ($\Delta=0.31$), $F(1,25)=5.7$, $p<.05$. The *Duration pattern* × *Original tone* interaction was also significant, $F(1,25)=4.7$, $p<.05$, reflecting larger LS vs. SL differentials for the

continua based on originally *yin* than *yang* syllables (*yin*: Δ=0.44, $F(1,25)=28.1$, $p<.0001$; *yang*: Δ=0.21, $F(1,25)=5.0$, $p<.05$).

5.2.2.2 Overall percentage of *yin* responses

Overall, 55.4% of the syllables were identified as *yin* for the LS pattern. This percentage went down to 50.6% for the SL pattern. The *yin* response rate thus was higher for the LS than the SL pattern for both unchecked (53.1 vs. 46.4%) and checked (57.8 vs. 54.9%) syllables, indicating again that the LS pattern induces a bias toward *yin* responses.

Figure 101 shows the overall *yin* response rate (averaged across all stimulus steps) according to duration pattern and original syllable-type (onset and rime). Recall that the labels used for the onsets are phonological (indexing *yin-yang* syllable-type) and all the onsets were produced without phonetic voicing.

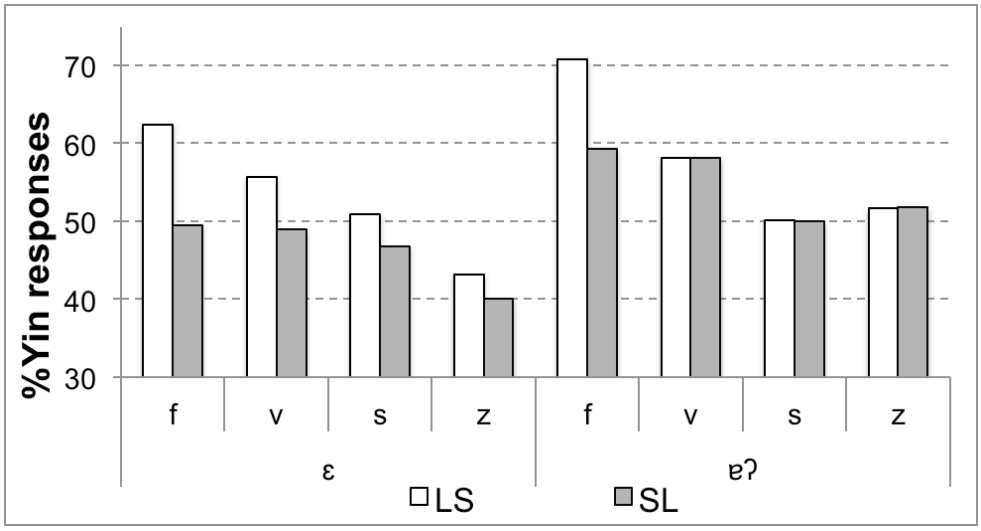


Figure 101. *yin* response rate across all stimulus steps according to duration pattern and syllable-type (onset and rime).

To test the significance of the above-mentioned factors, a series of generalized linear models (GLM) were fit to the binomial *yin*/*yang* response data, using the *lme4* package (Bates, Maechler & Dai, 2008) in the R software (R Core team, 2014). *Subject* was the sole random factor; in the full model, *Duration* (LS vs. SL), *Onset* (labial vs. dental), *Rime* (/ɛ/ vs. /aʔ/), *Original tone* (*yin* vs. *yang*: see §0.2.2.1) and *Step* (0 to 7) were the fixed factors. The inclusion of *Gender* (female, male) did not improve the

goodness of fit of the model, hence it was removed and taken as non-significant. The random factor structure included by-participant random intercept. Likelihood ratio tests were performed to compare the full model, which included all the fixed factors, to models with one fixed factor missing, using the *anova* function. Three other models including the *Duration* × *Onset*, *Duration* × *Rime,* or *Duration* × *Original tone* interactions were also compared with the model without interactions.

Duration was significant, $\chi^2(1)=15.5$, $p<.0001$. The *Duration* × *Onset* interaction was significant, $\chi^2(1)=59.3$, $p<.0001$, reflecting larger LS vs. SL differentials for labial ($\Delta=7.8\%$) than dental onsets ($\Delta=1.8\%$). The *Duration* × *Rime* interaction was significant, $\chi^2(1)=10.8$, $p<.005$, reflecting larger LS vs. SL differentials for /ɛ/ ($\Delta=6.7\%$) than /aʔ/ rime ($\Delta=2.9\%$). The *Duration* × *Original tone* interaction was significant as well, $\chi^2(1)=15.0$, $p<.0005$, reflecting larger LS vs. SL differentials for originally *yin* ($\Delta=7.2\%$) than *yang* syllables ($\Delta=2.4\%$). The detailed R outputs including the other predictors are shown in Appendix 3.

5.2.2.3 Response time

Figure 102 shows the identification response time (RT) data for /ɛ/ and /aʔ/ rime syllables, averaged across subjects and onset-types, according to duration pattern and continuum step. Because *yin* responses are clearly dominant in the first half of the continua (steps 0–3), while *yang* responses are infrequent, as can be seen in Figures 9-10, we only analyzed the RTs data for *yin* responses in this region of the continua. These RTs are displayed in Figure 102 from step 0 to 3. Symmetrically, we only analyzed the RTs for *yang* responses in the second half of the continua. These RTs are displayed in Figure 102 from step 4 to 7. For /ɛ/ rime syllables, the *yin* RTs at steps 2 and 3 (close to the ambiguous region of the continuum) were longer for the SL than LS pattern, suggesting that the LS pattern facilitates *yin* responses, whereas the *yang* RTs at steps 4 and 5 (again close to the ambiguous region) were longer for the LS than SL pattern, suggesting that the SL pattern facilitates *yang* responses. This pattern of facilitation was not observed for /aʔ/ rime syllables.

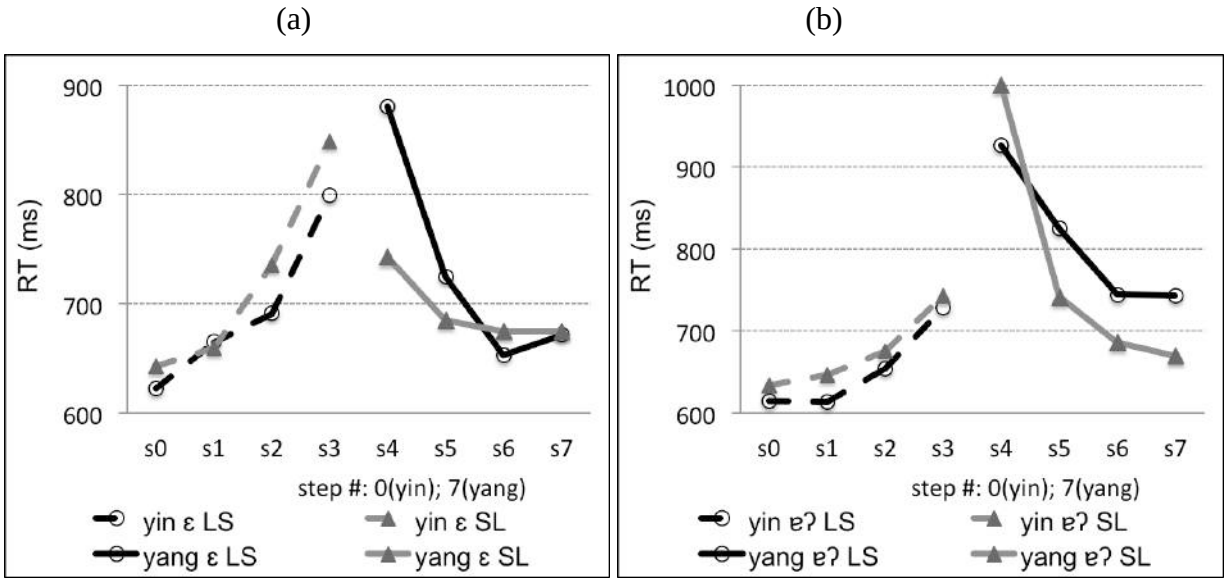


Figure 102. Average RTs for *yin* responses in the *yin*-dominant region (steps 0–3: dashed lines) and for *yang* responses in the *yang*-dominant region (steps 4–7: solid lines) for the LS (black) vs. SL (gray) duration patterns, for (a) /ɛ/ rime and (b) /aʔ/ rime syllables.

We examined the impact of the congruence between duration pattern and tone register on RTs. We define Congruent stimuli as those stimuli whose duration pattern matches the dominant response; Symmetrically, we define Incongruent stimuli as those stimuli whose duration pattern mismatches the dominant response. By this definition, the LS pattern stimuli are congruent (with *yin* responses) at steps 0-3 and incongruent (with *yang* responses) at steps 4-7, and vice versa for the SL pattern stimuli.

Overall, RTs were shorter for Congruent than Incongruent stimuli (697<728 ms), for both unchecked (693<719 ms) and checked syllables (702<736 ms).

In order to substantiate these observations, a by-subject ANOVA was run on the RT data. *RT* was the dependent variable and *Subject* was the random factor. *Congruence* (congruent vs. incongruent), *Dominant steps* (*yin*-dominant steps 0-3 vs. *yang*-dominant steps 4-7), *Place* of articulation (labial vs. dental), and *Rime* (/ɛ/ vs. /aʔ/) were within-subject factors.

Congruence was globally significant, with shorter RTs for *Congruent* than *Incongruent* stimuli (697<728 ms), $F(1,25)=15.4$, $p<.001$. *Dominant steps* also had a significant effect overall, with shorter RTs for *yin*-dominant than *yang*-dominant steps

(684<740 ms), $F(1,25)=14.0$, $p<.005$. *Congruence* did not interact with any of the other within-subject factors, Fs$(1,25)<1$.

5.2.3 Discussion

Experiment 4 showed that LS-pattern syllables (high C/V duration ratio) are more likely than SL-pattern syllables (low C/V duration ratio) to be perceived as *yin* syllables by Shanghai young listeners, both males and females. This conclusion is supported by the shift toward *yin* responses for high C/V ratio relative to low C/V ratio syllables in the *yin–yang* categorization of T2-T3 continua, as shown both by the shift of the intercepts toward the *yang* endpoint and by the overall increase in *yin* response rate, as well as the RT data,

The duration pattern effect was greater for unchecked than for checked syllables. This might be due to the fact that the checked syllable rimes were too short and the 30% duration modification on these rimes were less perceptible than for the unchecked syllable rimes. An alternative explanation is that the presence of a rime-final glottal stop reduced the perceptual effect of the vowel duration manipulation.

The effect of the duration pattern was larger for syllables with labial fricative onsets than those with dental fricative onsets. Although in production, in monosyllables, the *yin–yang* duration difference is similar for the labial and dental fricative series (§4.1.6.4), the difference is larger for labial than dental fricatives in the S1 context, probably leading to a similar trend in perception.

In spite of the different performances according to the rime and the onset type, the general trend we observed clearly is that high C/V duration ratios bias tone identification toward the *yin* category. We know that voiced obstruents have shorter durations than their voiceless counterparts (e.g., Umada, 1977). The role of duration in the perception of voicing also important. Short intervocalic closure duration biases listeners' perception toward the voiced category (Lisker, 1957). Cho & Giavazzi (2009) found that, in English, the duration of frication and that of the preceding vowel are the most salient perceptual cues for the perception of voicing in intervocalic position. In the same vein, Shinji Maeda (personal communication) once commented that when

he tried to synthesize voiced fricatives, he simply shortened their duration so that short fricatives would be heard as voiced. In Shanghai Chinese, *yang* obstruents are not consistently voiced, but the C/V duration pattern still plays a role, presumably a secondary role, in the perception of the phonological voicing which correlates with the *yin* vs. *yang* tone register distinction.

5.3 Experiment 5: the role of voice quality in tone perception[21]

This experiment was intended to test whether voice quality affects the identification of a syllable's tone register (*yin* vs. *yang*). We predict that modal phonation should facilitate the identification of *yin* tones, whereas breathy phonation should facilitate the identification of *yang* tones. According to our production data, the *yin–yang* phonation type difference, that is, the breathy voice accompanying *yang* tone syllables, tends to disappear in young speakers' productions (see Chapter 4). We therefore chose to use young listeners in order to test whether the loss of breathiness as a redundant feature precedes or follows the loss of its role in perception.

To our knowledge, the only study on the role of voice quality in tone perception in Shanghai Chinese was conducted by Ren Nianqi in his Ph.D dissertation (Ren, 1992). He found that breathy voice was a perceptual cue to *yang* tone identity. He tested listeners on their perception of an ambiguous disyllabic sequence as ending with a *yin* or *yang* second syllable, according to its F0 onset height and its degree of breathiness, as we explain below in more detail. We know that, in Shanghai Chinese disyllabic words, the realized tone contour of the second syllable S2 is determined by the underlying tone of the first syllable S1 by virtue of a tone sandhi rule (§2.1.3.2). That is, the tonal component of the original *yin* or *yang* identity of the S2 syllable, is, in principle, absent. Yet, after S1 in tone T2, the F0 onset height of S2 is still modulated by the voicing of its onset: the F0 onset of S2 is higher when its onset is phonologically and phonetically voiceless than voiced (§4.1.6.1.4). Ren asked whether these F0 onset height differences, together with voice quality differences, might help listeners to recover the presumably original tonal identity of S2. He used the potentially ambiguous sequence [ɦa.ta], which, he claimed, may correspond to either 鞋带 'shoelace' or 鞋大 'the shoe is big'. [ta] in the former and latter [ɦa.ta] originally bears the yin tone T2 and the yang tone T3, respectively, but the sandhi rule imposes a 22-33 contour on both [ɦa.ta] sequences, determined by the tone T3 of [ɦa]. Ren

[21] Part of the results of this section have been reported at the 27èmes Journées de Linguistique d'Asie Orientale in 2014 in Paris.

synthesized [ɦa.ta] stimuli, varying the [ta] syllable (whose onset phonetic voicing was ambiguous) along two dimensions. The first dimension was F0 onset height, which took three possible values: 110, 120, or 130 Hz. The F0 contour of [ta] ended at 135 Hz. The second dimension was the voice quality of [a] in [ta], which varied from breathy to modal on a voice quality continuum constructed by manipulation of the open quotient (OQ). Listeners were tested with these stimuli on a forced-choice identification task with 鞋带 'shoelace' /aɻ.taɬ/ and 鞋大 'the shoe is big' /aɻ.daɻ/ as the two possible responses. Both manipulated dimensions influenced listeners' responses. Higher F0 onsets induced more /aɻ.taɬ/, that is, *yin* /ta/ responses. For each F0 onset, most listeners perceived the voice quality continua in a categorical way from /aɻ.taɬ/ (*yin*) to /aɻ.daɻ/ (*yang*): stimuli at the breathy voice endpoint induced a large majority of /aɻ.daɻ/, that is, *yang* /da/ responses, whereas stimuli at the opposite endpoint induced a large majority of /aɻ.taɬ/, that is, *yin* /ta/ responses. These findings suggest that listeners perceived breathy voice as a cue to the *yang* identity of a syllable, at least at the time of Ren's study.

Unfortunately, Ren's study had some methodological shortcomings. As a main concern, the unique minimal pair he used is somewhat questionable. While 鞋带 'shoelace' is a frequent word, indeed produced [ɦaɻ.taɬ] due to the sandhi rule, 鞋大 'the shoe is big' is an artificially constructed verb phrase which would be seldomly produced by a Shanghai speaker: (1) when used as a noun, the morpheme 鞋 [ɦa] is usually suffixed with 子 [tsɻ]; (2) the morpheme 大 is pronounced /da/ in literary reading but /du/ in colloquial reading. In this context, speakers would tend to use the colloquial reading. Ren admitted it and instructed the listeners that this syllable would be read /da/. He wrote that no listener complained. Apart from the unbalanced frequencies between the two sequences, the fact that 鞋大 'the shoe is big' is an artificial sequence lead to a more drastic problem: whether the left-dominant sandhi rule should apply. If it applies, the *yang* /da/ syllable should surface as [da] but not [ta] in S2 position, and with little or no breathiness (Cao & Maddieson, 1992; also see our own data in Chapter 4). The important cues would then be phonetic voicing,

closure duration, and, potentially, F0 onset height. If it does not apply or if the right-dominant sandhi rule applies, the second syllable should not lose its original tone identity, thus the primary cue would be its F0 contour. In both cases, the critical ambiguity in Ren's material, segmental and tonal at the same time, is quite artificial.

To substantiate this point, we informally asked ten Shanghai speakers to read aloud the questionable sequence 鞋大. Eight young speakers in their twenties (all female) were native speakers of Shanghai Chinese. And the two other speakers, my parents, are in their fifties. My father is a native Shanghai speaker; my mother was born in the Jiangsu province and arrived in Shanghai when she married my father and speaks perfectly Shanghai Chinese since her youth. All these speakers pronounced the second syllable /du/. Five of them produced the sequence 鞋大 as [ɦa˩.du˦] and applied the left-dominant sandhi rule for lexicalized words, whereas the five others produced it as [ɦa˩.du˩] with the second syllable in its original tone T3 (the intervocalic dental stop was perceived as a sound between [t] and [d] by a native French speaker). Besides, all ten speakers judged this sequence unnatural. When asked to read this sequence, my father had a discussion with my mother, saying, "This is impossible. We don't say [ɦa.tu] but [ɦa.tsz̩.tu]."

Despite these shortcomings, Ren's study provided interesting data demonstrating that listeners used perceived breathiness as cue to *yang* tone identity, although one of them was influenced by voice quality in the opposite direction. We believe that the listeners might have focused on the second syllable as if it were an isolated syllable, ignoring the disyllabic context, and judged which of *isolated* 大 /da/ or 带 /ta/ the synthesized [ta] matched better.

In this Experiment 5, we aim at investigating the role of voice quality (more precisely, breathiness) in tone perception with improved materials over those used by Ren (1992). More importantly, we are interested in the relation between production and perception of voice quality in today's young Shanghai speakers. As shown in chapters 3 and 4, Shanghai Chinese changed since Ren's study toward a loss of *yang* tone breathy voice in young speakers.

Instead of breathiness continua constructed by manipulation of the open quotient, as in Ren's study, we constructed tone contour continua from *yin* rising T2 to

yang rising T3 imposed on monosyllabic words, with two patterns of voice quality: breathy and modal. We reasoned that breathy voice, if still used in perception, should bias stimulus identification toward *yang* responses. We tested monosyllabic targets, with stop, fricative, or nasal onsets. The continua we used were constructed from either natural or synthesized stimuli.

5.3.1 Method

We conducted two two-fold forced-choice identification tests using T2-T3 continua, one with synthesized stimuli, and the other with modified natural stimuli. For both types of stimuli, we contrasted breathy and modal voice. The stimuli were constructed from minimal pairs of monosyllables that differed in tone register (*yin* vs. *yang*) but shared identical segments and similar tone contour shapes (both T2 and T3 are rising), as in Experiment 4.

5.3.1.1 Speech materials

The stimuli were derived from 12 "base" monosyllabic words sharing the /ɛ/ rime, with six different onset (pair)s: zero onset (∅), labial or dental stop or fricative onsets, and labial nasal /∅, p-b, t-d, f-v, s-z, m/. The synthesized /m/ stimuli sounded unnatural, hence were not used. Checked syllables were not used since it was difficult to manipulate both breathy and glottalized qualities at the same time. Two eight-step tone contour continua (one breathy, one modal) were constructed for each onset from synthesized or natural base syllables. For each step, a contour between T2 and T3 was imposed on a base syllable. The endpoint base syllables of the continua (T2 and T3) are listed in Table 37. The meaning of each monosyllable can be found in Appendix 1.

Table 37. Materials of Experiment 5: endpoint base syllables at tones T2 and T3 (/m/-onset syllables were not retained for synthesized stimuli).

	∅	**stop**	**fricative**	**nasal**

T2 (Yin)	ɛ 爱	pɛ 板, tɛ 胆	fɛ 反, sɛ 伞	mɛ 美
T3 (Yang)	ɛ 咸	bɛ 办, dɛ 台	vɛ 烦, zɛ 馋	mɛ 梅

5.3.1.1.1 Synthesized syllables

In order to synthesize breathy and modal syllables, we used the VocalTractLab 2.1 software[22] to synthesize syllables from gestural scores, using a two-mass model triangular glottis (Birkholz, Kröger & Neuschaefer-Rube, 2011), in which each vocal fold was modeled by two clamp-shaped masses opening along the dorso-ventral axis. Compared to the usual two-mass model (Ishizaka & Flanagan, 1972), this model makes it possible to simulate incomplete closure of the glottis and thus different degrees of breathiness, as shown in Figure 103.

[22] http://www.vocaltractlab.de/

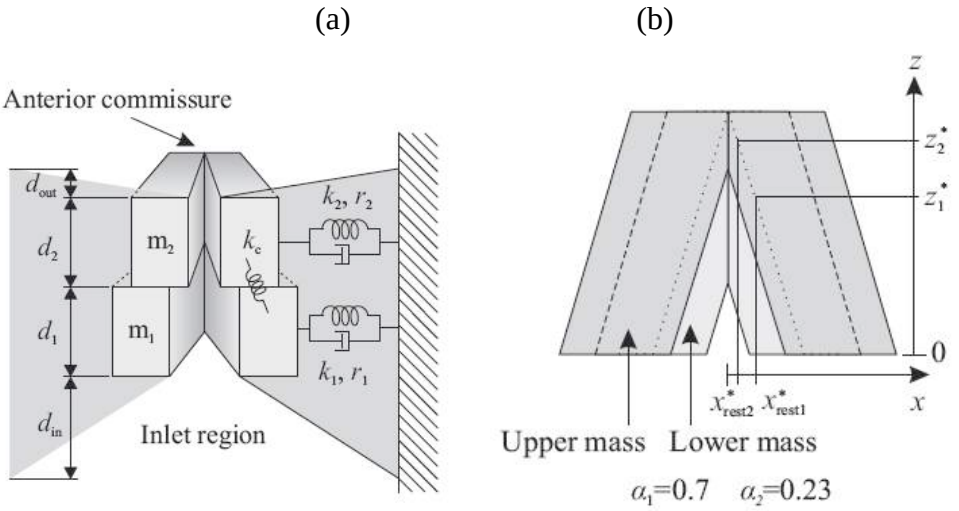


Figure 103. (a) Modelized vocal folds in rest position; (b) Top view of modelized vocal folds in partly closed position. (from Birkholz et al., 2011).

We synthesized the modal voice syllables with the default gestural score for modal phonation type in VocalTractLab 2.1. The controlled parameters were as follows: F0=120 Hz; Subglottal pressure=1000 Pa; Lower rest displacement=0.15 mm; Upper rest displacement=0.10 mm; Arythenoid area=0.00 mm^2; Aspiration strength= –40 dB. As for the breathy syllables, we set the *lower rest displacement* parameter to 0.60 mm, *upper rest displacement* to 0.55 mm, and the *subglottal pressure* to 1300 Pa. These settings resulted in two configurations of the two-mass model triangular glottis, as shown in Figure 104.

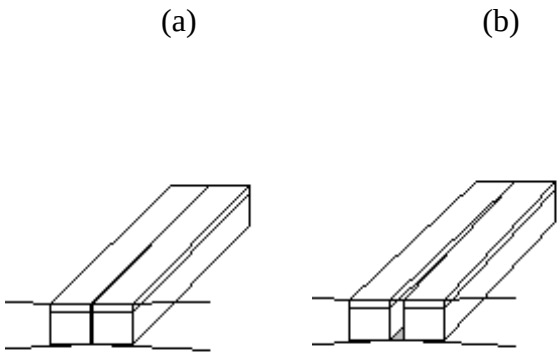


Figure 104. (a) modal and (b) breathy configurations of the two-mass model triangular glottis.

H1–H2 was measured in the synthesized base syllables to check their voice quality. H1–H2 was consistently higher for synthesized breathy than modal base syllables for all five onsets, and across the entire vowel. Table 38 shows the H1–H2 values of modal and breathy /ɛ/ at five time points in the vowel, according to syllable onset.

Table 38. H1–H2 values at five time points in the synthesized breathy (br) and modal (mo) syllables.

Time point	**P1**		**P2**		**P3**		**P4**		**P5**	
Onset	mo	br	mo	br	mo	br	mo	br	mo	br
∅	2.0	10.4	1.7	11.2	1.6	13.1	1.9	13.8	4.9	19.9
p	5.1	15.7	2.1	14.0	2.1	14.1	2.0	14.0	2.7	21.4
t	4.8	16.4	2.2	14.2	2.1	14.1	2.1	14.0	8.9	16.7
f	1.3	8.8	1.6	12.5	1.9	13.7	2.1	14.0	4.0	17.1
s	3.7	13.1	1.7	13.1	1.9	13.9	2.1	14.1	7.5	22.0
mean	3.4	12.9	1.9	13.0	1.9	13.8	2.0	14.0	5.6	19.4
mean diff. (breathy-modal)	**9.5**		**11.1**		**11.9**		**11.9**		**13.8**	

5.3.1.1.2 Natural syllables

T3 vs. T2 natural syllables served as natural base syllables for the breathy vs. modal continua, respectively. They were produced by the author of this dissertation, a female native speaker of Shanghai Chinese, aged 26 at the moment of the recording. As a trained phonetician, she deliberately produced breathy voice for T3 and modal voice for T2 syllables. H1–H2 was again measured on the two types of syllables to check for the two naturally produced phonation types. H1–H2 remained consistently higher for T3 than T2 syllables, for all six onsets, from vowel onset up to about 80% in the vowel. This differential was largest at vowel onset and decreased throughout the vowel, as observed in production. Table 39 shows the H1-H2 values for modal and breathy /ɛ/ at five time points in the vowel, detailed by syllable onset.

Table 39. H1–H2 values at five time points of natural base syllables. Bold values indicate higher values for breathy (br) syllables than their modal (mo) counterparts.

Time point	**P1**		**P2**		**P3**		**P4**		**P5**	
Onset	mo	br	mo	br	mo	br	mo	br	mo	br
∅	0.1	**16.4**	-0.9	**12.3**	-1.0	**10.2**	-0.7	**7.9**	3.7	**7.7**
p	-3.6	**20.7**	0.3	**11.1**	0.8	**11.8**	0.8	**11.2**	10.3	2.2
t	2.7	**13.0**	-2.4	**10.6**	-0.4	**9.3**	-0.7	**8.2**	4.6	**6.6**
f	5.7	**13.3**	-0.1	**8.3**	2.5	**13.0**	1.6	**7.4**	6.1	**6.4**
s	4.7	**8.9**	0.5	**10.7**	1.1	**11.3**	-1.0	**11.4**	15.3	7.8
m	0.4	**20.1**	2.8	**18.0**	3.4	**22.7**	0.5	**10.5**	5.6	0.4
mean	1.7	**15.4**	0.1	**11.8**	1.1	**13.1**	0.1	**9.4**	7.6	5.2
mean diff. (breathy-modal)	**13.7**		**11.8**		**12.0**		**9.3**		-2.4	

5.3.1.1.3 Tone contours and continua

All the base syllables, synthesized or natural, were intensity-equalized at 80 dB SPL. For each onset, the T2 and T3 syllables were time-scaled so that they had the same onset and vowel durations.

A continuum of eight equidistant, stylized tone contours between T2 and T3 were imposed on each base syllable, using the Praat (Boersma & Weenink, 1992-2015) implementation of PSOLA (Valbert, et al., 1992), following the same procedure as in Experiment 4. For the synthesized continua, the endpoint T2 and T3 contours were taken from the productions of a 74-year-old male speaker. For the natural continua, they were those produced by the trained phonetician for each onset. For each continuum, the eight contours from *yin* endpoint to *yang* endpoint were numbered from 0 to 7, as shown in Figure 98 (§5.1.1).

This yielded 10 synthesized and 12 natural continua (2 voice qualities × 6 or 5 onsets), hence a total of 80 (10*8) synthesized stimuli and 96 (12*8) modified natural stimuli.

5.3.1.2 Participants

Sixteen native speakers of Shanghai Chinese (11 female and 5 male), aged 18 to 26 years (mean age: 22), participated in the experiment.

All were born in Shanghai urban area, except one female who was born in Japan and came back to Shanghai at the age of 1 year old. All of them have spent most of their lifetime in Shanghai. All learned Standard Chinese before the age of 8. All learned English, six of them spoke Japanese at the beginner-intermediate level and one of them spoke Spanish at an advanced level. Four of them spoke some other Chinese dialects, among which Changsha dialect (Xiang), Cantonese (Yue), Wuxi and other Wu dialects spoken in Jiangsu and Zhejiang. Twelve of the participants were recruited from Tongji University in Shanghai where they majored in English, Japanese, or Environmental Engineering. No participant reported any hearing or reading disorder. If a participant was curious about the purpose of the experiment, she/he was given a short debriefing after the experiment. Before the experiment, however, all the subjects were naïve as to the purpose of the experiment. They were simply told the study was about the perception of words pronounced in Shanghai Chinese.

When questioned about their frequency of language use, most of them declared they frequently used Shanghai Chinese in their daily life with a subjective rating of 4 or 5 on a 1-5 scale;, only two of them rated their frequency of use of Shanghai Chinese below 3. The use of Standard Chinese was frequent for all of them (rated 5). English was one of their working languages for three of them (rated 3 to 5). Other dialects or languages were used only occasionally. Their self-evaluation of language competence was slightly lower for Shanghai than Standard Chinese: they rated Shanghai Chinese from 3 to 5, and Standard Chinese from 3.5 to 5. All of them evaluated themselves to speak and understand Shanghai Chinese no better than Standard Chinese.

5.3.1.3 Procedure

The identification tests were conducted using the E-Prime 2.0 software. Participants were tested individually in a quiet room, in front of a laptop. Stimuli

were presented to the participants through professional quality headphones. The testing procedure was the same as used in the identification test of Experiments 3 and 4. The stimuli were blocked by construction type. Half the participants were tested first on the natural stimuli, and the other half on the synthesized stimuli. Within each block, the stimuli were presented in random order. The experiment was preceded by a training phase of six trials, with syllable stimuli bearing canonical *yin* or *yang* tones. Participants received accuracy and response time feedback during the training but not during the test phase.

5.3.2 Results

For synthesized stimuli, the data from one 19-year-old male subject were excluded. He perceived the entire T2-T3 /fɛ/-/vɛ/ continua as tone T2 /fɛ/ syllable. Although there is a general trend for phonetically voiceless /f/-onset syllables (as were all the syllables of the /fɛ/-/vɛ/ continua) to be perceived as *yin* syllables (presumably because labial fricative onsets are often phonetically voiced in *yang* syllables, as explained in §4.1.6.2), we preferred to discard his data, which seemed extremely biased compared to the other subjects'.

For natural stimuli, the data from the same 19-year-old male subject and another 21-year-old male subject also had to be excluded. The first one had a missing rate of 6.8% during the experiment phase, compared to an average 0.62% for the other subjects. The second one perceived all of the breathy [sɛ] and [mɛ] syllables as *yang* tone T3 syllables through the entire T2-T3 continua, which is an as undesirable bias as the opposite bias mentioned earlier.

The *yang* identification curves for the synthesized stimuli, according to voice quality (VQ), averaged across onset-types and across 15 participants, are shown in Figure 105. These identification functions are quite categorical and show a category boundary shift toward the *yin* endpoint (T2) for the breathy relative to the modal continuum.

The *yang* identification curves for the natural stimuli, computed in the same way as for the synthesized stimuli are shown in Figure 106. The same category boundary

shift toward the *yin* endpoint (T2) for the breathy relative to the modal continuum can be observed.

Similar *yang* identification curves, detailed by onset-type, are shown in Figure 107 for synthesized stimuli, and Figure 108 for natural stimuli. A category boundary shift toward the *yin* endpoint (T2) for breathy relative to modal continua can be observed for stop and fricative onsets. For zero onset syllables, the category boundary shift is found for synthesized stimuli only. For nasal onset syllables (used in natural stimuli only), there is no category boundary shift.

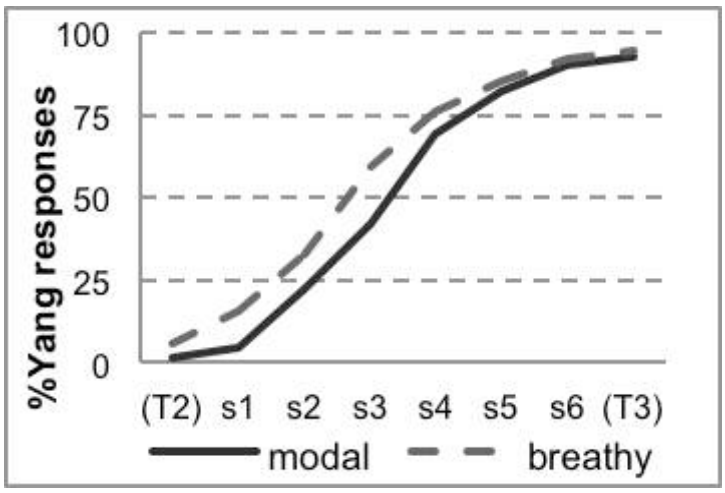


Figure 105. Averaged *yang* identification functions according to voice quality (synthesized stimuli).

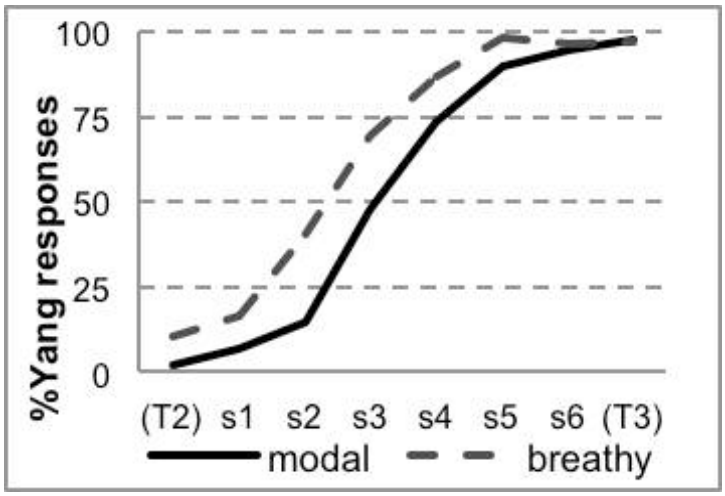


Figure 106. Averaged *yang* identification functions according to voice quality (natural stimuli).

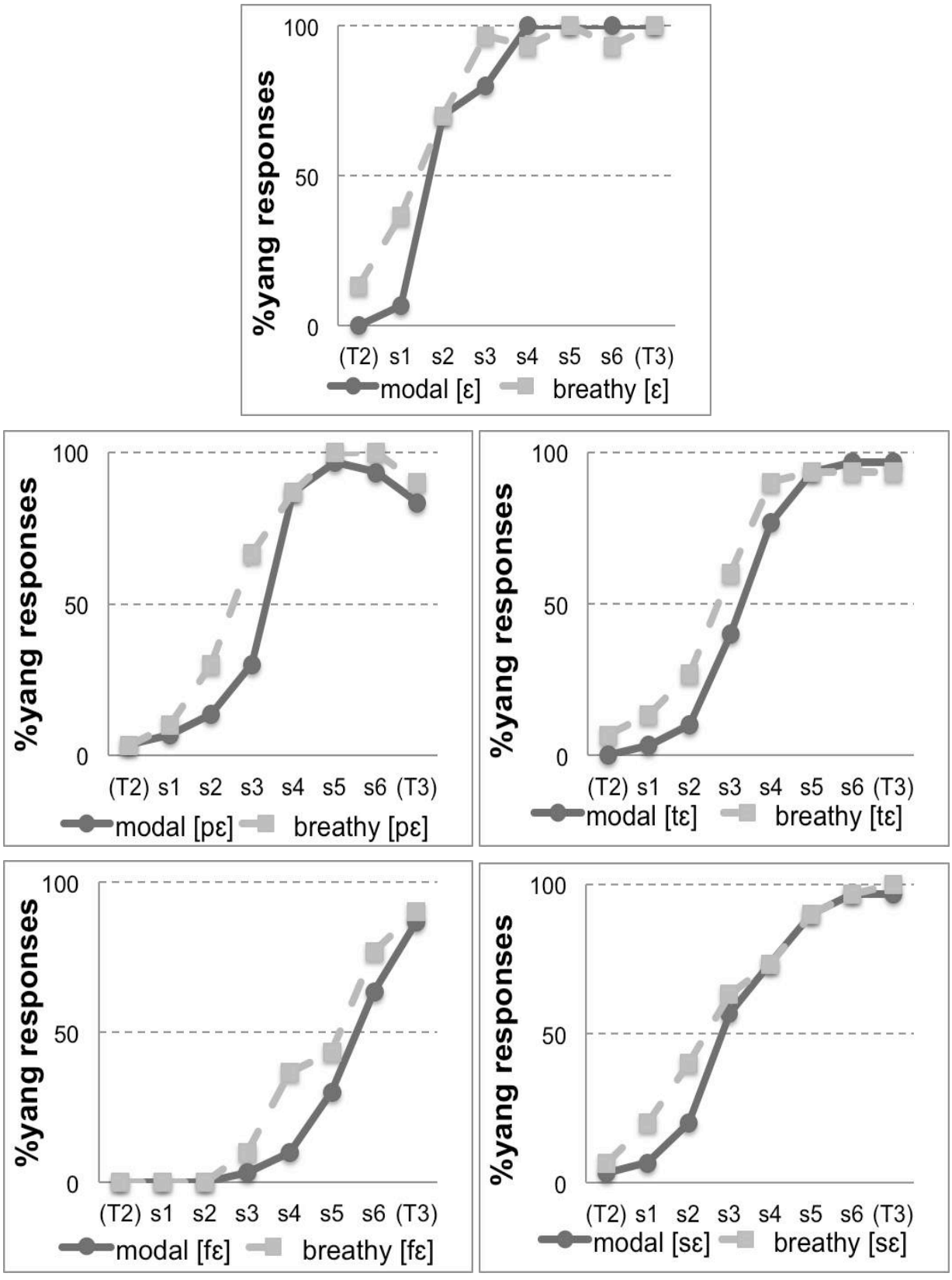


Figure 107. Averaged *yang* identification functions according to voice quality (synthesized stimuli); 1st line: zero onset; 2nd line: stop onsets; 3rd line: fricative onsets.

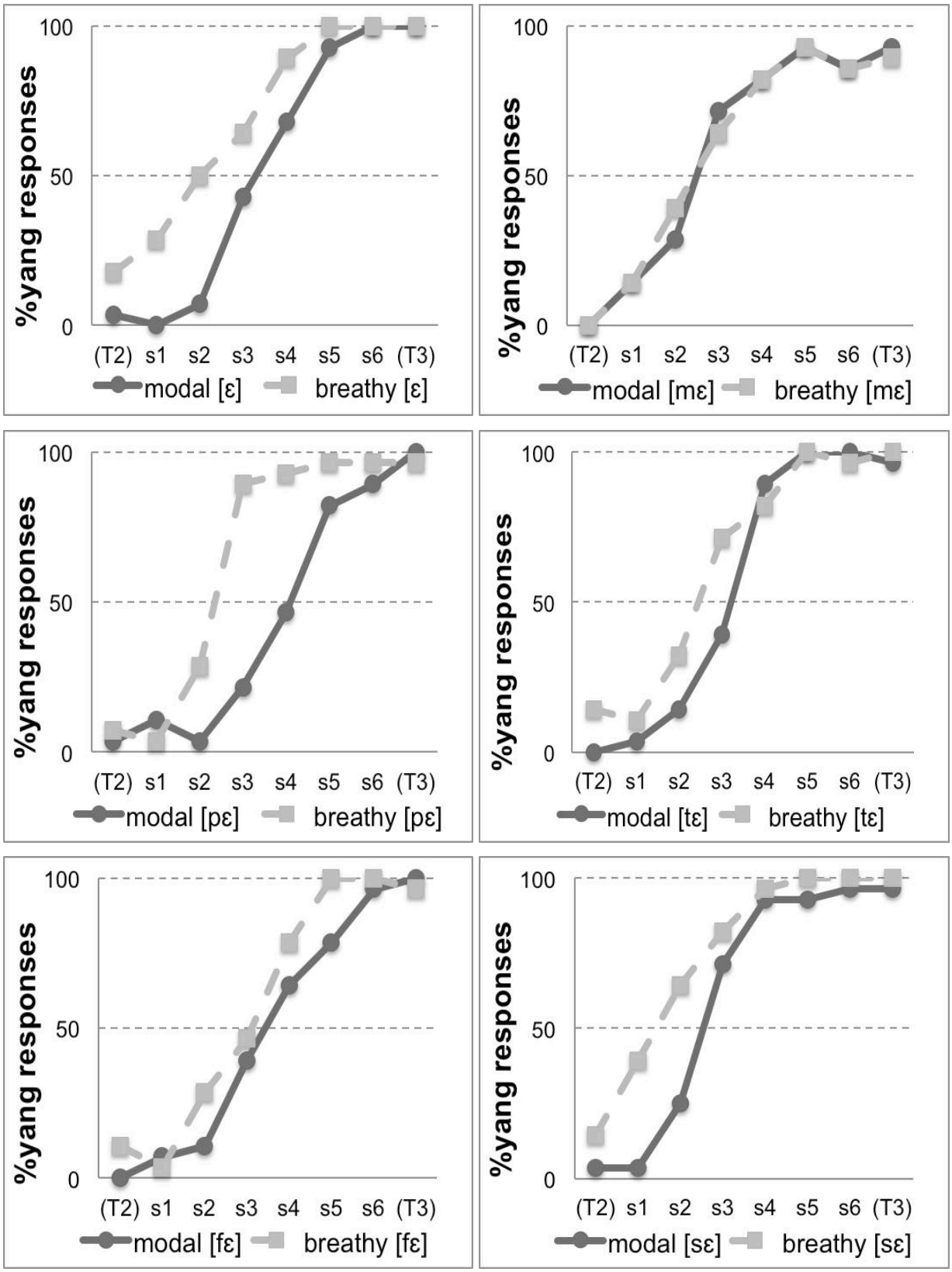


Figure 108. Averaged *yang* identification functions according to voice quality (natural stimuli); 1st line: /m/ and zero onsets; 2nd line: stop onsets; 3rd line: fricative onsets.

5.3.2.1 50% *yang* category boundary location

As in Experiment 4, we estimated category boundary locations (for the perceived *yin* and *yang* categories), or intercepts, as the stimulus (step) number where the *yang* identification rate reached 50%, for each continuum and each listener, using probit analyses fitting short ogive Gaussians to the raw data, thus yielding intercepts and slopes (Best & Strange, 1992; Hallé et al., 1999)[23]. For the synthesized stimuli, the intercepts averaged to 2.60 vs. 3.21 (step number) for breathy vs. modal continua, respectively. For natural stimuli, similar values obtained: 2.51 vs. 3.06. These differences indicate a shift toward *yin* responses for breathy compared to modal voice continua. Table 40 shows the intercepts (step numbers from 0 to 7) averaged across participants as a function of onset-type and voice quality for natural and synthesized stimuli. For each onset, the category boundary location was closer to the *yin* endpoint for the breathy than modal continuum, suggesting that breathy voice biases syllable identification toward *yang* tone syllables.

Two by-subject ANOVAs were conducted separately for synthesized and natural stimuli. *Boundary location* (in step number) was the dependent variable and *Subject* was the random factor. *Voice quality* (modal vs. breathy) and *Onset* type (synthesized: /∅, p-b, t-d, f-v, s-z/; natural: /∅, p-b, t-d, f-v, s-z, m/) were within-subject factors.

For synthesized stimuli, *Voice quality* was significant, with the category boundary closer to the *yin* endpoint for breathy than modal syllables (2.60<3.21), $F(1,14)=23.3$, $p<.0005$. *Onset* was highly significant, $F(4,56)=39.6$, $p<.0001$. For zero-onset syllables, the category boundary was the closest to the *yin* endpoint (1.7), suggesting they were more likely to be perceived as *yang* syllables than for the other onsets. In contrast, ,the boundary was the closest to the *yang* endpoint (4.8) for /f/ onset syllables, suggesting the opposite bias. The *Voice quality* × *Onset* interaction was not significant, $F(4,56)=1.1$, $p=.34$.

For natural stimuli, *Voice quality* was again significant, with the boundary closer to the *yin* endpoint for breathy than modal syllables (2.51<3.06), $F(1,13)=22.1$,

[23] We also tried to fit Gaussian cumulative distribution functions to all the eight data points of each individual continuum and found similar intercept values.

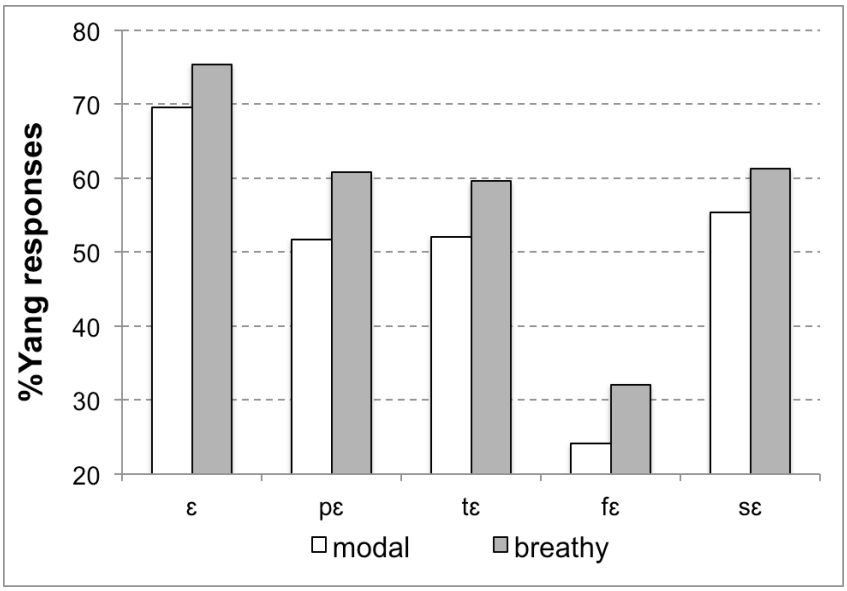
p<.0005. *Onset* was also significant, $F(4,56)=6.7$, p<.0001. Onset types can be ordered as follows, based on their distance from the *yin* endpoint (/s/ (2.05) < /m/ (2.56) < /t/ (2.75) < ∅ (3.03) < /p/ (3.18) = /f/ (3.18)). The *Voice quality* × *Onset* interaction was significant, $F(5,65)=2.8$, p<.05, reflecting that the breathy vs. modal difference was significant for the zero, /p/ and /s/ onsets (∅: Δ=0.65, $F(1,13)=7.7$, p<.05; /p/: Δ=1.14, $F(1,13)=9.4$, p<.01; /s/: Δ=0.97, $F(1,13)=15.1$, p<.005), but not for the other onsets (/t/: Δ=0.22, $F(1,13)<1$; /f/: Δ=0.36, $F(1,13)=2.4$, p=.15; /m/: Δ=–0.11, $F(1,13)<1$).

Table 40. Averaged intercept data (in step number) as a function of voice quality for synthesized and natural stimuli. Significance levels for the breathy vs. modal paired comparisons: * for p<.05, ** for p<.01.

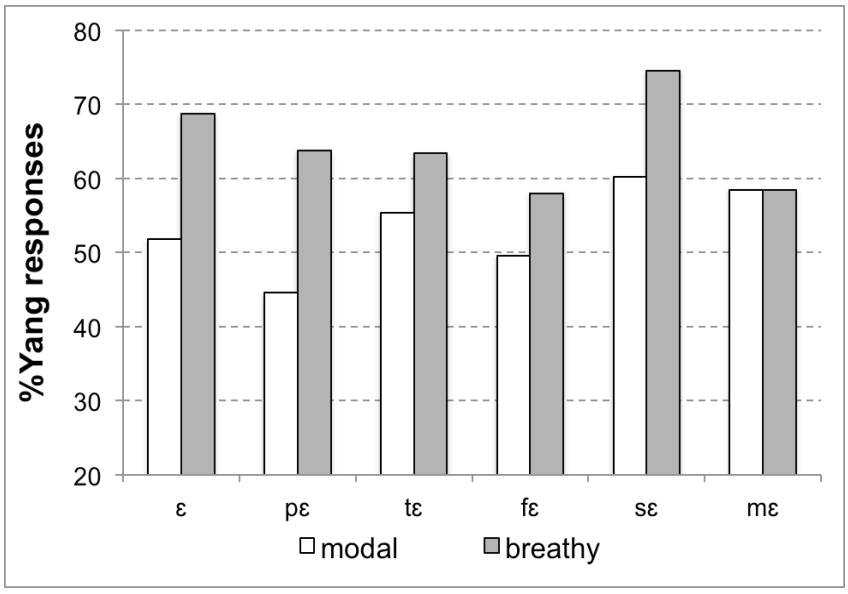
	Synthesized		**Natural**	
Onset	**Breathy**	**Modal**	**Breathy**	**Modal**
zero	1.63	1.77	2.71	* 3.36
p	2.17	** 3.00	2.62	** 3.75
t	2.44	* 3.20	2.64	2.86
f	4.47	5.11	3.00	3.36
s	2.40	** 3.07	1.57	** 2.54
m	—	—	2.61	2.50
Mean	**2.60**	**** 3.21**	**2.51**	**** 3.06**

5.3.2.2 Overall percentage of *yang* responses

In the synthesized stimulus data, the overall rate of yang responses was 57.8% for breathy syllables vs. 50.6% for modal syllables. In the natural stimulus data, these rates were 64.5% for breathy syllables vs. 53.3% for modal syllables. Figure 109 shows the *yang* response rate averaged across all stimulus steps and across subjects, as a function of onset-type and voice quality. For the /m/ onset (natural stimuli), there was no advantage for breathy over modal voice quality.



(a)



(b)

Figure 109. *yang* response rate averaged across all stimulus steps as a function of voice quality and onset-type for (a) synthesized and (b) natural stimuli.

A series of generalized linear models (GLM) were fit to the binomial *yin*/*yang* response data, using the *lme4* package (Bates, Maechler & Dai, 2008) in R (R Core team, 2014), separately for synthesized and natural stimuli. *Subject* was the random factor; in the full model, *Voice quality* (modal vs. breathy), *Onset* (synthesized: /∅, p-b, t-d, f-v, s-z]; natural: /∅, p-b, t-d, f-v, s-z, m]), *Step* (0 to 7) and *Gender* (male, female) were the fixed factors. The random factor structure included by-participant random intercept. Likelihood ratio tests were performed to compare the full model, which included all the fixed factors, to models with one fixed factor missing, using the *anova*

function. Another model including *Voice quality* × *Onset interaction* effect was also compared with the model without this interaction effect.

In the synthesized stimuli data, *Voice quality* was significant, $\chi^2(1)=36.4$, $p<.0001$, and the *Voice quality* × *Onset* interaction was not, $\chi^2(4)=0.95$, $p=.92$, indicating similar effects for the five onsets /∅, p, t, f, s/. The detailed output of this GLM analysis with R, including the other predictors appears in Appendix 3.

In the natural stimuli data, *Voice quality* was significant, $\chi^2(1)=89.3$, $p<.0001$. The *Voice quality* × *Onset* interaction was also significant, $\chi^2(5)=29.5$, $p<.0001$. We checked the *z* values for the fixed factors in the model including the *Voice quality* × *Onset* interaction, and found that the /m/ onset differed from the others concerning the *Voice quality* effect. We thus extracted the data subset for the /m/ onset and ran a likelihood ratio test on this subset data, comparing a model including the Voice quality factor with the model without this factor. The two models did not differ, showing that *Voice quality* was not significant for the /m/-onset syllables, $\chi^2(1)=0.005$, $p=.95$. The detailed output of this GLM analysis with R appears in Appendix 3.

5.3.2.3 Response time

Figure 110 shows the average identification response time (RT), according to voice quality: As for the RT data of Experiment 4, only the *yin* response RTs are shown (and analyzed) from step 0 to 3 (where *yin* responses are dominant) and only the *yang* response RTs are shown and analyzed from step 4 to 7 (where *yang* responses are dominant). For natural stimuli, the *yin* RTs at steps 1-3 were clearly longer for the breathy than modal syllables whereas the *yang* RTs at steps 4-7 were clearly longer for the modal than breathy syllables, suggesting that modal voice quality impedes *yang* responses, whereas breathy voice quality impedes *yin* responses. The results were less clear-cut for synthesized stimuli, although similar trends were observed.

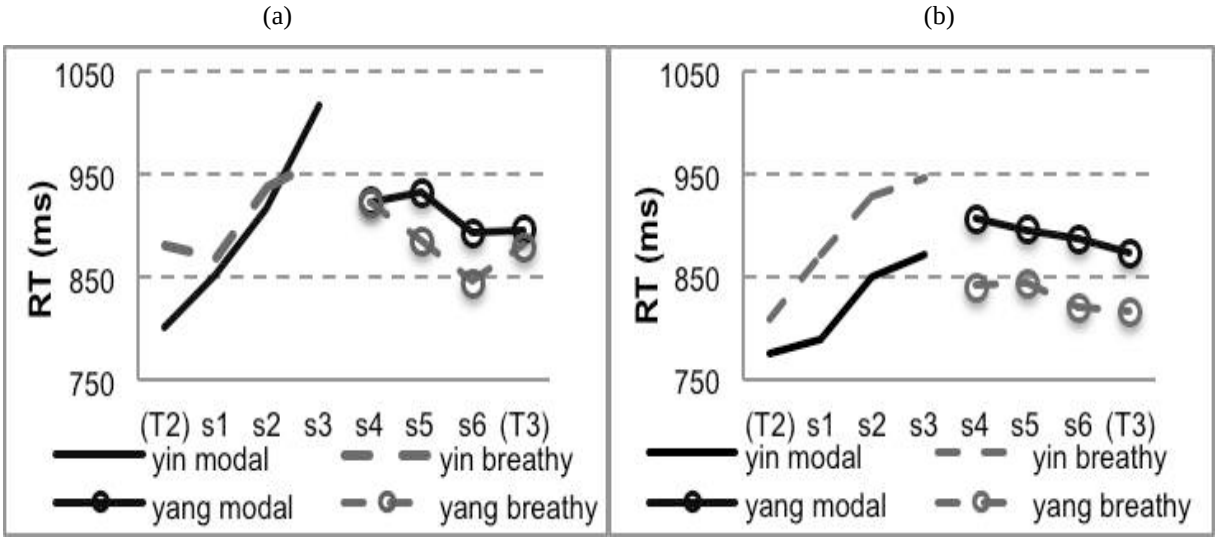


Figure 110. Average *yin* responses RTs in the *yin*-dominant region (steps 0–3: dashed lines) and *yang* responses RTs in the *yang*-dominant region (steps 4–7: solid lines) for modal (black) vs. breathy (gray) voice, and (a) synthesized vs. (b) natural stimuli.

We examined the impact of the congruence between voice quality and tone register on RTs. Reasoning as for the RT data of Experiment 4, we defined Congruent stimuli as those stimuli whose voice quality matches the dominant response, and Incongruent stimuli as those stimuli whose voice quality mismatches the dominant response. By this definition, the modal voice stimuli are congruent (with *yin* responses) at steps 0-3 and incongruent (with *yang* responses) at steps 4-7, and vice versa for the breathy voice stimuli.

RTs were shorter overall for Congruent than Incongruent stimuli, for both synthesized (876<914 ms) and natural stimuli (833<893 ms).

In order to substantiate these observations, two by-subject ANOVAs were run separately for synthesized and natural stimuli. *RT* was the dependent variable and *Subject* was the random factor. *Congruence* (congruent vs. incongruent), *Dominant steps* (*yin*-dominant steps 0-3 vs. *yang*-dominant steps 4-7) and *Onset* type (synthesized: /∅, p-b, t-d, f-v, s-z/; natural: /∅, p-b, t-d, f-v, s-z, m/) were within-subject factors.

For synthesized stimuli, *Congruence* was marginally significant, with shorter RTs for Congruent than Incongruent stimuli (833<893 ms), $F(1,14)=4.3$, $p=.057$, and did not interact with either *Onset* or *Dominant steps*, Fs<1.

For natural stimuli, *Congruence* was significant, with shorter RTs for Congruent than Incongruent stimuli (876<914 ms), $F(1,13)=9.6$, $p<.01$. *Congruence* did not interact with *Dominant steps*, $F(1,13)<1$, but the *Congruence* × *Onset* interaction was significant, $F(5,65)=2.9$, $p<.05$: *Congruence* was significant for the /∅, f/ onset-types and marginal for the /s/ onset-type, with an advantage of Congruent over Incongruent stimuli, but *Congruence* was not significant for the other onsets, with a numerical advantage of Incongruent over Congruent stimuli for some onsets.

5.3.3 Discussion

Experiment 5 showed that breathy voice biases tone perception in young Shanghai listeners, both males and females, toward the *yang* category. This conclusion is supported by the shift toward *yang* responses, for most onsets, induced by breathy compared to modal voice in the identification of T2-T3 continua, as shown both by the shift of the intercepts toward the *yin* endpoint and by the overall increase in *yang* response rate, as well as by the RT data, although the RT effects were not found for all onset-types. Also, the results were less clear-cut for synthesized than natural stimuli.

The greater effect observed in natural than synthesized stimuli could be explained by the fact that (1) the stimuli's naturalness helped the participants to identify the target syllables' tone[24]; (2) other uncontrolled cues in the natural stimuli might play a role in tone identification; (3) natural stimuli were longer in duration than synthesized ones, so that the acoustic cues accompanying the tone (here, breathy vs. modal voice quality) were more salient and easier to detect.

There was a notable exception for the /m/-onset syllables. The identification functions of the T2–T3 continua for the breathy and modal voice versions of /mɛ/ were virtually undistinguishable, and no RT effect was observed that could suggest a *yang* bias induced by breathy voice. Our production also showed no difference in voice

[24] This was also a general comment made by most participants: they found the natural stimuli easier to identify than the synthesized stimuli.

quality between *yin* and *yang* syllables with a nasal onset. In other words, for nasal onset syllables such as /mɛ/, there is little use of breathy voice in both production and perception. Two scenarii are possible. Either *yang* nasal onsets never developed breathiness since Middle Chinese, or they did develop breathiness, perhaps in analogy with the *yang* obstruents, and are on the track of losing it nowadays.

According to Rose's (1989, 2002) description, although without phonetic measures, breathiness existed with all *yang* tone syllables in (Northern) Wu dialects, including zero and sonorant onset syllables. In fact, it is not unknown that nasal onset syllables are among the first to be affected by the loss of a phonation difference between high and low tones. The production data of eight Yue dialects examined by Tsuji (1977, [cited in Yip, 1980: 139]) illustrate different stages in the loss of breathy voice for low tone syllables. Rongxian 容县 dialect preserved breathy voice, whereas Cangwu 苍梧 dialect lost it in low tone for all onsets. Between these two stages, which we may view as initial and final, Cenxi 岑溪 dialect lost breathy voice only for nasal onset low tone syllables, representing an intermediate stage.

Our perception data suggest that Shanghai Chinese might be on the same tracks as the Cenxi dialect, the loss of breathy voice being more clearly achieved in production, and emerging with nasal onset syllables in perception. This said, the voice quality was found to be an important secondary cue in tone identification. Breathy voice facilitates *yang* tone perception, although our production data (chapter 4) suggest a trend toward the disappearance of the phonation type (or voice quality) difference between *yin* and *yang* syllables.

6 GENERAL DISCUSSION

After we have presented our experimental results, we discuss and attempt to address the issues and answer the questions asked in §3.4.

6.1 A description of acoustic and articulatory correlates of Shanghai tones based on production and perception data

We investigated the production and perception of some phonetic correlates of Shanghai tones, in particular, of the two tone registers *yin* and *yang*, which are intricately associated with the phonologically voiceless and voiced obstruent onsets (henceforth *yin* and *yang* obstruent onsets), respectively. Of course, we are aware that the list of the correlates we investigated is far from exhaustive.

We first summarize the issues we addressed and our main results in this section. We then discuss the phonetic and phonological motivations in the following section.

We first asked the following questions related to the production aspects:

(1a) What are the articulatory and acoustic correlates of Shanghai tones?

(1b) Which phonetic measures are the most efficient to distinguish Shanghai Chinese breathy voice from modal voice?

(1c) To what extent are there intra- and/or inter- speaker variations in the realization of these phonetic correlates?

Prior experimental studies (e.g., Cao & Maddieson, 1992; Ren, 1992) mostly confirmed the traditional description of word-initial *yang* obstruents as 清音浊流 'clear (i.e., voiceless) sound followed by muddy (i.e., voiced) airflow': an impressionistic but lucid description first proposed by Chao Yuen-Ren (1928). These studies also described word-medial *yang* obstruents as phonetically voiced. Our production study (Experiments 1 and 2) attempted to qualify these descriptions and provide additional precisions. To start with, we included zero and nasal onsets in our study, to clarify the

issue of whether breathy voice is restricted to *yang* obstruent onset syllables or, rather, is a common characteristic of all *yang* tone syllables.

- **Voicing**

Concerning the phonetic voicing associated with the two tone registers, *yin* and *yang*, our data partly confirm the 清音 'clear sound' part in the above mentioned description of *yang* onsets in word-initial position. *Yang* stop onsets are predominantly realized phonetically as voiceless in word-initial position (and as voiced in word-medial position as we will explain in the following). However, *yang* fricative onsets, especially labial fricative onsets, are often realized as voiced in word-initial position and, as it seems, regardless of the word length: at least, this is what we observed for both monosyllabic and disyllabic words.

As for the word-medial, that is, intervocalic context, we not only found predominant voicing of *yang* obstruents, but also found spirantized forms of *yang* stops.

Inter-speaker variations. The voiced realization of word-initial fricatives is more frequently observed in young speakers compared to elderly speakers, and in female speakers compared to male speakers. The spirantized realization of word-medial fricatives is, on the contrary, more frequently observed in elderly speakers compared to young speakers.

- **Phonation type**

Concerning the phonation type correlated with tone register, the 浊流 'muddy airflow' part of the traditional description, suggesting a breathy phonation, is confirmed by several objective measures. The Harmonics to Noise Ratio (HNR) data indicate more breathiness during the release of *yang* than *yin* stops, but not during the production of *yang* nasal stops. Longer VOT values of *yang* stops also suggest a delayed vibration of the vocal folds in the production of the following vowel, presumably due to a slack configuration of the folds. Among all the acoustic measurements on the vocalic part, the H1–H2 data most robustly show breathier phonation in vowels following *yang* than *yin* onsets (henceforth *yang* vowels) in the monosyllabic context, with substantial inter-speaker variations, which are related to

age and gender: the phonation type difference in terms of breathy voice is larger for elderly than young speakers, and for male than female speakers. Electroglottographic (EGG) measurements of the vocal fold vibration reveal open quotient (OQ) values that also indicate breathier phonation in *yang* than *yin* vowels, a difference which is consistently found only for elderly male speakers. (Admittedly though, the number of speakers in each of our speaker groups in our EGG study was quite limited.)

Time course. The HNR data show breathier phonation in the release part of word-initial *yang* than *yin* oral stops for both monosyllabic and disyllabic words. Several acoustic measures (spectral tilt measures such as H1–H2, noise-level measure, F1 height) and the OQ physiological measure are diversely successful in revealing breathier phonation for *yang* than *yin* vowels. When they are, they indicate that breathy phonation is most prominent at vowel onset and is maintained until vowel midpoint, before it decreases in the second half of the vowel. Breathy phonation thus is not limited to the consonant release and vowel onset parts, as proposed by Cao and Maddieson (1992) and Ren (1992).

Degree of breathiness. Limiting the domain of observation to the first 40% of the vowel (our "time points" 1 and 2) and to unchecked monosyllables, the H1–H2 differential between *yin* and *yang* syllables approximately reaches an average 6 dB for elderly male speakers, 3.5 dB for elderly females, and about 3 dB for young speakers. For stop onset syllables, the H1–H2 differential is even higher for elderly speakers (around 9 dB for elderly male speakers, and 6.5 dB for elderly female speakers). Previous studies reported an average H1–H2 differential of 6 dB between breathy and modal vowels in Gujarati, and of 9.7 dB in !Xóõ (Bickley, 1982). Hence, the size of the H1–H2 difference between *yin* and *yang* syllables in Shanghai elderly male speakers' productions is of the same order of magnitude as in those languages that make a contrastive usage of breathy vs. modal voice. However, a lesser degree of breathiness is observed with the three other speaker groups of our study.

Onset context. Breathy voice quality in *yang* syllables is most prominent after *stop* onsets, but is also observed after fricative onsets as well as the *zero* (null) onset. It should be noted that the *yang* zero onset (when not followed by a glide) evolved from the Middle Chinese /ɦ/ onset. In fact, *yang* zero onset syllables are often transcribed phonetically with an initial [ɦ]. Moreover, a short breathy friction part distinct from

the following vowel can be observed for some speakers, although this pattern is far from being systematic. Breathy voice thus is not restricted to obstruent onset syllables. However, it is not systematically associated to the *yang* tone register. Indeed, nasal onsets stand out as an exception. Most acoustic measures show little or no difference between *yin* and *yang* syllables with nasal onsets, except the H1–A2 measure. The nasal onset exception is also found in our perception data: Shanghai listeners are insensitive to the (artificial) presence of breathiness in nasal onset syllables.

Within-word position. The breathy voice phonation difference is greatest in monosyllabic words. In disyllabic words, the difference is small when the target syllable is the first syllable, and is not related to the underlying *yin* vs. *yang* tone register when the target syllable is the second syllable. In this case (second syllable of a disyllabic word), due to the application of a tone sandhi rule, the original phonological tone of the target syllable is neutralized and the realized tone contour is solely determined by the first syllable's tone. The pattern that seems to emerge in this case from our data is that H1–H2 is slightly larger when the target syllable is realized with a low than a high F0.

Inter- and intra- speaker variation. All the acoustic and physiological measures, supported by statistical analyses, show that the breathy voice phonation difference between *yin* and *yang* syllables, both in the consonantal part (stop release) and the vocalic part, is clearly realized by elderly male speakers, less consistently so by elderly female and young male speakers, and hardly realized at all by young female speakers. For speakers who do not produce acoustically breathier phonation on *yang* than *yin* tones (as indexed, mostly, by the H1–H2 measure), a reversed OQ pattern is observed. That is, larger OQs characterize higher rather than lower F0 syllables, as was observed by Koreman (1996) for Dutch. We explain this inter-speaker difference by a tradeoff between phonation type and F0. Aside from elderly male speakers, those speakers who exhibit a large F0 dynamics tend to rely less on breathy voice phonation to distinguish *yin* and *yang* syllables; in stark contrast, those speakers who exhibit a small F0 dynamics tend to produce breathier phonation on *yang* than *yin* tone syllables. A particular voice quality was observed in several *yang* syllables produced by two elderly male speakers without being systematic. It could be defined as "harsh

whispery," suggesting that the phonation type that contrasts with modal voice may have variable realizations, as long as it can be distinguished from the modal voice.

- **Duration**

Globally, the duration pattern related to tone register is as follows: *yin* (phonologically voiceless) obstruent onsets are longer than *yang* (phonologically voiced) obstruent onsets, and *yin* rimes are shorter than *yang* rimes, all other things being equal.

Onset duration. Fricatives in all positions, phonetically voiced or not, follow the "*yin* longer than *yang*" pattern. Intervocalic stops follow this pattern as well. Word-initial stops, in our pilot articulatory study with only two speakers, show a tendency of longer *yin* than *yang* duration as well. Nasal onsets, again, are an exception and show no sign of differential duration according to tone register. Thus, at least for obstruent consonant onsets (the case of zero onsets is debatable), the duration difference between phonologically voiceless and voiced consonants is preserved, although the *phonetic* voicing contrast is partly lost in word-initial position. We propose two explanations, which are not mutually exclusive. First, since *yang* obstruent onsets are associated with slacker articulatory setting of the folds than yin obstruents, due to breathier phonation in word-initial position or to voicing in word-medial position, the short duration might be attributable to a laxer articulation. The same account would apply to Korean: closure duration is the longest for fortis stops and the shortest for lenis stops (e.g., Silva, 1992). That is, stop obstruents are longer for tense than lax articulation. Second, and perhaps more importantly, that the duration pattern characteristic of the phonological voicing contrast is maintained in all contexts in spite of partial phonetic neutralization, suggests that, to some extent, speakers still maintain the supralaryngeal articulatory phasing relationship typical of the phonological voicing contrast.

Rime duration. In citation form, rimes (/ɛ/ or /aʔ/ in our study) are the shortest when carrying T1 and the longest when carrying T3. In disyllables, the same difference is found but to a lesser extent. Among three explanations for these rime duration differences (see §3.3), namely, the consonant/vowel compensation explanation (Gao & Hallé, 2012), the tone height explanation (Gandour, 1977), and

the phonation explanation (e.g., Fischer-Jørgensen, 1967), we can discard the consonant/vowel compensation explanation, since vowels after a zero or nasal onset are shorter for *yin* than *yang* syllables, in the absence of difference in consonant duration. Besides, to our knowledge, this pattern is rarely found in other languages (but see Cho & Giavazzi, 2009). Since young speakers produce less breathy voice on *yang* syllables, and at the same time, less duration difference between *yin* and *yang* rimes, we are tempted to attribute the rime duration difference mainly to the phonation difference. Breathy vowels tend to be longer than modal ones. That the difference is even more robust for checked than unchecked syllables further support the phonation account. In *yang* syllables, laryngealized vowels (/aʔ/ in our study) begin with a slack vocal folds setting (breathy or voiced) and end with a constricted folds setting (glottal stop), whereas, in *yin* syllables, these vowels begin with a critical constriction of the folds, a setting closer to the syllable-final constriction than the slack setting. Thus, a more complex laryngeal maneuver, presumably requiring more time, needs to be executed for *yang* than *yin* checked syllables. The phonation account might thus also explain the robust *yin–yang* difference in vowel duration we found for /aʔ/ compared to /ɛ/. Tone height might somewhat contribute to the *yin–yang* difference in vowel duration: for nasal onsets, for which little or no *yin–yang* difference in phonation is found, high tone *yin* rimes are still shorter than low tone *yang* rimes. Thus, low tone vowels might tend to be longer than high tone vowels, as suggested by Gandour (1977).

Inter-speaker variation. All speaker groups share the same duration pattern for onsets: longer *yin* than *yang* obstruent onsets. The duration pattern differs among speakers with respect to rimes in unchecked syllables. (It is robust and stable across speakers for checked syllables.) As mentioned earlier, the difference between rimes carrying T2 (*yin*) and T3 (*yang*) is larger for elderly than young speakers. This may be related to the loss of T3 breathy voice in T3 syllables found in young speakers, supporting again the phonation account discussed in the preceding subsection. Finally, we observed longer VOTs for male than female speakers and longer rimes overall for female than male speakers. We explain this cross-gender variation by different strategies used for contrasting the *yin* and *yang* categories. Female speakers lengthen

the rime, which carries the tone contour information, whereas male speakers lengthen the onset, which carries an important part of the phonation type information.

- **F0 onset**

We compared the F0 onset height of the second syllable of a disyllable (S2) between voiced and voiceless onsets. Theoretically, in case of application of the left-dominant tone sandhi rule, S2 loses its tone identity and receives the tone component spread from the preceding syllable. Hence, the tone contour is not related to the voicing of the S2 onset. However, prior studies (Ren, 1992: 51; Chen, 2011; Chen & Downing, 2011) found a depressor effect of intervocalic voiced stops, that is, the F0 onset is lowered by voiced stop onsets compared to voiceless counterparts.

Our data showed a global depressor effect of voiced stops and fricatives, but a further examination of the data provided more precisions. The depressor effect is found when the realized tone is high, that is, when the first tone carries T2, but not when the realized tone is low, that is, when the first tone carries T1 (we did not investigate the post-T3 sandhi context). Our explanation for the post-T1 situation is that the depressor effect of voiced consonants in post-T1 sandhi context is compensated by the ballistic F0 trajectory effect programmed on the whole prosodic word. The S2 vowel onset occurs earlier when preceded by a voiced onset because of its shorter duration, and thus is assigned a higher F0 onset, since the F0 trajectory falls monotonically at a constant velocity. If this explanation is correct, it suggests that the F0 trajectory is programmed on the whole prosodic word at an independent level of the syllable segments.

Inter-speaker variations. The ballistic F0 trajectory effect and the voiced onsets' depressor effect are observed in all speaker groups. The difference lies in the duration of the depressor effect between young and elderly speakers. This duration is dependent on the syllable duration for young speakers, but not for elderly speakers.

We now turn to the issues concerning the perception aspects:

(1d) How do Shanghai listeners perceive all these cues?

The results from Experiment 3, in which we interchanged F0 tone contours between *yin* syllables and their *yang* counterparts, thereby constructing stimuli whose

F0 contour mismatched all characteristics other than F0 contour. We tested Shanghai Chinese listeners' tone identification of these "mismatched" stimuli compared to "matched" stimuli constructed with the same speech transformation technique but whose tone contour was qualitatively maintained. The tone identification accuracy data show that the dominant cue for tone identification is F0 contour and height, as well as voicing in the case of labial fricatives, for both young and elderly, female and male listeners. However, listeners responded more slowly to mismatched than matched stimuli and rated lower their naturalness, suggesting sensitivity to cues other than F0 contour and height in the native perception of Shanghai Chinese syllables, even though such cues are largely ignored in a tone categorization task. We thus proceeded to investigate the perceptual role of the phonetic cues that were examined in the production experiment part of this work.

- **Duration**

In Experiment 4, we created tone continua from *yin* T2 and *yang* T3 imposed on fricative onset syllables with two duration patterns: long consonant followed by short vowel (LS), and short consonant followed by long vowel (SL). The results show an overall facilitation of *yin* tone identification with the LS compared to SL duration pattern, suggesting that duration pattern is an important secondary cue in tone perception. Only young listeners participated in the experiment, so we may only speculate that elderly speakers exhibit similar perception behavior, based on their production data.

- **Phonation type**

In Experiment 5, we created tone continua from *yin* T2 and *yang* T3 with two phonation types: breathy and modal. The results show an overall facilitation of *yang* tone identification with the breathy compared to modal syllables, suggesting that phonation type is another secondary cue to tone perception. Nasal onset syllables, however, make a notable exception: tone identification of nasal onset syllables is not affected by their phonation type. Participants were, again, young listeners, who, according to our production data, produced much less phonation difference than

elderly speakers. Our study thus suggests that the loss of a redundant feature in production is not necessarily triggered by the loss of its perceptual function.

- **Voicing**

We did not investigate directly the voicing cue in tone identification. However, the results of all three perception experiments indirectly show that the voicing cue is important in tone perception, at least for labial fricative onset syllables. The phonetically voiced labial fricative onsets we used in Experiment 3 greatly facilitate *yang* tone register identification, overweighing the imposed tone contour. Stimuli with phonetically voiceless labial fricative onset, in all three experiments, yield a high *yin* response rate, which is not found with the other syllable onsets. This perception behavior is probably related to the fact that the *yang* labial fricative is frequently produced as voiced. It remains to be discovered whether stimuli with phonetically voiced onsets would all facilitate perception of *yang* identity. At this time, we cannot conclude yet whether the voicing cue is important only for labial fricative onsets or for any obstruent onsets.

6.2 Classification of redundant features

Voicing, phonation type, and duration (in monosyllables), and F0 onset (in sandhi-modified S2 position) all correlate with Shanghai tone register and all participate in its perception by native listeners, some cues being more important than others. We did not investigate F0 onset in the S2 context but Ren (1992) did and found that this parameter influences tone identification. All these cues contribute to the definition of the *yin–yang* register contrast.

Mazaudon (2012) distinguishes between two types of "redundant" features. She avoided to oppose redundant features to "distinctive" features, but rather integrated them in the definition of a "toneme": (1) cues related to coarticulatory effects; (2) cues that are remnants of a diachronic transphonologization process, previously distinctive and later replaced, thus becoming redundant. In the second case, there should be a period in the evolution of the language at issue in which it is difficult to define which feature is distinctive and which is redundant, since several features co-exist and

overlap. The Tamang-Gurung-Thakali-Manangke languages seem to be now at this stage of their evolution (Mazaudon, 2012; Mazaudon & Michaud, 2008).

We thus further asked:

(2) Are redundant features due to coarticulatory effects or are they remnants of diachronic changes?

Shanghai Chinese, as the other Chinese dialects, has undergone the tone split initially triggered by a voicing distinction (or, as has been argued, by a phonation type distinction). In most Chinese dialects, the tone split resulted in a *yin–yang* tone-register opposition and the loss of the voicing distinction. But Shanghai Chinese, like the other Wu dialects, has maintained the voicing contrast. It may be related to the left-dominant tone sandhi rule in Wu dialects, after which tonal distinctions are mainly maintained in the word-initial syllable (Ballard, 1989). In word-initial position, the *yin–yang* distinction surfaces as a tone-register distinction, as in, for example, Cantonese. Breathy voice is therefore unnecessary to the *yin–yang* distinction and can thus be viewed as a redundant feature that accompanies *yang* tones. In other positions, the *yin–yang* distinction surfaces as a phonetic voicing distinction and *yang* tone breathy voice is absent.

Whereas the phonetic voicing contrast is undoubtedly present in non-initial, intervocalic position, it is generally neutralized in word-initial position but may also occasionally appear in this position. We would normally interpret the occasional presence of word-initial voiced onsets as a remnant of an ancient stage of the language when the voicing contrast was still alive (presumably before the tone split). However, instead of the gradual loss which the "remnant" hypothesis suggests, we observe a gradual revival of the word-initial voicing contrast as far as fricatives are concerned, since it is more often produced by young than elderly speakers. Furthermore, that the *yang* labial fricative is much more often produced as voiced than the *yang* dental fricative is not easily explained by the "remnant" hypothesis. An alternative, tentative explanation may be proposed, which is specific to labial fricative onsets. As explained in §4.1.6.2.4, /v/ in Shanghai Chinese has two main origins in Middle Chinese: (1) *bɦ,

(which became *fɦ), as in the syllable 饭 /vɛ/ (T3) (< MC *bjonH)[25]; and (2) *m/w, as in the syllable 万 /vɛ/ (T3) (< MC *mjonH). In a recent past, the two kinds of /v/ may have formed minimal pair syllables in the word-initial context, differing solely in voicing of their onset. The /v/s evolved from MC *m/w were probably pronounced as voiced and the other /v/s evolved from MC *bɦ were probably not, as the other previously voiced obstruent onsets that lost their voicing after the tone split, in word-initial position only. Instead of merging the two labial fricatives in all tones, and producing both of them as voiceless, speakers may tend to merge them only in the *yang* tone and to produce both of them as voiced in word-initial position, possibly due to their common underlying phonological voicing, which surfaces phonetically in word-medial position. To put it slightly differently, the voiced /v/s evolved from *m/w might contaminate the voiceless /v/s evolved from *bɦ, a process encouraged by the homophony of the two fricatives in word-medial position. It also seems, looking at the young speakers' productions, that the phonetic voicing in word-initial labial fricatives is in the process of being generalized, by analogy, to dental fricatives. One problem with this account is that the timing of the change remains vague. What is assumed is that "in a recent past" Shanghai speakers (perhaps from the generation of our elderly speakers, or some even older generation) contrasted, for example, 饭 (/v/< *bɦ) and 万 (/v/<*m/w) as [v]-[f], and that younger generations of Shanghai speakers began to merge the two /v/s into [v]. This revival of phonetic voicing would thus be newly introduced from a different source. However, we cannot fully test this speculative scenario since the /v/s in the /v/-initial syllables we used in the production experiments all had the MC *bɦ origin. What remains to be shown is that the /v/s evolved from MC *m/w were realized [v] by Shanghai speakers from some older generations.

In our view, the onset duration pattern (*yin* longer than *yang*), is mainly related to phonological voicing. Indeed, although the phonetic voicing is not realized in certain contexts, speakers still maintain the timing pattern of the phonological voicing contrast. The rime duration pattern (*yin* shorter than *yang*) is attributable to a

[25] The Middle Chinese (MC) reconstructed forms come from Baxter and Sagart (2011): the final H stands for the MC rising tone (上声).

presumably phonetic effect related to modal vs. breathy phonation in unchecked syllables. In checked syllables, this phonetic effect may be reinforced by a difference in laryngeal articulation constraint: the shift toward the constricted larynx required for final laryngealization is more demanding, hence takes more time, when the initial setting of the vocal folds is slack (*yang* onsets) than when it is tense (*yin* onsets).

We interpret the intervocalic F0 depressor effect as a coarticulatory effect induced by the intervocalic phonetic voicing, which is universally attested. However, the ballistic F0 trajectory effect we propose, which substantially modulates the depressor effect, is specific to the prosodic structure of Shanghai Chinese. In Shanghai Chinese, the prosodic word is a prominent accentual unit, with tight internal cohesion. F0 contours are thus likely programmed at the prosodic word level, independent of the segmental timing at the syllable level. This is also why we observe spirantized realizations of voiced stops in word-medial position. Indeed, lenition is a common process inside a prosodic constituent, but is very rare at its edges (Kingston, 2008). Lenition is however much less often produced by young than elderly speakers. We surmise that young speakers give more weight to the syllable unit than to the prosodic word unit, especially perhaps in a reading task. Elderly speakers' productions are driven by deeply imprinted lexical representations of frequent words. That is, they give more weight to the prosodic word than syllable unit. For young speakers, the reading task might have been similar to school work, in which a careful reading of each character, that is, each syllable is required.

The redundant features resulting from coarticulatory effects will probably be present as long as the conditioning distinctive feature is maintained and no higher level structure, morphological, prosodic, or syntactic interferes with that feature.

The redundant features that are remnants of a diachronic process may tend to fall in disuse. But their loss may take a very long time, especially in some or all of the following cases: (1) the feature is compatible with a coarticulatory effect (As a corollary, features that are *not* compatible with a coarticulatory effect would rapidly disappear; for example, *yang* tone breathiness, which is compatible with low F0 and not high F0, is lost (or at least does not occur) when *yang* tone is realized with a high F0 by virtue of the left-dominant sandhi rule.); (2) the feature has a perceptual role (Breathiness biases identification toward *yang* identity: yet, its role in perception is

probably insufficient to prevent its loss, which seems to be in progress.); (3) although the feature is redundant in some contexts, it is contrastive in others (This is the case of the phonetic voicing, which is contrastive in word-medial position and not in word-initial position.)

6.3 A panchronic account of transphonologization

We asked the following question:

(3) How do features, distinctive or redundant, evolve in modern Shanghai Chinese?

Transphonologization from an initial voicing contrast into a tone register contrast, as a diachronic process already completed or in progress, has been attested in many languages in which low tone is synchronically associated to breathy voice, suggesting that during this transphonologization process, there was an intermediate stage in which breathy vs. modal phonation was distinctive (see §2.2.3). It is a classical example of Panchronic Phonology. Panchronic phonology is a universalist and inductive approach, aiming at discovering general laws in sound change, independently of the language (Hagège & Haudricourt, 1978; Mazaudon & Michailovsky, 2007; Michaud, Jacques, & Rankin, 2012). It analyses synchronic facts observed in different languages, related or unrelated, at different times and locations, and explains these facts by general laws of diachronic evolution.

More specifically, Mazaudon (2010: 169; 171) proposed three panchronic laws of sound change concerning transphonologization, two of which are of interest for us :

Law 1: the phonologization of phonetic material follows the order: voice > breathy > phonologized F0;

Law 3: when breathiness is not, or is no more, a separate feature orthogonal to a tone system, but has become a feature of the tone, it does not survive when the tone becomes phonetically high.

Our synchronic analyses of Shanghai Chinese data from two generations show an ongoing trend towards the loss of breathiness in *yang* tone syllables, suggesting that breathy voice has indeed appeared at an earlier stage than *yang* tone (Law 1).

What is specific to Shanghai Chinese, or even to most Northern Wu dialects, is its retention of the ancient voicing contrast at the time of the disappearance of the breathiness. As we explained earlier, these Wu dialects share a common prosodic structure in which the prosodic word has a heavy-light pattern and the non-initial light syllables have their tone neutralized. In the tone neutralization context, the voicing contrast needs to be maintained to distinguish underlying *yin* and *yang* tone syllables. But since *yang* tone syllables can be realized with a high pitch in this context, breathiness is lost (or at least is absent) (Law 3). Therefore, the prosodic structure common to most Wu dialects might have rendered the voicing contrast (the original contrast) and the tone register contrast (the most recent contrast) equally important, and the phonation difference less important, leading to its predictable loss.

6.4 More on cross-gender variations: internal or external factors?

Our hypothesis that loss of the breathy voice redundant feature would be in progress in the young generation of Shanghai speakers but not yet completed in the older generation was largely borne out among male speakers. We did not make any hypothesis on possible cross-gender variations but, somewhat surprisingly, our results showed a sizeable difference between male and female speakers' production.

If we look at the spectral tilt data (e.g., H1–H2) in word-initial syllables (monosyllabic and S1 contexts), we find that the *yin–yang* difference in breathy voice is the largest for elderly male speakers, followed by elderly female, themselves followed by young speakers hardly distinguished between male and female. This ordering is confirmed, and summarized, by the LDA outcome shown in the confusion matrices (Chapter 4, Table 29). Two differences are of interest here: that between elderly and young speakers, and that between men and women within elderly speakers. Young speakers tend to lose breathy voice compared to elderly speakers, and elderly women seem to be on the same path. These trends are much more dramatic in the open quotient data.

If we only consider the open quotient data as an indicator of phonation type, there is a clear interaction between age and gender: elderly male speakers have a consistent behavior, with systematic production of breathy *yang* tone; young female speakers also have a consistent behavior but, contrary to elderly male speakers, hardly produce breathy voice at all with *yang* tone; in between, both elderly female and young male speakers produce *yang* tone breathiness, and with greater variability. To account for the OQ data, we proposed that there is a tradeoff between F0 and phonation information, with the lack of breathy voice in *yang* syllables compensated by more distinctiveness in tone contours to differentiate *yin* and *yang* tones. Conversely, speakers with a small F0 range might tend to put more weight on breathy voice than on tone contours to signal *yang* syllables. Partly due to biological reasons, male speakers have smaller F0 range than female speakers. Thus, the greater weight on breathy voice than tone contour distinctiveness might reflect an internal factor that distinguishes men and women.

Another factor is more external. As we explained earlier, breathy voice is a conservative, redundant feature. Both young and elderly male speakers produce more systematically breathy voice than female speakers of their age. Male speakers thus have a more conservative behavior than female speakers. Our data thus suggest that female speakers started to lose the breathy feature before male speakers.

This pattern reminds us of William Labov's (2001) analyses of social factors in linguistic change, particularly the gender factor. (Note that social factors are also taken into account in the panchronic approach.) In a great majority of the studied cases, women play the leading role in sound change, often a full generation ahead of men. This is consistent with our data. But why do female speakers innovate more than male speakers? Labov (2001: 274ff) distinguishes two forms of change: (1) "change from above," which is caused by the adoption of well-known and well described linguistic variables, accomplished at a high level of social awareness, such as the fronting of the back /ɑ/ vowel in words ending in *-action* in Quebec French, which is an adoption of the norm of Continental Standard French (Kemp & Yaeger-Dror, 1991); (2) "change from below," which is the primary form of linguistic change within the system, accomplished below the level of social awareness, such as the raising of non-high vowels in English of New York City in the past decades (Labov,

1966). In the first form of change, women, more often than men, tend to adopt more prestigious and, as a corollary, to avoid stigmatized linguistic features, from outside the speech community, or from the upper classes inside the speech community. In the second form of change, women tend to use more innovative forms than men do, probably due to their desire for novelty. These two forms of change yield what Labov called the "Gender paradox," a paradox with respect to conformity:

"Women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not." (Labov, 2001: 293).

What about the loss of breathiness in Shanghai Chinese? At first sight, it corresponds to neither of the two forms of change. It is not a sociolinguistic norm overtly prescribed and Shanghai naïve speakers are probably not aware of the breathy phonation related to the *yang* tone, not even of the *yin–yang* tone difference, not to mention a sociolinguistic connotation of the usage of breathy voice. Moreover, the "changes from above" proposed by Labov are assumed to remain within a single language with prestigious vs. stigmatized varieties. Here, if we stick strictly to Labov's classification and only consider Shanghai Chinese varieties, or even more broadly, Wu dialect varieties, breathy voice phonation is certainly not stigmatized. As for the "changes from below," speakers more likely innovate, in our view, by using more marked variants of pronunciations than the standard, frequent variants. With respect to phonation, breathy phonation is, by definition, more marked than modal phonation, hence its loss can hardly be viewed as an innovation. The case of breathy voice disappearance in Shanghai Chinese is naturally reminiscent of a case going in the opposite direction: the increasing use of creaky voice, a strongly marked phonation type, by young American females (e.g., Yuasa, 2010). This is a clear illustration of a "change from below" in Labov's classification. In contrast, it would be difficult to argue that the loss of breathiness, a non-modal phonation feature, is an innovative change.

If we expand the domain of potentially influential languages to other languages than Shanghai Chinese varieties or, more broadly, Wu dialect varieties, the picture is different and the case of breathy voice disappearance in Shanghai Chinese may fall in the "change from above" category. An obvious influential language is indeed Standard Chinese.

In the modern Chinese society, under the political promotion of Standard Chinese to the disadvantage of local dialects, the prestige is represented by Standard Chinese. But this needs some qualification. What is specific to the situation in a non-Mandarin-speaking Chinese region is that, Standard Chinese and local dialects are not simply co-existing as language varieties of different levels of prestige. Rather, the situation is closer to that of language contact, although the languages (so called "dialects") in contact have many common features and are unified by the same writing system. Shanghai speakers are constantly adopting words from Standard Chinese so that many local lexical forms are disappearing (Qian, 2003: 142ff). The adoption of Standard Chinese forms also occurs at the phonological and the syntactical levels.

The adoption of Standard Chinese forms is not limited to a number of overtly prescribed sociolinguistic norms, but extends to the integration of the entire system of the prestigious Standard Chinese, so that Shanghai speakers have the general feeling of a global impact of Standard Chinese on Shanghai Chinese, and young speakers judge themselves speaking a "degraded" dialect, although they are not aware of which specific linguistic variables are concerned. This is reflected by an overall lower self-evaluation of linguistic competence in Shanghai than Standard Chinese by Shanghai young speakers, which is presumably highly related to their frequent usage of Standard Chinese. Shanghai women, as women from other cultures, are more conventional than men regarding prestigious linguistic forms. Among the 12 young speakers we tested in Experiment 1, five out of six female speakers judged they were less competent in Shanghai than Standard Chinese (by 0.5 point on a 1-5 scale); only one male speaker out of six made the same judgment, while four others rated themselves as equally competent in both languages, and the sixth one rated himself as more competent in Shanghai than Standard Chinese (by 0.5 point). A similar pattern obtained for the self-ratings on the frequency of language use. The same five female speakers reported using more frequently Standard than Shanghai Chinese (by 1 to 2 points on a 1-5 scale); two of the six male speakers gave the same ratings, two others reported an opposite advantage for Shanghai Chinese and the last two male speakers reported they used the two languages with the same frequency. To summarize, Shanghai female speakers were less confident than male speakers in their competence in Shanghai Chinese compared to the assumedly more prestigious Standard Chinese.

Whether or not these female speakers were truly less familiar with and less competent in Shanghai than Standard Chinese remains to be proven, although we rather believe their judgments were biased by an underestimation of their competence in their native language and an overestimation of their competence in Standard Chinese, regarded as more prestigious. At any rate, Shanghai female speakers seem more willing than male speakers to accept, perhaps non-consciously, the prestigious Standard Chinese linguistic system, in which tone is contrastive but phonation type does not play any linguistic role. This would fit nicely with the strategy we found some of them used for distinguishing *yin* and *yang* tones, whereby all the weight is given to F0 contour and voice quality is ignored.

As we can see, the loss of a marked phonation type can take a very long time (nearly thousand years in the case of the breathy voice in Shanghai Chinese), but can be accelerated by external factors such as language contact and sociolinguistic prestige.

6.5 Concluding remarks and perspectives

This dissertation provides an account of the interdependence between segments, tones and phonation types in Shanghai Chinese, with transversal data comparing young (between 20 and 30 years) and elderly (between 60 and 80 years) speakers. We observe the ongoing disappearance of a redundant feature associated with the *yang*, low tone register: breathy voice. This might be, for the time being, specific to Shanghai Chinese and future work is needed to examine whether it might generalize to other Wu dialects.

At a more general level, we provide evidence for the panchronic law according to which transphonologization from a voicing contrast to a tone register contrast first goes through an association between voiced onsets and breathy phonation (e.g., Mazaudon, 1992).

We now would like to address a more theoretical issue: the modeling of transphonologization. In monosyllabic languages, a transphonologization process is often completed through the initial spreading of a feature from one segment to another within the syllable unit so that distinctiveness is maintained. (Note that the

initial spreading of a feature may be followed by its dissociation from the source segment or the disappearance of the source segment.) For example, a consonantal feature may be replaced by a vowel feature, or vice versa. This may occur because consonants and vowels within a syllable are tightly bound by coarticulation (e.g., Michaud, 2012).

In the case of the evolution from a voicing contrast to a phonation difference, and later to a tone register contrast, as in Chinese languages, or to a vowel quality contrast, as in Mon-Khmer languages (see §2.2.3.2), the common sound change process begins with the spreading of a feature from the initial consonant to the following vowel. It is, however, always a challenge to specify and describe the feature that was transferred, especially in the case of a transfer from the suprasegmental the segmental level or vice versa. Unlike sound changes due to an assimilation process, such as nasality spreading (e.g., French: [ɑ̃] < [ɑ̃n] < [an]) in which the [nasal] feature is sufficient to describe the sound change, the situation is more complex when tone or phonation type is involved. For example, it seems somewhat awkward to describe, using distinctive feature representations, the sound change resulting from the transfer of the [+voice] feature on the initial consonant to, say, a L feature on the tone, or to a [-tense] feature on the vowel?

Another challenge is posed by extreme cases of feature migration, in which a feature not only is spread from one segment to another, but several features become shared by or exchanged between different segments within the syllable. In such cases, it may be difficult to figure out which feature is distinctive and which is redundant. Eugénie J. Henderson (1985) noticed many such examples in monosyllabic languages. Forms in free variation, or in different dialects, or at different stages of a dialect's evolution, may share the same set of features at the syllable level but not at the phoneme level. [Ru]~[wi] in Bwe Karen ([R] for a back unrounded glide); [oŋ]>[ɤŋ^{m}] in Vietnamese illustrate such cases: In these examples, the [+round] feature is realized on the vowel in one form, and on the consonant in the other. Henderson (1985) called these alternations "feature shuffling." Henderson proposed an interesting concept to explain "feature shuffling": the phasing relationship between articulatory gestures. Interestingly, at about the same time, a new phonological approach emerged: articulatory phonology (Browman & Goldstein, 1986, 1992). In articulatory phonology,

the notions of "gesture," as representational primitives, and of "phasing relationship," as a means to describe their interactions, are central. Contrary to features that are static and linked to entire segments, gestures are dynamic and may overlap to different extents.

The notion of phasing relationship may be crucial in our Shanghai data, as is suggested by the maintenance of intrinsic timing relationships between segments in phonological voicing contrasts in which phonetic voicing has been lost (see §4.1.6.4). A phasing relationship account also seems quite straightforward to explain the timing between the segments of a disyllabic prosodic word and the F0 contour planned to apply to that prosodic word (see §4.1.6.1). The set of laryngeal and supralaryngeal gestures required are timed so that they simply start and end "in phase," without further specification (see Man Gao's, 2008, dissertation for an articulatory phonology modeling of Mandarin tones in terms of phasing relationships between consonant, vowel, and tone tier gestures.)

We see in the articulatory phonology approach a promising capacity to model sound changes such as segmental to suprasegmental transphonologization. This approach might indeed be able to account for the sound changes whereby what is modified is the assignment to segments and tones of qualitatively different features. Instead of postulating e.g. the artificial interchange between features of intrinsically different nature, the articulatory phonology approach would explain the segmental to suprasegmental changes just in the same way as it explains segmental changes: by changes in phasing relationships and in degree of overlap between consonant, vowel, and tone tier gestures. What would be new here is the integration in the articulatory phonology modeling of a gestural tier accounting for tones (see M. Gao, 2008) and, more broadly, laryngeal settings that are not easily captured by segment-specific features.

REFERENCES

Abdelli-Beruh, N. B. (2004). The stop voicing contrast in French sentences: Contextual sensitivity of vowel duration, closure duration, voice onset time, stop release and closure voicing. *Phonetica*, *61*(4), 201-219.

Anastaplo, S., & Karnell, M. P. (1988). Synchronized videostroboscopic and electroglottographic examination of glottal opening. *The Journal of the Acoustical Society of America*, *83*(5), 1883-1890.

Andruski, J. E., & Ratliff, M. (2000). Phonation types in production of phonological tone: The case of Green Mong. *Journal of the International Phonetic Association*, *30*(1-2), 37-61.

Anonymous. (1960). **江苏省和上海市方言概况** [*A survey of the dialects of Jiangsu and Shanghai*]. Nanjing: Jiangsu Renmin Chubanshe.

Ballard, W. L. (1989). Progress in tone sandhi analysis. In D. Bradley, E. Henderson, M. Mazaudon (Eds.), *Prosodic analysis and Asian linguistics : to honour R. K. Sprigg* (pp. 95-108). The Australian National Univeristy.

Bao, Z. 包智明 (1990). Fanqie languages and reduplication. *Linguistic Inquiry*, *21*, 317-350.

Bao, Z. 包智明 (1996). The syllable in Chinese. *Journal of Chinese Linguistics*, *24*(2), 312-353.

Bahl, K. C. (1957). Tones in Punjabi. *Indian Linguistics 17*, 139-147.

Bates, D., Maechler, M., & Dai, B. (2008). *lme4: Linear mixed-effects models using Eigen and S4.* Version 1.1-7, http://CRAN.R-project.org/package=lme4.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology*. Trends in Linguistics: Studies and Monographs 102. Berlin: Mouton de Gruyter.

Baxter, W. H., & Sagart, L. (2011). Baxter-Sagart Old Chinese Reconstruction (Version 1.00).

Baxter, W. H., & Sagart, L. (2014). *Old Chinese: A new reconstruction*. Oxford: Oxford University Press.

Belotel-Grenié, A., & Grenié, M. (1994). Phonation types analysis in Standard Chinese. *Proceedings of International Conference on Spoken Language Processing*, 343-346. Yokohama, Japan.

Belotel-Grenié, A., & Grenié, M. (1995). Consonants and vowels influence on phonation types in isolated words in Standard Chinese. *Proceedings of the 13th International Congress of Phonetic Sciences*, 400-403. Stockholm, Sweden.

Benguerel, A. P., & Bhatia, T. K. (1980). Hindi stop consonants: an acoustic and fiberscopic study. *Phonetica*, *37*(3), 134-148.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of phonetics*, *20*(3), 305-330.

Bickley, C. (1982). Acoustic analysis and perception of breathy vowels. *Speech communication group working papers*, *1*, 71-81.

Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011). Synthesis of breathy, normal, and phonation using a two-mass modal with a triangular glottis. *Proceedings of the Interspeech 2011*, 2681-2684, Florence, Italy.

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics, 30*(2), 163-191.

Boersma, P. & Weenink, D. (1992-2014). Praat: doing phonetics by computer [Computer program]. Version 5.4.01. http://www.praat.org/

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology, 3*(01), 219-252.

Browman, C. P., & Goldstein, L. M. (1992). Articulatory phonology: An overview.*Haskins Laboratories Status Report on Speech Research, SR-111, 112*, 23-42.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language.* Cambridge: Cambridge University.

Brown, W. S., Morris, R. J., Hollien, H., & Howell, E. (1991). Speaking fundamental frequency characteristics as a function of age and professional singing. *Journal of voice, 5*(4), 310-315.

Brunelle, M. (2005). *Register in Eastern Cham: Phonological, Phonetic and Sociolinguistic approaches*. Ph.D dissertation, Cornell University.

Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics, 37*(1), 79-96.

Brunelle, M., & Finkeldey, J. (2011). Tone perception in Sgaw Karen. *Proceedings of the 17th International Congress of Phonetic Sciences*, 372-375. Hong Kong, China.

Cao, J. 曹剑芬 (1987). The ancient initial "voiced" consonants in modern Wu dialects. *Proceedings of the 11th International Congress of Phonetic Sciences*, Vol. 4, 169-172. Tallinn, U.S.S.R.

Cao, J. 曹剑芬, & Maddieson, I. (1992). An exploration of phonation types in Wu dialects of Chinese. *Journal of Phonetics, 20*, 77-92.

Cao, Z. 曹志耘 (2002). **南部吴语语音研究** [*Research on Southern Wu phonetics*]. Beijing: Shangwu Yin Shuguan.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Edinburgh: Edinburgh University Press.

Chatterji, S. K. (1940). *Indo-Aryan and Hindu*. Ahmedabad.

Chambers, J. K., & Trudgill, P. (1990). *Dialectology* (4th ed.). Cambridge University Press.

Chao, Y. R. 赵元任 (1928). *Studies in the modern Wu dialects*. Monograph No.4. Peking: Tsinghua College Research Institute.

Chao, Y. R. 赵元任 (1930). A system of tone-letters. *Le Maître Phonétique 30*, 24-27.

Chao, Y. R. 赵元任 (1934). 音位标音法的多能性 [The non-uniqueness of phonemic solutions of phonetic systems]. *Bulletin of the National Research Institute of History and Philology, Academia Sinica 4*, 363-397.

Chao, Y. R. 赵元任 (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, M. Y. (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge: Cambridge University Press.

Chen, T. Y., & Tucker, B. V. (2013). Sonorant Onset Pitch as a Perceptual Cue of Lexical Tones in Mandarin. *Phonetica, 70*(3), 207-239.

Chen, Y. 陈轶亚 (2008a). The acoustic realization of vowels of Shanghai Chinese. *Journal of Phonetics*, *36*(4), 629-648.

Chen, Y. 陈轶亚 (2008b). Revisiting the phonetics and phonology of Shanghai tone sandhi. *Proceedings of the Fourth Conference on Speech Prosody*, 253-256, Campinas, Brazil.

Chen, Y. 陈轶亚 (2011). How does phonology guide phonetics in segment–*f0* interaction?. *Journal of Phonetics*, *39*(4), 612-625.

Chen, Y. 陈轶亚 & Downing, L. J. (2011). All depressors are not alike: a comparison of Shanghai Chinese and Zulu. In S. Frota & G. Elordieta (Eds.), *Prosodic Categories: Production, Perception and Comprehension* (pp. 243-265). Springer Netherlands.

Chen, Z. 陈忠敏 (2003). *Studies on dialects in the Shanghai area: Their phonological systems and historical developments*. Lincom Europa.

Childers, D. G., Naik, J. M., Larar, J. N., Krishnamurthy, A. K., & Moore, G. P. (1983). Electroglottography, speech and ultra-high speed cinematography. In I. Titze & R. Scherer, (Eds.), *Vocal fold physiology: biomechanics, acoustics and phonatory control* (pp. 202-220), Denver, CO: Denver Center for the Performing Arts.

Childers, D. G., Hicks, D. M., Moore, G. P., Eskenazi, L., & Lalwani, A. L. (1990). Electroglottography and vocal fold physiology. *Journal of Speech, Language, and Hearing Research*, *33*(2), 245-254.

Cho, H., & Giavazzi, M. (2009). Perception of voicing in fricatives. *Proceedings of the 18th International Congress of Linguists (CIL XVIII).* Seoul, South Korea.

Chung, H. (2002). Segment duration in spoken Korean. *Proceedings of the 7th International Conference on Spoken Language Processing.* Denver, Colorado, USA.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, *2*, 225-252.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (6th ed.). Maldon: Blackwell Publishing Company.

Davison, D. S. (1991). An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working papers in Phonetics*, *78*, 50-57.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonic-to-noise ratio in speech signals. *Journal of Speech and Hearing Research, 36,* 254-266.

Denning, K. (1989). *The diachronic development of phonological voice quality, with special reference to Dinka and the other Nilotic languages.* Ph.D dissertation, Stanford University.

Downing, L. J. (2009). On pitch lowering not linked to voicing: Nguni and Shona group depressors. *Language Sciences*, *31*(2), 179-198.

Duanmu, S. 端木三 (1988). *Shanghai Tone: Representation and Spreading.* Ms. MIT.

Duanmu, S. 端木三 (1990). *A Formal Study of Syllable, Tone, Stress, and Domain in Chinese Languages.* Ph.D dissertation, MIT.

Edmondson, J. A., & Esling, J. H. (2006). The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. *Phonology*, *23*(02), 157-191.

Edmondson, J. A., & Gregerson, K. J. (1993). Western Cham as a register language. *Oceanic Linguistics Special Publication, 24,* 61-74.

Esling, J. H., & Harris, J. G. (2003). An expanded taxonomy of states of the glottis. *Proceedings of the 15th International Congress of Phonetic Sciences*, Vol. 1, 1049-1052.

Ewan, W. G. (1976). *Laryngeal behavior in speech*. Ph.D dissertation, University of California, Berkeley.

Ewan, W. G., & Krones, R. (1974). Measuring larynx movement using the thyroumbrometer. *Journal of Phonetics, 2*, 327-335.

Fagan, J. L. (1988). Javanese intervocalic stop phonemes: the light/heavy distinction. In R. McGinn (Ed.), *Studies in Austronesian Linguistics* (pp. 173-200). Ohio University Center for International Studies, Center for Southeast Asia Studies, Athens, Ohio.

Ferlus, M. (1979). Formation des registres et mutations consonantiques dans les langues mon-khmer. *Mon-Khmer Studies, 8*, 1-76.

Fischer-Jørgensen, E. (1967). Phonetic analysis of breathy (murmured) vowels in Gujarati. *Indian linguistics*, *28*, 71-139.

Fu, G., Cai, Y., Bao, S., Fang, S., Fu, Z., & Zhengzhang, S. 傅国通，蔡勇飞，鲍士杰，方松熹，傅佐之，郑张尚芳 (1986). 吴语的分区 [The grouping of Wu phonetics]. **方言** [*Dialect*], *8*(1), 1-7.

Fux, F. (2012). *Vers un système indiquant la distance d'un locuteur par transformation de sa voix.* Ph.D dissertation, Université de Grenoble, Grenoble.

Gandour, J. (1977). On the interaction between tone and vowel length: Evidence from Thai dialects. *Phonetica*, *34*(1), 54-65.

Gao, J. 高佳音 (2011). *Etude acoustique des syllabes (C)V en shanghaïen : redondance et complémentarité des caractéristiques tonales et segmentales*. Master's thesis, Université Sorbonne Nouvelle – Paris 3.

Gao, J. 高佳音, Hallé, P., Honda, K., Maeda, S., & Toda, M. (2011). Shanghai slack voice: acoustic and EPGG data. *Proceedings of the 17th International Congress of Phonetic Sciences*, 719-722. Hong Kong, China.

Gao, J. 高佳音, & Hallé, P. (2012). Caractérisation acoustique des obstruantes phonologiquement voisées du dialecte de Shanghai. *Actes de JEP-TALN-RECITAL*, 145-152. Grenoble, France.

Gao, M. (2008). *Tonal alignment in Mandarin Chinese: An Articulatory Phonology account*. Ph.D thesis, Yale University.

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, *41*(02), 185-205.

Georgeton, L., & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics*, *44*, 83-95.

Glover, W. W. (1970). Gurung tone and higher levels. *Occasional papers of the Wolfenden Society on Tibeto-Burman linguistics*, *3*, 52-73.

Goldsmith, J. A. (1976). *Autosegmental phonology*. Indiana University Linguistics Club.

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383-406.

Hagège, C., & Haudricourt, A.-G. (1978). *La Phonologie Panchronique*. Paris: Presses Universitaires de France.

Halle, M., & Stevens, K. N. (1971). A note on laryngeal features. *Quaterly Progress Report*, Research Laboratory of Electronics, MIT, *101*, 198-213.

Hallé, P. (1994). Evidence for tone-specific activity of the sternohyoid muscle in modern standard Chinese. *Language and Speech, 37,* 103-123.

Hallé, P., Best, C. T., & Levitt, A. (1999). Phonetic vs. phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics, 27*(3), 281-306.

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, *106*(2), 1064-1077.

Haudricrout, A.-G. (1954). De l'origine des tons en vietnamien. *Journal Asiatique, 242*, 68-82.

Haudricourt, A.-G. (1961). Bipartition et tripartition des systèmes de tons dans quelques langues d'Extrême-Orient. *Bulletin de la Société de Linguistique de Paris*, *56*(1), 163-80.

Haudricourt, A.-G. (1965). Les mutations consonantiques des occlusives initiales en mon-khmer. *Bulletin de la Société de Linguistique de Paris*, *60*(1), 160-172.

Haudricourt, A.-G. (1972). Tones in Punjabi. *Pakha Sanjam 5*, [xxi-xxii].

Henderson, E. J. (1952). The main features of Cambodian pronunciation. *Bulletin of the School of Oriental and African Studies*, *14*(1), 149-174.

Henderson, E. J. (1985). Feature shuffling in Southeast Asian languages. In *Southeast Asian Linguistic Studies presented to André-G. Haudricourt.* (pp. 1-22). Bangkok: Mahidol University.

Henrich, N. (2001). *Study of the glottal source in speech and singing: Modeling and estimation, acoustic and electroglottographic measurements, perception.* Ph.D dissertation, Université Pierre et Marie Curie – Paris 6. <tel-00123133>

Henrich, N., d'Alessandro, C., Doval, B., & Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America*, *115*(3), 1321-1332.

Henrich, N., Gendrot, C., Michaud, A. (2004). Tools for Electroglottographic Analysis: Software, Documentation and Databases [Web page]. <http://voiceresearch.free.fr/egg/>.

Henton, C. G. (1986). Creak as a sociophonetic marker. *The Journal of the Acoustical Society of America*, *80*(S1), S50-S50.

Henton, C. G. & Bladon, R. A. (1985). Breathiness in normal female speech: Inefficiency versus desirability. *Language and Communication*, *5*, 221-227.

Hess, M. M., & Ludwigs, M. (2000). Strobophotoglottographic transillumination as a method for the analysis of vocal fold vibration patterns. *Journal of Voice*, *14*(2), 255-271.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research, 37*(4), 769-778.

Hirai, H., Honda, K., Fujimoto, I., & Shimada, Y. (1994). Analysis of magnetic resonance images on the physiological mechanisms of fundamental frequency control. *Journal of the Acoustical Society of Japan, 50*, 296-304.

Hirano, M., Vennard, W., & Ohala, J. (1970). Regulation of register, pitch and intensity of voice. *Folia Phoniatrica et Logopaedica*, *22*(1), 1-20.

Hirose, H., & Gay, T. (1972). The activity of the intrinsic laryngeal muscles in voicing control. *Phonetica*, *25*(3), 140-164.

Holmberg, E. B., Hillman, R. E., & Perkell, J. S. (1989). Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch. *Journal of Voice*, *3*(4), 294-305.

Holmberg, E. B., Hillman, R. E., Perkell J. S., Guiod, P. C., & Goldman, S. L. (1995). Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech, Language, and Hearing Research* 38, 1212-1223.

Hombert, J. M. (1975). *Towards a theory of tonogenesis: an empirical, physiologically and perceptually based account of the development of tonal contrasts in languages.* Ph.D dissertation, University of California, Berkeley.

Hombert, J. M. (1977). Difficulty of producing different F0 in speech. *UCLA Working Papers in Phonetics*, *36*, 12-19.

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: from global to local prosody. *ISCA International Conference on Speech Prosody 2004*, Nara, Japan.

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech*, *42*(4), 401-411.

Howie, J. M. (1974). On the domain of tone in Mandarin. *Phonetica*, *30*(3), 129-148.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press.

Hyman, L. M. (2008). Enlarging the scope of phonologization. *University of California Berkeley Phonology Lab Annual Report*, 382-409.

Hyman, L. M., & Schuh, R. G. (1974). Universals of tone rules: evidence from West Africa. *Linguistic Inquiry*, *5*(1), 81-115.

Iseli, M., & Alwan, A. (2004). An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. *Proceedings of ICASSP*, vol. 1, 669-672. Montreal, Canada.

Iseli, M., Shue, Y. L., Epstein, M. A., Keating, P. A., Kreiman, J., & Alwan, A. (2006). Voice source correlates of prosodic features in American English: a pilot study. *Proceedings of the Interspeech 2006*, Pittsburgh.

Ishizaka, K., & Flanagan, J. L. (1972). Synthesis of Voiced Sounds From a Two‐Mass Model of the Vocal Cords. *The Bell System Technical Journal*, *51*(6), 1233-1268.

Iwata, R., Hirose, H., Niimi, S., & Horiguchi, S. (1991). Physiological properties of "breathy" phonation in a Chinese dialect – a fiberoptic and electromyographic study on Suzhou dialect. *Proceedings of the 12th Conference on Phonetic Sciences*, Vol. 3, 162-165, Aix-en Provence, France.

Jakobson, R. (1931 [1972]). Principles of historical phonology. In A. R. Keiler (Ed. & tr.), *A reader in historical and comparative linguistics* (pp. 121-138), New York: Holt, Rinehart & Winston.

Jessen, M., & Roux, J. C. (2002). Voice quality differences associated with stops and clicks in Xhosa. *Journal of Phonetics*, *30*(1), 1-52.

Kagaya, R., & Hirose, H. (1975). Fiberoptic electromyographic and acoustic analyses of Hindi stop consonants. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, *9*, 27-46.

Kao, D. L. (1971). *Structure of the syllable in Cantonese*. The Hague: Mouton.

Karlgren, B. (1915–1926). Études sur la phonologie chinoise. *Archives d'études orientales, 15*. Upplasa: K. W. Appelberg; Leiden: E. J. Brill.

Kemp, W., & Yaeger-Dror, M. (1991). Changing realizations of A in (a)tion in relation to the front A-Back A opposition in Quebec French. *New ways of analyzing sound change*, 127-84.

Khan, S. U. D. (2012). The phonetics of contrastive phonation in Gujarati. *Journal of Phonetics, 40*(6), 780-795.

Kingston, J. (2005). The phonetics of Athabaskan tonogenesis. In S. Hargus & K. Rice (Eds.) *Athabaskan prosody* (pp. 137-184). Amsterdam & Philadelphia: John Benjamins.

Kingston, J. (2008). Lenition. In *Selected proceedings of the 3rd conference on laboratory approaches to Spanish phonology* (pp. 1-31). Somerville, MA: Cascadilla Press.

Kingston, J. (2011). Tonogenesis. In M. van Oestendrop, C. J. Ewen, E. Hume & K. Rice (Eds.), (pp. 2304-2333), *The Blackwell companion to phonology,* Blackwell Reference Online.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, 419-454.

Kirk, P. L., Ladefoged, J., & Ladefoged, P. (1993). Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In A. Mattina & T. Montler (Eds.), *American Indian linguistics and ethnography in honor of Laurence C. Thompson*, (pp. 435-450). Missoula, MT: University of Montana Press.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820-857.

Kreiman, J., Iseli, M., Neubauer, J., Shue, Y. L., Gerratt, B. R., & Alwan, A. (2008). The relationship between open quotient and H1*–H2*. *The Journal of the Acoustical Society of America*, *124*(4), 2495-2495.

Kohler, K. J. (1984). Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica, 41*(3), 150-174.

Koreman, J. (1996). *Decoding linguistic information in the glottal airflow*. Ph.D dissertation, University of Nijmegen.

Kuang, J. (2012). Registers in tonal contrasts. *UCLA Working Papers in Phonetics, 110,* 46-64.

Kuang, J. (2013a). The Tonal Space of Contrastive Five Level Tones. *Phonetica, 70*(1-2), 1-23.

Kuang, J. (2013b). *Phonation in Tonal Contrasts*. Ph.D dissertation, UCLA.

Labov, W. (1966). *The social stratification of English in New York City.* Washington, D.C.: Center for Applied Linguistics.

Labov, W. (2001). *Principles of linguistic change, Volume 2: Social factors.* Oxford: Blackwell Publishers.

Ladefoged, P. (1967). *Three areas of experimental phonetics*. London: Oxford University Press.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago.

Ladefoged, P. (1975). *A course in phonetics.* New York: Harcourt.

Ladefoged, P. (1983). The linguistic use of different phonation types. *Vocal fold physiology: Contemporary research and clinical issues,* 351-360.

Ladefoged, P., & Antoñanzas-Barroso, N. (1985). Computer measures of breathy voice quality. *UCLA Working Papers in Phonetics, 61*, 79-86.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages,* Blackwell publishers.

Laver, J. (1980). *The phonetic description of voice quality.* Cambridge University Press.

Li, F. (2009). *Parlons shanghaïen.* Paris: l'Harmattan.

Li, F.-K. 李方桂 (1937). Languages and dialects. *The Chinese yearbook.* Reprinted with revisions in the *Journal of Chinese Lingustics* (1973), *1*, 1-13.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, *33*(1), 42-49.

Liu, F. 刘复 (1923). 守温三十六字母排列法之研究 [Study on the 36 initials of Shouwen]. **国学季刊，** *1*(3), 451-464.

Liu, F. 刘复 (1925). Étude expérimentale sur les tons du chinois. *Collection de l'Institut de Phonétique des Archives de la Parole, Fascicule 1*. Paris: Société d'édition.

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *The journal of the acoustical society of America*, *85*(3), 1314-1321.

Macchi, M., Altom, M. J., Kahn, D., Singhal, S., & Spiegel, M. F. (1993). Intelligibility as a function of speech coding method for template-based speech synthesis. *Proceedings of the 3rd European Conference on Speech Communication and Technology* (Eurospeech), 893-896, Berlin, Germany.

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge University Press, Cambridge.

Maddieson, I. (1996). Phonetic universals. *UCLA Working Papers in Phonetics*, 160-178.

Maddieson, I., & Ladefoged, P. (1985). 'Tense' and 'Lax' in four minority languages of China. *UCLA Working papers in Phonetics*, *60*, 59-83.

Maddieson I., & Precoda K. (1990). Updating UPSID. *UCLA Working Papers in Phonetics*, 104-111.

Maspero, H. (1912). Etudes sur la phonétique historique de la langue annamite. Les initiales. *Bulletin de l'École française d'Extrême-Orient*, *12*(1), 1-124.

Mazaudon, M. (1977). Tibeto-Burman tonogenetics. *Linguistic of the Tibeto-Burman Area, 3*, 1-123. In M. M. J. Fernandez-Vest (Ed.), *Combats pour les langues du monde: Hommage à Claude Hagège* (pp. 351–362), Paris: L'Harmattan.

Mazaudon, M. (2012). Path to tone in the Tamang branch of Tibeto-Burman (Nepal). In G. de Vogelaer & G. Seiler (Eds.), *The Dialect Laboratory: Dialects As a Testing Ground for Theories of Language Change*, (pp. 139-177), Amsterdam & Philadelphia: John Benjamins.

Mazaudon, M., & Michailovsky, B. (2007). La phonologie panchronique aujourd'hui: Quelques repères.

Mazaudon, M., & Michaud, A. (2008). Tonal contrasts and initial consonants: a case study of Tamang, a 'missing link' in tonogenesis. *Phonetica*, *65*(4), 231-256.

Meynadier, Y., & Gaydina, Y. (2012). Contraste de voisement en parole chuchotée. *Actes de JEP-TALN-RECITAL*, 361-368. Grenoble, France.

Michaud, A. (2005). *Prosodie de langues à tons (naxi et vietnamien), prosodie de l'anglais: éclairages croisés*. Ph.D dissertation, Université de Paris III–Sorbonne Nouvelle.

Michaud, A. (2012). Monosyllabicization: patterns of evolution in Asian languages. *Monosyllables: from phonology to typology*, 115-130.

Michaud, A., Jacques, G., & Rankin, R. L. (2012). Historical transfer of nasality between consonantal onset and vowel: from C to V or from V to C?. *Diachronica*, *29*(2), 201-230.

Mei T.-L. 梅祖麟 (1970). Tones and prosody in Middle Chinese and the origin of the rising tone. *Harvard Journal of Asiatic Studies, 30,* 86-110.

Mysak, E. D. (1959). Pitch and duration characteristics of older males. *Journal of Speech & Hearing Research, 2,* 46-54.

Moeller, J., & Fischer, J. F. (1904). Observation on the action of the cricothyroideus and thyro-arytenoideus internus. *Annals of Otology, Rhinology, and Laryngology*, *13*, 42-46.

Norman, J. (1988). *Chinese*. Cambridge: Cambridge University Press.

Norman, J. (2003). Phonology of Chinese languages. In G. Thurgood & R.J. LaPolla (eds.), *The Sino-Tibetan Languages*, 72-83, London: Routledge.

Ohala, J. J. (1972). How is pitch lowered?. *The Journal of the Acoustical Society of America*, *52*, 124-124.

Ohala, J. J. (1993a). The phonetics of sound change. In C. Jones (Ed.), *Historical linguistics: Problems and perspectives* (pp. 237-278), London: Longman.

Ohala, J. J. (1993b). Sound change as nature's speech perception experiment. *Speech Communication*, *13*(1), 155-161.

Ohala, J. J. (1997). Aerodynamics of phonology. *Proceedings of the 4th Seoul International Conference on Linguistics*, 92-97.

Omori, K., Kojima, H., Kakani, R., Slavit, D. H., & Blaugrund, S. M. (1997). Acoustic characteristics of rough voice: subharmonics. *Journal of Voice*, *11*(1), 40-47.

O'Shaugnessy, D. (1981). A study of French vowel and consonant durations. *Journal of Phonetics, 9,* 385–406.

Pandit, P. B. (1957). Nasalisation, aspiration and murmur in Gujarati. *Indian Linguistics, 17*, 165-172.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America, 32*(6), 693-703.

Phu, V. H., Edmondson, J. A., & Gregerson, K. J. (1992). East Cham as a tone language. *Mon-Khmer Studies, 20*, 31-43.

Pulleyblank, E. G. (1978). The nature of the Middle Chinese tones and their development to Early Mandarin. *Journal of Chinese Linguistics*, *6*(2), 173-203.

Qian, N. 钱乃荣 (2003). **上海语言发展史** [*A history of the language of Shanghai*]. Shanghai: Shanghai Renmin Chubanshe.

Qian, Z. 钱曾怡 (ed.) (2010). **汉语官话方言研究** [*Study on Mandarin dialects*]. Jinan: Qilu Chubanshe.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsey, S. R. (1989). *The languages of China* (2nd ed.). Princeton University Press.

Ren, N. 任念麒 (1988). A fiberoptic and transillumination study of Shanghai stops. Paper presented at *International Conference on Wu Dialects*, Hong Kong, China.

Ren, N. 任念麒 (1992). *An acoustic study of Shanghai stops.* Ph.D dissertation, University of Connecticut.

Ridouane, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association, 37*, 119-142.

Rose, P. (1982a). Acoustic characteristics of the Shanghai-Zhenhai syllable-types. In D. Bradley (Ed.), *Papers in Southeast Asian Linguistic, No 8: Tonation, (Pacific Linguistics, Series A, No 62),* (pp. 1-53), Canberra: Australian National University.

Rose, P. (1982b). *An Acoustic Based Phonetic Description of the Syllable in the Zhenhai Dialect.* Ph.D dissertation, Cambridge University.

Rose, P. (1988). On the non-equivalence of fundamental frequency and pitch in tonal description. In D. Bradley, E. J. Henderson & M. Mazaudon (Eds.), *Prosodic Analysis and Asian Linguistics: To Honour RK Sprigg*, (pp. 55-82), Pacific linguistics.

Rose, P. (1989). Phonetics and phonology of Yang tone: phonation types in Zhenhai. *Cahiers de linguistique-Asie orientale*, *18*(2), 229-245.

Rose, P. (1993). A linguistic-phonetic acoustic analysis of Shanghai tones. *Australian Journal of Linguistics*, *13*(2), 185-220.

Rose, P. (2002). Independent depressor and register effects in Wu dialect tonology. *Journal of Chinese Linguistics*, *30*(1), 39-81.

Rothenberg, M., & Mahshie, J. J. (1988). Monitoring vocal fold abduction through vocal fold contact area. *Journal of Speech, Language, and Hearing Research*, *31*(3), 338-351.

Sagart, L. (1986). On the Departing Tone. *Journal of Chinese Linguistics*, *14*(1), 90-113.

Sagart, L. (1989). Glottalised tones in China and Southern Asia. In D. Bradley, E. Henderson, M. Mazaudon (Eds.), *Prosodic analysis and Asian linguistics : to honour R. K. Sprigg* (pp. 83-93). The Australian National Univeristy.

Sagart, L. (1998). On distinguishing Hakka and non-Hakka dialects. *Journal of Chinese Linguistics*, *26*(2), 281-302.

Sagart, L. (1999). The origin of Chinese tones. *Proceedings of the Symposium "Cross-Linguistic Studies of Tonal Phenomena: Tonogenesis, Typology and Related Topics"*, 91-104. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies..

Samely, U. (1991). *Kedang (Eastern Indonesia), some aspects of its grammar.* Hamburg: Helmut Buske Verlag.

Selkirk, E., & Shen, T. 沈同 (1990). Prosodic domains in Shanghai Chinese. In S. Inkelas & D. Zec, *The phonology-syntax connection*, (pp. 313-337), Chicago and London: University of Chicago Press.

Shen, Z. W. 沈钟伟, Wooters, C., & Wang, W. Y 王士元 (1987). Closure Duration in the Classification of Stops: A Statistical Analysis. *Ohio State University Working Papers in Linguistics*, *35*, 197-209.

Sherard, M. (1972). *Shanghai phonology*. Ph.D dissertation, Cornell University.

Shi, F. 石锋 (1983). Acoustic features of muddy initials in Suzhou dialect. **语言研究** [*Linguistic study*], *1*, 49-83.

Shue, Y. L., Keating, P., Vicenik, C., & Yu, K. (2011). Voice Sauce: A program for voice analysis. *Proceedings of the 17th International Congress of Phonetic Sciences*, 1846-1849. Hong Kong, China.

Silva, D. J. (1992). *The Phonetics and phonology of stop lenition in Korean.* Ph.D dissertation, Cornell University.

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, *23*(02), 287-308.

Silverman, D. (1995). *Phasing and recoverability*. Ph.D dissertation, UCLA.

Shorto, H. L. (1967). The register distinctions in Mon-Khmer languages. In Wissenschaftliche Zeitschrift der Karl-Marx-Universitat, Leipzig, Gesellschafts-und sprachwissenschaftliche Reich, Heft 1/2, 245-248.

Sundberg, J. (1987). *The Science of Singing Voice.* Northen Illinois University Press.

Sundberg, J., Gu, L., Huang, Q., & Huang, P. (2012). Acoustical study of classical Peking Opera singing. *Journal of Voice*, *26*(2), 137-143.

Svantesson, J. O. (2009). Shanghai vowels. *Lund Working Papers in Linguistics*, *35*, 191-202.

't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, *69*(3), 811-821.

Thongkum, T. (1988). Phonation types in Mon-Khmer languages. In O. Fujimura (Ed.), *Vocal fold physiology*: *voice production*, *mechanisms and functions* (pp. 319-334), New York: Raven Press.

Thurgood, G. (2002). Vietnamese and tonogenesis: Revising the model and the analysis. *Diachronica*, *19*(2), 333-363.

Titze, I. R. (1994). *Principles of voice production.* New Jersey: Prentice Hall.

Tsuji, N. (1977). *Eight Yue dialects in Guangxi Province, China, and reconstruction of proto-Yue phonology.* Ph.D dissertatoin, Cornell University.

Umeda, N. (1977). Consonant duration in American English. *The Journal of the Acoustical Society of America*, *61*(3), 846-858.

Vaissière, J., Honda, K., Amelot, A., Maeda, S., & Crevier-Buchman, L. (2010). Multisensor platform for speech physiology research in a phonetics laboratory. *The Journal of the Phonetic Society of Japan*, *14*(2), 65-78.

Valbret, H., Moulines, E., & Tubach, J. P. (1992). Voice transformation using PSOLA technique. *Speech Communication*, *11*(2), 175-187.

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Wan, I. P. 万依萍 (1997). The status of prenuclear glides in Mandarin Chinese: Evidence from speech errors. *Chicago Linguistics Society*, *33*, 417-428.

Wang, H. S., & Chang, C. 王旭, 张之玲 (2001). On the status of the prenuclear glides in Mandarin Chinese. **語言暨語言學** [*Lanugage and Linguistics*], *2.2*, 243-260.

Wayland, R., & Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics*, *31*(2), 181-201.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (Eds.) *Data Analysis and Decision Support*, 335-343, Springer-Verlag, Berlin.

Whiteside, S. P., Hanson, A., & Cowell, P. E. (2004). Hormones and temporal components of speech: sex differences and effects of menstrual cyclicity on speech. *Neuroscience Letters*, *367*, 44–47.

Whiteside, S. P., & Marshall, J. (2001). Developmental trends in Voice Onset Time: some evidence for sex differences. *Phonetica, 58*,196–210.

Wurm, S. A., Li, R., Baumann, T. & Lee, M. W. (1987). *Language Atlas of China: Parts I and II*. Hong Kong: Longman Group (Far East) Limited.

Xu, B., & Tang, Z. 徐宝华, 汤珍珠 (1988). **上海市区方言志** [*Urban Shanghai dialect records*]. Shanghai: Shanghai Education Publication House.

Yip, M. (1980). *The tonal phonology of Chinese.* Ph.D dissertation, MIT.

Yip, M. (1993). Tonal Register in East Asian Languages. In H. van der Hulst & K. Snider (Eds.) *The Phonology of Tones* (pp. 245-268). Berlin: Mouton de Gruyter.

Yip, M. (2003). Casting doubt on the Onset–Rime distinction. *Lingua*, *113*(8), 779-816.

Yu, K. M., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *The Journal of the Acoustical Society of America*, *136*(3), 1320-1333.

Yuan, J. 袁家骅 (1961). **汉语方言概要** [*An outline of the Chinese dialects*]. Beijing: Wenzi Gaige Chubanshe.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, *85*(3), 315-337.

Zee, E. 徐云扬 (2003). Shanghai phonology. In G. Thurgood & R. J. LaPolla (Eds.), *The Sino-Tibetan Languages* (pp. 131-138), London: Routledge.

Zee, E. 徐云扬, & Maddieson, I. (1979). Tones and tone sandhi in Shanghai: phonetic evidence and phonological analysis. *UCLA Working Papers in Phonetics*, *45*, 93-129.

Zhang, J. 张吉生 (2006). *The Phonology of Shaoxing Chinese*. Ph.D dissertation, Netherlands Graduate School of Linguistics LOT, Utrecht.

Zhengzhang, S. 郑张尚芳 (2000). *The phonological system of Old Chinese* (translated by Laurent Sagart from Chinese to English). Paris: École des Hautes Études en Sciences Sociales, Centre de Recherches Linguistiques sur l'Asie Orientale.

Zhu, X. 朱晓农 (1999). *Shanghai tonetics* (Vol. 32). Lincom Europa.

Zhu, X. 朱晓农 (2006). *A grammar of Shanghai Wu* (Vol. 66). Lincom Europa.

Zhu, X. 朱晓农 (2010). Slack voice: the phonetic nature of the voiced consonants in ancient Chinese. **语言研究** [*Linguistic study*], 30(3), 1-19.

Appendix 1. Glose for speech materials in each experiment.

Experiment 1: monosyllables (Table 12)

T1:	ɛ 哀	‘grief’	pɛ 杯	‘cup’	tɛ 堆	‘stack’			
T2:	ɛ 爱	‘love’	pɛ 板	‘board’	tɛ 胆	‘gallbladder’			
T3:	ɛ 咸	‘salty’	bɛ 办	‘handle’	dɛ 谈	‘talk’			
T4:	aʔ 鸭	‘duck’	paʔ 八	‘eight’	taʔ 搭	‘build’			
T5:	aʔ 盒	‘box’	baʔ 白	‘white’	daʔ 踏	‘tread’			
T1:	fɛ 翻	‘turn over’	sɛ 三	‘three’	mɛ 蛮	‘quite’	nɛ 拿	‘take’	
T2:	fɛ 反	‘reverse’	sɛ 伞	‘umbrella’	mɛ 美	‘beautiful’			
T3:	vɛ 饭	‘rice’	zɛ 才	‘talent’	mɛ 梅	‘plum’	nɛ 难	‘difficult’	
T4:	faʔ 发	‘deliver’	saʔ 杀	‘kill’					
T5:	vaʔ 罚	‘punish’	zaʔ 石	‘stone’	maʔ 麦	‘wheat’	naʔ 纳	‘accept’	

Experiment 1: first syllable as the target syllable (Table 13)

T1:	**pɛ**.tsɻ 杯子	‘cup’	**tɛ**.ɕiŋ 担心	‘worry’
T2:	**pɛ**.tɕɪʔ 背脊	‘back (of the body)’	**tɛ**.tsɻ 胆子	‘courage’
T3:	**bɛ**.koŋ 办公	‘work’	**dɛ**.tsɻ 台子	‘table’
T4:	**paʔ**.paʔ 八百	‘eight hundreds’	**taʔ**.sɛ 搭讪	‘hit on (someone)’
T5:	**baʔ**.pɛ 白板	‘white board’	**daʔ**.zaʔ 踏实	‘steady and sure’
T1:	**fɛ**.ɪʔ 翻译	‘translate’	**sɛ**.ɕi 三鲜	‘shredded sea foods’
T2:	**fɛ**.wɛ 返回	‘return’	**sɛ**.piŋ 伞柄	‘handle of the umbrella’
T3:	**vɛ**.wø 饭碗	‘bowl’	**zɛ**.nəŋ 才能	‘talent’
T4:	**faʔ**.lɪʔ 法律	‘law’	**saʔ**.ts^{h}u 塞车	‘traffic jam’
T5:	**vaʔ**.kø 罚款	‘monetary penalty’	**zaʔ**.dɤ 石头	‘stone’

Experiment 1: second syllable as the target syllable (Table 14)

T1-high:	tsɔ.**pɛ** 早班	‘morning shift’	tɕi.**tɛ** 简单	‘easy’
T1-low:	ku.**pɛ** 科班	‘professional training’	sɻ.**tɛ** 私单	‘private deal’
T2-high:	ɕi.**pɛ** 死板	‘stubborn’	tɕiɔ.**tɛ** 校对	‘proofread’
T2-low:	kø.**pɛ** 干贝	‘dried scallop’	tsɻ.**tɛ** 猪胆	‘pig’s gallbladder (Chinese medicine)’
T3-high:	p^{h}ɛ.**bɛ** 配备	‘equip’	tsɻ.**dɛ** 子弹	‘bullet’
T3-low:	ts^{h}ɔ.**bɛ** 操办	‘manage’	tɕi.**dɛ** 鸡蛋	‘chicken egg’
T4-high:	sɻ.**paʔ** 四百	‘four hundreds’	pɔ.**taʔ** 报答	‘return back’
T4-low:	sɛ.**paʔ** 三百	‘three hundreds’	i.**taʔ** 医德	‘medical ethics’

T5-high:	ɕiɔ.**baʔ** 小白	(a frequent nickname)	tsø.**daʔ** 转达	‘transmit’
T5-low:	kɔ.**baʔ** 茭白	(a Chinese food plant)	ɔ.**daʔ** 凹凸	‘concavity’

T1-high:	tsʰɔ.**fɛ** 吵翻	‘quarrel’	tʰɔ.**sɛ** 套衫	‘sweaters’
T1-low:	sɛ.**fɛ** 三番	‘time and again’	i.**sɛ** 衣衫	‘clothes’
T2-high:	tɕʰi.**fɛ** 遣返	‘repatriate’	tsɤ.**sɛ** 走散	‘lost’
T2-low:	sɛ.**fɛ** 三反	‘three-anti campaigns’	ɕiɔ.**sɛ** 消散	‘disappear’
T3-high:	tsɔ.**vɛ** 早饭	‘breakfast’	tɕiɔ.**zɛ** 教材	‘textbook’
T3-low:	tsʰʐ̩.**vɛ** 糍饭	‘stuffed rice ball’	tʰi.**zɛ** 天才	‘genius’
T4-high:	zu.**faʔ** 做法	‘method’	pɛ.**saʔ** 板刷	‘scrubbing brush’
T4-low:	kɛ.**faʔ** 开发	‘develop’	tsʐ̩.**saʔ** 知识	‘knowledge’
T5-high:	tʰi.**vaʔ** 体罚	‘corporal punishment’	ɕy.**zaʔ** 选择	‘choice’
T5-low:	i.**vaʔ** 衣物	‘clothing’	ɕy.**zaʔ** 虚实	‘actual status’

Experiment 2 (Table 31)

T1:	pi 边	‘border’	pɛ 班	‘class’	pa 巴	‘gullible’		
T2:	pi 比	‘compare’	pɛ 板	‘board’	pa 摆	‘put’		
T3:	bi 皮	‘skin’	bɛ 办	‘handle’	ba 排	‘row’		
T1:	pø 搬	‘move’	pu 波	‘wave’	po 疤	‘scar’	pɔ 包	‘bag’
T2:	pø 半	‘half’	pu 布	‘cloth’	po 把	‘a handful of’	pɔ 宝	‘treasure’
T3:	bø 盘	‘disc’	bu 步	‘step’	bo 爬	‘climb’	bɔ 暴	‘violent’
T1:	pɑŋ 帮	‘help’	pəŋ 奔	‘run’	piŋ 冰	‘ice’		
T2:	pɑŋ 榜	‘published list’	pəŋ 本	‘origin’	piŋ 饼	‘cookies’		
T3:	bɑŋ 碰	‘touch’	bəŋ 笨	‘stupid’	biŋ 病	‘disease’		
T4:	pɪʔ 笔	‘pen’	pʊʔ 剥	‘peel’	paʔ 百	‘hundred’		
T5:	bɪʔ 鼻	‘nose’	bʊʔ 薄	‘thin’	baʔ 白	‘white’		

Experiment 3 (Table 32)

T1:	pi 边	‘border’	pɛ 杯	‘cup’	pu 波	‘wave’
T2:	pi 比	‘compare’	pɛ 板	‘board’	pu 布	‘cloth’
T3:	bi 皮	‘skin’	bɛ 办	‘handle’	bu 步	‘step’
T1:	ti 低	‘low’	tɛ 堆	‘stack’	tu 多	‘many’
T2:	ti 底	‘bottom’	tɛ 胆	‘gallbladder’	tu 堵	‘block’
T3:	di 提	‘lift (v.)’	dɛ 台	‘table’	du 涂	‘paint’
T1:	fi 妃	‘concubine’	fɛ 翻	‘turn over’	fu 夫	‘hunsband’
T2:	fi 费	‘fee’	fɛ 反	‘reverse’	fu 付	‘pay’
T3:	vi 维	‘maintain’	vɛ 烦	‘annoying’	vu 扶	‘help up’

T1:	sɻ 丝	‘silk’	sɛ 三	‘three’	su 苏	(a surnam)
T2:	sɻ 试	‘try’	sɛ 伞	‘umbrella’	su 所	‘place’
T3:	zɻ 是	‘yes’	zɛ 馋	‘greedy’	zu 坐	‘sit’

Experiment 4 (Table 35)

T2:	fɛ 反	‘reverse’	sɛ 伞	‘umbrella’
T3:	vɛ 烦	‘annoying’	zɛ 馋	‘greedy’
T4:	faʔ 发	‘deliver’	saʔ 杀	‘kill’
T5:	vaʔ 罚	‘punish’	zaʔ 十	‘ten’

Experiment 5 (Table 37)

T2:	ɛ 爱	‘love’	pɛ 板	‘board’	tɛ 胆	‘gallbladder’
T3:	ɛ 咸	‘salty’	bɛ 办	‘handle’	dɛ 台	‘table’

T2:	fɛ 反	‘reverse’	sɛ 伞	‘umbrella’	mɛ 美	‘beautiful’
T3:	vɛ 烦	‘annoying’	zɛ 馋	‘greedy’	mɛ 梅	‘plum’

Appendix 2. Detailed acoustic results

1. Detailed stop onsets' VOT in word-initial posotion (§4.1.6.2).

Table 41. Stop onsets' VOT (ms) of /ɛ/ rime in monosyllables according to onset and tone (T1-3).

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial stop	11.8 (3.7)	11.6 (4.2)	12.0 (3.3)	16.3 (2.4)	16.2 (4.8)	19.7 (5.2)
dental stop	12.2 (2.2)	13.0 (3.1)	12.9 (2.5)	17.5 (4.7)	17.9 (5.0)	23.8 (10.6)
Mean	**12.0 (2.9)**	**12.3 (3.6)**	**12.5 (2.9)**	**17.0 (3.6)**	**17.0 (4.8)**	**21.7 (8.2)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial stop	13.4 (4.5)	13.9 (1.3)	16.1 (4.1)	16.5 (4.1)	15.3 (2.9)	19.5 (4.4)
dental stop	11.9 (1.1)	13.4 (3.7)	17.9 (6.8)	15.6 (3.9)	13.8 (2.5)	20.5 (2.4)
Mean	**12.7 (3.2)**	**13.3 (2.8)**	**17.0 (5.5)**	**15.8 (3.9)**	**14.4 (2.3)**	**20.4 (3.5)**

Table 42. Stop onsets' VOT (ms) of /aʔ/ rime in monosyllables according to onset and tone (T4-5).

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
labial stop	11.0 (4.4)	11.3 (4.2)	14.4 (4.3)	17.9 (12.3)
dental stop	10.2 (1.1)	11.5 (1.9)	13.8 (5.4)	16.8 (6.3)
Mean	**10.6 (3.1)**	**11.4 (3.1)**	**14.1 (4.7)**	**17.4 (9.4)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
labial stop	8.9 (1.2)	14.3 (4.4)	12.9 (2.7)	18.3 (2.8)
dental stop	9.3 (3.1)	12.9 (3.8)	9.9 (1.7)	16.5 (4.3)
Mean	**9.1 (2.2)**	**13.6 (4.0)**	**11.0 (2.3)**	**17.8 (3.6)**

Table 43. VOTs of stop onsets in the S1 context for /ɛ/-rime syllables (T1-3), according to place of articulation and tone.

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial stop	12.1 (3.1)	13.1 (3.5)	11.3 (4.3)	18.4 (6.1)	20.8 (6.0)	21.7 (5.7)
dental stop	12.5 (2.0)	11.7 (3.1)	12.2 (1.8)	20.3 (6.7)	21.0 (6.5)	22.7 (5.8)
Mean	**12.3 (2.4)**	**12.4 (3.2)**	**11.8 (3.2)**	**19.3 (6.2)**	**20.9 (5.9)**	**22.1 (5.5)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial stop	10.3 (1.4)	12.3 (1.7)	13.3 (3.0)	13.2 (1.8)	12.7 (4.5)	19.2 (3.3)
dental stop	11.3 (3.7)	13.2 (4.1)	15.5 (4.8)	13.9 (1.3)	16.1 (5.3)	15.6 (4.5)
Mean	**10.8 (2.7)**	**12.8 (3.0)**	**14.4 (4.1)**	**13.4 (1.4)**	**15.6 (4.9)**	**17.1 (4.2)**

Table 44. VOTs of stop onsets in the S1 context for /aʔ/-rime syllables (T4-5), according to place of articulation and tone.

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
labial stop	10.5 (3.5)	12.0 (3.6)	14.3 (4.9)	18.9 (10.2)
dental stop	10.2 (1.4)	12.3 (2.0)	20.3 (6.6)	19.1 (4.6)
Mean	**10.3 (2.6)**	**12.4 (2.8)**	**17.3 (6.3)**	**19.0 (7.6)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
labial stop	11.4 (2.7)	10.1 (2.6)	13.2 (2.5)	15.7 (2.0)
dental stop	11.1 (4.9)	13.4 (2.9)	17.6 (8.8)	17.8 (6.8)
Mean	**11.2 (3.8)**	**12.0 (3.1)**	**15.7 (6.8)**	**16.9 (5.1)**

2. Detailed v-ratios in word-initial fricatives et word-medial obstruents (§4.1.6.2).

Table 45. v-ratios of fricatives onsets in monosyllables with the /ɛ/ rime (T1-3), according to place of articulation and tone.

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial fric	0.10 (0.06)	0.09 (0.04)	0.86 (0.13)	0.05 (0.03)	0.05 (0.03)	0.40 (0.36)
dental fric	0.08 (0.02)	0.09 (0.02)	0.44 (0.21)	0.06 (0.03)	0.06 (0.04)	0.25 (0.37)
Mean	**0.09 (0.04)**	**0.09 (0.03)**	**0.65 (0.27)**	**0.05 (0.03)**	**0.06 (0.03)**	**0.36 (0.20)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial fric	0.15 (0.10)	0.09 (0.02)	0.58 (0.35)	0.13 (0.05)	0.12 (0.03)	0.62 (0.41)
dental fric	0.07 (0.02)	0.07 (0.02)	0.09 (0.05)	0.07 (0.02)	0.06 (0.01)	0.13 (0.06)
Mean	**0.11 (0.08)**	**0.08 (0.02)**	**0.34 (0.35)**	**0.10 (0.05)**	**0.09 (0.04)**	**0.38 (0.38)**

Table 46. v-ratios of fricatives onsets in monosyllables with the /aʔ/ rime (T4-5), according to place of articulation and tone.

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
labial fric	0.09 (0.04)	0.89 (0.20)	0.04 (0.01)	0.67 (0.43)
dental fric	0.06 (0.02)	0.40 (0.33)	0.04 (0.02)	0.19 (0.33)
Mean	**0.08 (0.03)**	**0.64 (0.36)**	**0.04 (0.02)**	**0.43 (0.44)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
labial fric	0.08 (0.02)	0.51 (0.36)	0.08 (0.01)	0.36 (0.16)
dental fric	0.05 (0.01)	0.07 (0.02)	0.06 (0.01)	0.07 (0.02)
Mean	**0.06 (0.02)**	**0.29 (0.34)**	**0.07 (0.02)**	**0.56 (0.19)**

Table 47. v-ratios of fricative onsets (S1 context) for /ɛ/-rime syllables (T1-3), according to place of articulation and tone.

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial fric	0.14 (0.06)	0.14 (0.07)	0.80 (0.21)	0.08 (0.03)	0.09 (0.08)	0.36 (0.34)
dental fric	0.08 (0.02)	0.09 (0.02)	0.24 (0.24)	0.08 (0.04)	0.07 (0.03)	0.10 (0.07)
Mean	**0.12 (0.05)**	**0.11 (0.06)**	**0.52 (0.37)**	**0.08 (0.03)**	**0.08 (0.06)**	**0.24 (0.28)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial fric	0.11 (0.03)	0.16 (0.11)	0.66 (0.43)	0.18 (0.13)	0.16 (0.09)	0.65 (0.42)
dental fric	0.08 (0.02)	0.08 (0.02)	0.10 (0.04)	0.08 (0.03)	0.07 (0.02)	0.12 (0.05)
Mean	**0.09 (0.03)**	**0.12 (0.09)**	**0.38 (0.41)**	**0.13 (0.10)**	**0.11 (0.07)**	**0.39 (0.40)**

Table 48. v-ratios of fricative onsets (S1 context) for /aʔ/-rime syllables (T4-5), according to place of articulation and tone.

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
labial fric	0.11 (0.04)	0.59 (0.30)	0.05 (0.03)	0.35 (0.42)
dental fric	0.08 (0.03)	0.08 (0.03)	0.05 (0.03)	0.04 (0.02)
Mean	**0.10 (0.04)**	**0.34 (0.33)**	**0.05 (0.03)**	**0.15 (0.29)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
labial fric	0.11 (0.04)	0.48 (0.33)	0.11 (0.04)	0.24 (0.14)
dental fric	0.06 (0.01)	0.08 (0.33)	0.06 (0.02)	0.07 (0.02)
Mean	**0.09 (0.04)**	**0.28 (0.31)**	**0.09 (0.04)**	**0.15 (0.13)**

Table 49. Average v-ratios of S2 onsets for /ɛ/-rime syllables (T1-3), according to tone and place of articulation.

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
lab. stop	0.32 (0.15)	0.42 (0.22)	0.89 (0.24)	0.27 (0.07)	0.32 (0.09)	0.73 (0.28)
den. stop	0.33 (0.15)	0.30 (0.14)	0.80 (0.23)	0.29 (0.07)	0.30 (0.11)	0.87 (0.23)
lab. fric	0.28 (0.10)	0.33 (0.13)	1.00 (0.00)	0.25 (0.09)	0.29 (0.08)	1.00 (0.00)
den. fric	0.23 (0.06)	0.26 (0.06)	0.94 (0.19)	0.20 (0.07)	0.21 (0.05)	0.94 (0.12)
Mean	**0.29 (0.12)**	**0.33 (0.15)**	**0.91 (0.20)**	**0.25 (0.08)**	**0.28 (0.09)**	**0.89 (0.21)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
lab. stop	0.23 (0.10)	0.24 (0.06)	0.93 (0.14)	0.40 (0.30)	0.28 (0.06)	0.99 (0.02)
den. stop	0.25 (0.07)	0.28 (0.08)	0.98 (0.05)	0.30 (0.14)	0.29 (0.15)	0.98 (0.05)
lab. fric	0.29 (0.11)	0.28 (0.09)	1.00 (0.00)	0.48 (0.15)	0.43 (0.24)	1.00 (0.00)
den. fric	0.22 (0.06)	0.21 (0.05)	0.99 (0.03)	0.28 (0.07)	0.27 (0.06)	1.00 (0.00)
Mean	**0.25 (0.09)**	**0.25 (0.08)**	**0.98 (0.07)**	**0.36 (0.18)**	**0.32 (0.16)**	**0.99 (0.03)**

Table 50. Average v-ratios of S2 onsets for /aʔ/-rime syllables (T4-5), according to tone and place of articulation.

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
lab. stop	0.31 (0.11)	0.80 (0.18)	0.24 (0.09)	0.72 (0.24)
den. stop	0.56 (0.41)	0.91 (0.11)	0.23 (0.18)	0.83 (0.23)
lab. fric	0.25 (0.11)	1.00 (0.00)	0.24 (0.07)	0.99 (0.03)
den. fric	0.13 (0.06)	0.92 (0.18)	0.15 (0.04)	0.92 (0.13)
Mean	**0.28 (0.24)**	**0.91 (0.15)**	**0.21 (0.11)**	**0.86 (0.20)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
lab. stop	0.19 (0.07)	0.99 (0.03)	0.21 (0.09)	0.95 (0.08)
den. stop	0.14 (0.08)	0.91 (0.14)	0.19 (0.10)	0.95 (0.12)
lab. fric	0.21 (0.10)	1.00 (0.00)	0.29 (0.18)	1.00 (0.00)
den. fric	0.16 (0.07)	0.86 (0.32)	0.21 (0.04)	1.00 (0.00)
Mean	**0.17 (0.09)**	**0.94 (0.18)**	**0.23 (0.13)**	**0.98 (0.07)**

3. Detailed HNR results for word-initial stops (§4.1.6.3).

Table 51. HNR of stop onsets in monosyllables for /ɛ/-rime (T1-3) syllables, according to place of articulation and tone.

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial stop	7.6 (0.7)	7.6 (2.2)	6.1 (2.3)	5.0 (2.0)	3.9 (2.1)	2.8 (1.5)
dental stop	7.3 (2.3)	8.2 (2.6)	5.6 (2.3)	5.1 (1.3)	4.3 (2.1)	3.1 (1.8)
Mean	**7.5 (1.6)**	**7.8 (2.4)**	**5.9 (2.2)**	**5.1 (1.6)**	**4.1 (2.0)**	**3.0 (1.6)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial stop	10.7 (3.4)	9.9 (2.6)	5.2 (2.6)	8.9 (2.7)	8.2 (3.3)	0.7 (3.8)
dental stop	10.0 (2.6)	9.3 (1.6)	5.4 (2.4)	8.4 (2.0)	7.4 (1.9)	4.9 (2.3)
Mean	**10.3 (2.9)**	**9.5 (2.1)**	**5.3 (2.4)**	**8.4 (2.1)**	**7.6 (2.4)**	**3.7 (3.6)**

Table 52. HNR of stop onsets in monosyllables for /aʔ/-rime (T4-5) syllables, according to place of articulation and tone.

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
labial stop	6.5 (1.9)	4.1 (1.8)	3.2 (2.6)	2.0 (2.0)
dental stop	7.3 (1.8)	4.4 (0.8)	2.3 (2.0)	1.3 (1.4)
Mean	**6.9 (1.8)**	**4.3 (1.3)**	**2.8 (2.2)**	**1.7 (1.7)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
labial stop	9.3 (1.4)	3.1 (2.2)	4.9 (2.7)	2.7 (2.5)
dental stop	7.4 (1.0)	4.8 (1.7)	0.8 (2.7)	0.6 (1.2)
Mean	**8.4 (1.6)**	**4.0 (2.1)**	**2.3 (3.3)**	**1.3 (2.1)**

Table 53. HNR of stop onsets in the S1 context for /ɛ/-rime (T1-3) syllables, according to place of articulation and tone.

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial stop	8.0 (3.0)	6.2 (3.9)	6.1 (1.4)	5.1 (1.3)	5.5 (2.2)	2.7 (2.3)
dental stop	8.4 (2.2)	7.9 (2.0)	5.1 (2.1)	4.8 (2.9)	4.1 (1.4)	2.8 (1.6)
Mean	**8.3 (2.5)**	**7.1 (3.1)**	**5.6 (1.8)**	**4.9 (2.2)**	**4.8 (1.9)**	**2.8 (1.9)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial stop	11.1 (1.4)	9.0 (1.4)	5.0 (3.4)	7.3 (1.0)	6.2 (1.8)	1.4 (4.0)
dental stop	9.9 (2.1)	9.9 (1.4)	6.3 (3.6)	7.3 (2.6)	7.8 (1.4)	7.0 (6.3)
Mean	**10.5 (1.8)**	**9.5 (1.4)**	**5.6 (3.4)**	**7.2 (1.9)**	**7.2 (1.7)**	**4.0 (5.8)**

Table 54. HNR of stop onsets in the S1 context for /aʔ/-rime (T4-5) syllables, according to place of articulation and tone.

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
labial stop	7.4 (1.5)	3.7 (2.4)	1.4 (2.1)	0.4 (1.1)
dental stop	6.8 (1.8)	4.4 (2.0)	2.0 (1.3)	1.9 (1.5)
Mean	**7.1 (1.6)**	**3.9 (2.1)**	**1.7 (1.7)**	**1.2 (1.5)**
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
labial stop	7.0 (4.0)	4.0 (2.6)	3.4 (1.8)	2.5 (1.9)
dental stop	7.4 (2.1)	4.0 (2.8)	4.6 (3.1)	3.4 (2.4)
Mean	**7.3 (3.0)**	**3.9 (2.6)**	**4.4 (2.5)**	**3.1 (2.0)**

4. Detailed H1–H2 results for vowels in the S1 and S2 contexts (§4.1.6.3).

Table 55. H1-H2 of /ɛ/ rime syllables (T1-3) in the S1 context according to onset and tone, pooled across the first two time points.

T1-T3 vowel	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
p	1.77 (6.03)	-8.40(12.78)	0.84 (2.78)	1.83 (2.65)	2.61 (2.60)	4.48 (2.55)
t	-5.38(13.24)	-8.56(11.67)	0.28 (3.35)	1.68 (3.41)	1.28 (5.48)	4.27 (2.08)
f	-0.47 (5.84)	-0.76(4.00)	1.55 (1.70)	2.77 (2.98)	4.04 (1.42)	5.19 (1.67)
s	4.01 (5.56)	-5.23(12.09)	0.34 (2.95)	1.44 (2.65)	3.53 (2.14)	5.70 (3.21)
Mean	**-0.02 (8.86)**	**-5.74 (10.99)**	**0.75 (2.75)**	**1.93 (3.52)**	**2.87 (3.39)**	**4.91 (2.46)**
T1-T3 vowel	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
p	3.46 (10.52)	-1.03 (5.12)	1.95 (2.80)	0.61(14.19)	-3.77 (5.43)	4.07 (7.42)
t	1.70 (8.29)	-1.11 (8.51)	1.56 (2.62)	0.87 (9.17)	-5.03 (3.35)	2.58 (7.05)
f	4.50 (9.27)	0.81 (8.50)	1.86 (2.90)	-0.33 (7.16)	-5.14 (1.64)	3.40 (1.76)
s	6.75 (9.64)	1.63(10.37)	1.68 (2.16)	-2.89 (8.00)	-5.33 (4.59)	2.25 (3.95)
Mean	**4.10 (9.44)**	**0.07 (8.32)**	**1.76 (2.58)**	**-0.43 (9.80)**	**-4.82 (3.91)**	**3.08 (5.38)**

Table 56. H1-H2 of /aʔ/ rime syllables (T4-5) in the S1 context according to onset and tone, pooled across the first two time points.

T4-T5 vowel	*Young female*		*Young male*	
	T4	T5	T4	T5
p	0.12 (3.54)	3.49 (1.96)	3.07 (3.07)	6.60 (2.78)
t	-0.14 (3.06)	3.55 (3.38)	3.03 (8.64)	7.27 (2.31)
f	-0.77 (2.82)	2.53 (2.11)	4.69 (1.47)	3.73 (6.16)
s	1.06 (4.03)	2.79 (2.06)	1.94 (3.83)	4.94 (1.82)
Mean	**0.07 (3.35)**	**3.09 (2.41)**	**3.18 (4.99)**	**5.63 (3.83)**
T4-T5 vowel	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
p	2.40 (2.38)	5.86 (2.40)	2.93 (0.95)	5.94 (1.66)
t	3.08 (2.95)	7.14 (3.38)	3.58 (1.37)	4.58 (9.02)
f	2.80 (3.13)	4.37 (5.53)	1.06 (1.79)	7.73 (4.12)
s	0.34 (4.44)	4.41 (2.65)	-0.40 (3.35)	4.22 (4.52)
Mean	**2.16 (3.38)**	**5.45 (3.79)**	**1.79 (2.53)**	**5.62 (5.42)**

Table 57. H1-H2 of /ɛ/ rime syllables (T1-3) in the S2 context according to onset, tone and preceding tone, pooled across the first three time points.

T1-T3 vowel	*Young female*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
p	0.00 (3.45)	0.37 (3.63)	1.11 (3.00)	-1.60 (4.43)	-1.75 (4.09)	-2.45 (4.58)
t	-1.17 (4.45)	-0.69 (4.20)	-0.84 (3.87)	-2.45 (5.39)	-5.18 (7.70)	-3.42 (4.79)
f	1.01 (2.30)	0.63 (2.27)	1.57 (3.34)	-1.67 (4.18)	-1.31 (4.88)	-1.36 (4.08)
s	0.05 (2.46)	0.57 (2.72)	0.14 (2.91)	-0.76 (5.09)	-2.92 (9.23)	-2.26 (4.45)
Mean	**-0.03 (3.30)**	**0.22 (3.27)**	**0.49 (3.36)**	**-1.62 (4.73)**	**-2.79 (6.85)**	**-2.37 (4.45)**
T1-T3 vowel	*Young male*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
p	1.86 (3.02)	2.47 (1.51)	2.66 (1.72)	1.58 (4.08)	-0.19 (9.19)	1.35 (2.14)
t	2.70 (2.05)	2.12 (1.88)	2.15 (2.05)	0.90 (5.69)	-1.12 (3.69)	2.74 (1.36)
f	2.41 (1.63)	3.38 (1.42)	2.08 (3.28)	-0.03 (8.53)	1.58 (4.90)	3.35 (1.36)
s	2.77 (2.17)	2.27 (2.76)	1.74 (1.83)	3.48 (2.20)	2.57 (2.36)	2.90 (1.39)
Mean	**2.43 (2.26)**	**2.56 (1.99)**	**2.16 (2.29)**	**1.48 (5.66)**	**0.71 (5.72)**	**2.58 (1.74)**
T1-T3 vowel	*Elderly female*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
p	-1.06 (3.86)	-0.79 (3.03)	0.10 (3.42)	-9.36 (16.27)	-2.60 (6.92)	-0.66 (8.87)
t	-1.49 (4.20)	-1.64 (3.30)	-0.39 (4.89)	0.02 (7.92)	-1.33 (8.42)	-1.14 (7.47)
f	0.44 (2.65)	0.45 (2.14)	-1.26 (3.62)	-0.79 (6.86)	0.28 (7.46)	-0.04 (6.79)
s	0.03 (4.36)	0.96 (5.54)	1.19 (5.26)	0.28 (7.73)	0.32 (8.28)	0.69 (9.14)
Mean	**-0.52 (3.84)**	**-0.26 (3.78)**	**-0.09 (4.37)**	**-2.46(10.96)**	**-0.83 (7.73)**	**-0.29 (7.98)**
T1-T3 vowel	*Elderly male*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
p	-1.97 (3.61)	-2.77 (1.90)	-0.94 (2.13)	-6.25 (2.28)	-5.49 (2.58)	-5.01 (2.37)
t	-2.80 (3.24)	-3.70 (2.92)	-3.12 (2.70)	-6.57 (2.28)	-6.37 (2.75)	-6.26 (2.53)
f	-2.11 (1.98)	-1.75 (1.34)	-2.44 (2.73)	-4.53 (1.89)	-3.13 (2.21)	-2.82 (0.83)
s	-2.40 (1.95)	-0.73 (2.35)	-1.95 (2.78)	-2.40 (2.06)	-3.01 (3.78)	-4.10 (2.27)
Mean	**-2.32 (2.72)**	**-2.24 (2.42)**	**-2.11 (2.64)**	**-4.94 (2.99)**	**-4.50 (3.16)**	**-4.55 (2.41)**

5. Detailed durations for onsets (§4.1.6.4).

Table 58. Average fricative and nasal onset durations in *yin* monosyllables (T1-3), according to Tone and Place of articulation.

T1-3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
lab. fric.	190 (30)	173 (34)	112 (14)	206 (30)	219 (46)	115 (30)
den. fric.	185 (33)	185 (24)	135 (30)	211 (40)	221 (43)	151 (24)
lab. nas.	108 (55)	95 (18)	91 (15)	96 (29)	91 (9)	91 (36)
den. nas.	78 (15)	—	80 (17)	96 (21)	—	88 (17)
T1-3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
lab. fric.	190 (75)	181 (27)	101 (47)	151 (48)	148 (32)	113 (13)
den. fric.	242 (60)	238 (36)	174 (49)	202 (30)	203 (31)	136 (44)
lab. nas.	95 (15)	98 (25)	100 (20)	79 (18)	79 (27)	90 (12)
den. nas.	91 (22)	—	129 (37)	77 (22)	—	101 (14)

Table 59. Average fricative and nasal onset durations in *yang* monosyllables (T4-5), according to Tone and Place of articulation.

T4-5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
lab. fric.	196 (56)	141 (32)	216 (35)	124 (20)
den. fric.	235 (41)	178 (35)	225 (43)	192 (24)
lab. nas.	—	144 (23)	—	119 (29)
den. nas.	—	135 (23)	—	122 (13)
T4-5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
lab. fric.	192 (57)	129 (67)	146 (42)	125 (19)
den. fric.	252 (58)	224 (77)	204 (45)	187 (31)
lab. nas.	—	106 (21)	—	130 (16)
den. nas.	—	132 (43)	—	110 (16)

Table 60. Average fricative onset durations for unchecked S1 syllables (tones T1-3).

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
fric. lab.	150 (26)	170 (53)	76 (28)	173 (27)	167 (35)	90 (21)
fric. den.	146 (20)	166 (31)	117 (21)	174 (30)	181 (23)	144 (19)
Mean	**148 (22)**	**168 (42)**	**87 (32)**	**173 (27)**	**174 (28)**	**117 (34)**
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
fric. lab.	149 (28)	150 (11)	81 (30)	140 (35)	131 (33)	92 (13)
fric. den.	185 (32)	188 (38)	136 (32)	164 (20)	187 (23)	122 (34)
Mean	**167 (34)**	**169 (33)**	**108 (41)**	**152 (29)**	**159 (40)**	**107 (28)**

Table 61. Average fricative onset durations for checked S1 syllables (tones T4-5).

T4-5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
fric. lab.	159 (29)	79 (29)	170 (30)	56 (28)
fric. den.	167 (41)	131 (30)	199 (31)	154 (23)
Mean	**163 (34)**	**105 (39)**	**184 (33)**	**105 (57)**
T4-5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
fric. lab.	135 (32)	66 (30)	122 (27)	57 (15)
fric. den.	183 (30)	147 (11)	172 (41)	126 (13)
Mean	**159 (39)**	**107 (48)**	**147 (42)**	**91 (39)**

Table 62. Average release and closure durations of stop onsets and durations of fricative onsets in S2, according to place of articulation and speaker group (unchecked syllables).

T1-T3 onset	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
stop lab. rel.	14 (3.6)	13 (2.8)	10 (1.7)	20 (7.4)	20 (6.9)	15 (4.3)
stop lab. clos.	94 (18)	98 (22)	59 (19)	88 (15)	96 (25)	56 (16)
stop dent. rel.	14 (1.8)	13 (1.9)	12 (2.7)	25 (10.1)	22 (7.3)	17 (5.6)
stop dent. clos.	87 (17)	94 (17)	51 (15)	85 (17)	82 (17)	55 (19)
fricative lab.	122 (16)	120 (14)	54 (12)	119 (11)	118 (14)	63 (12)
fricative den.	128 (19)	125 (15)	59 (9)	134 (15)	128 (12)	58 (10)
T1-T3 onset	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
stop lab. rel.	13 (3.1)	13 (2.3)	10 (2.6)	16 (2.2)	17 (3.4)	11 (2.9)
stop lab. clos.	112 (27)	121 (24)	55 (13)	105 (12)	106 (9)	53 (8)
stop dent. rel.	15 (4.2)	14 (5.0)	12 (4.1)	17 (5.8)	15 (4.3)	12 (3.7)
stop dent. clos.	120 (21)	111 (26)	49 (15)	95 (18)	92 (16)	45 (8)
fricative lab.	134 (22)	132 (27)	55 (11)	98 (12)	94 (10)	50 (7)
fricative dent.	146 (26)	146 (33)	60 (8)	120 (10)	110 (15)	56 (11)

Table 63. Average release and closure durations of stop onsets and durations of fricative onsets in S2, according to place of articulation and speaker group (checked syllables).

T4-T5 onset	*Young female*		*Young male*	
	T4	T5	T4	T5
stop lab. rel.	11 (2.4)	11 (3.2)	17 (7.6)	14 (5.6)
stop lab. clos.	115 (17)	58 (11)	112 (15)	64 (10)
stop dent. rel.	11 (2.9)	11 (2.4)	18 (7.0)	17 (6.3)
stop dent. clos.	89 (43)	53 (12)	103 (26)	57 (15)
fricative lab.	146 (19)	69 (13)	134 (17)	75 (16)
fricative dent.	145 (24)	65 (16)	145 (18)	72 (18)
T4-T5 onset	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
stop lab. rel.	9 (1.9)	9 (2.9)	13 (3.3)	11 (4.5)
stop lab. clos.	148 (28)	55 (17)	114 (16)	57 (10)
stop dent. rel.	12 (3.8)	12 (5.0)	14 (4.9)	12 (4.6)
stop dent. clos.	139 (20)	47 (16)	119 (28)	47 (15)
fricative lab.	156 (36)	70 (24)	109 (17)	60 (11)
fricative dent.	167 (30)	75 (29)	139 (36)	65 (12)

6. Detailed durations for rimes (§4.1.6.4).

Table 64. Average /ɛ/ rime durations in monosyllables, according to tone, syllable onset, and speaker group.

/ɛ/ rime	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
zero	272 (58)	321 (45)	348 (30)	237 (56)	273 (68)	268 (45)
lab. stop	237 (33)	300 (54)	323 (24)	214 (36)	277 (81)	265 (39)
den. stop	251 (44)	311 (35)	305 (30)	214 (44)	293 (58)	273 (50)
lab. fric	243 (55)	295 (39)	307 (40)	200 (36)	270 (57)	276 (52)
den. fric	229 (30)	283 (22)	286 (22)	202 (40)	268 (57)	262 (50)
lab. nasal	238 (55)	296 (32)	294 (59)	187 (26)	278 (61)	266 (43)
den. nasal	245 (67)	—	306 (34)	216 (74)	—	262 (37)
Mean	**245 (48)**	**301 (38)**	**310 (38)**	**210 (46)**	**277 (60)**	**267 (42)**
/ɛ/ rime	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
zero	304 (71)	329 (40)	372 (41)	270 (41)	257 (39)	349 (91)
lab. stop	244 (37)	289 (29)	344 (70)	209 (47)	255 (40)	284 (39)
den. stop	253 (37)	319 (38)	313 (39)	196 (49)	263 (37)	291 (57)
lab. fric	246 (26)	298 (49)	327 (43)	208 (68)	263 (44)	276 (50)
den. fric	235 (26)	314 (33)	336 (29)	186 (76)	245 (45)	283 (47)
lab. nasal	230 (29)	301 (30)	335 (49)	204 (53)	259 (37)	273 (48)
den. nasal	256 (31)	—	337 (49)	197 (56)	—	270 (41)
Mean	**252 (43)**	**308 (37)**	**338 (47)**	**210 (56)**	**257 (36)**	**289 (55)**

Table 65. Average /aʔ/ rime durations in monosyllables, according to tone, syllable onset, and speaker group.

/aʔ/ rime	*Young female*		*Young male*	
	T4	T5	T4	T5
zero	144 (34)	201 (43)	103 (12)	161 (23)
lab. stop	90 (18)	151 (24)	97 (11)	143 (12)
den. stop	86 (17)	166 (25)	88 (8)	135 (12)
lab. fric	106 (22)	167 (39)	101 (17)	138 (28)
den. fric	92 (25)	166 (27)	93 (22)	151 (15)
lab. nasal	—	141 (30)	—	128 (18)
den. nasal	—	140 (31)	—	135 (17)
Mean	**104 (31)**	**175 (22)**	**98 (15)**	**148 (20)**
/aʔ/ rime	*Elderly female*		*Elderly male*	
	T4	T5	T4	T5
zero	126 (10)	212 (27)	118 (13)	172 (23)
lab. stop	107 (11)	175 (23)	100 (13)	141 (24)
den. stop	99 (14)	180 (19)	91 (8)	147 (24)
lab. fric	133 (26)	180 (15)	116 (21)	145 (25)
den. fric	115 (11)	198 (22)	97 (13)	148 (18)
lab. nasal	—	162 (21)	—	127 (18)
den. nasal	—	177 (28)	—	137 (8)
Mean	**116 (19)**	**183 (26)**	**104 (17)**	**145 (22)**

Table 66. Average duration of the /ɛ/ rime (T1-3), detailed by S1 onset types.

/ɛ/ rime	*Young female*			*Young male*		
	T1	T2	T3	T1	T2	T3
labial stop	169 (22)	207 (41)	215 (56)	158 (15)	168 (31)	183 (38)
dental stop	179 (23)	195 (43)	205 (55)	164 (21)	172 (31)	177 (35)
labial fricative	202 (35)	217 (67)	221 (53)	165 (20)	178 (25)	195 (17)
dental fricative	183 (43)	186 (52)	217 (59)	162 (23)	163 (22)	185 (47)
Mean	**183 (32)**	**201 (50)**	**215 (52)**	**162 (19)**	**170 (26)**	**185 (34)**
/ɛ/ rime	*Elderly female*			*Elderly male*		
	T1	T2	T3	T1	T2	T3
labial stop	192 (38)	173 (40)	198 (30)	173 (53)	188 (51)	244 (62)
dental stop	224 (44)	203 (35)	215 (39)	200 (46)	203 (53)	234 (56)
labial fricative	237 (33)	251 (66)	237 (38)	172 (51)	196 (24)	231 (53)
dental fricative	209 (38)	198 (24)	229 (38)	197 (68)	185 (45)	227 (42)
Mean	**216 (40)**	**210 (50)**	**220 (37)**	**186 (51)**	**193 (41)**	**234 (48)**

Table 67. Average duration of the /aʔ/ rime (T4-5), detailed by S1 onset types.

/aʔ/ rime	*Young female*		*Young male*	
	T4	T5		T4
labial stop	67 (5)	70 (15)	labial stop	67 (5)
dental stop	81 (15)	120 (16)	dental stop	81 (15)
labial fricative	93 (8)	96 (12)	labial fricative	93 (8)
dental fricative	85 (8)	78 (14)	dental fricative	85 (8)
Mean	**81 (13)**	**91 (24)**	**Mean**	**81 (13)**
/aʔ/ rime	*Elderly female*		*Elderly male*	
	T4	T5		T4
labial stop	57 (12)	67 (13)	labial stop	57 (12)
dental stop	66 (14)	109 (16)	dental stop	66 (14)
labial fricative	80 (21)	77 (15)	labial fricative	80 (21)
dental fricative	71 (13)	93 (19)	dental fricative	71 (13)
Mean	**68 (17)**	**86 (22)**	**Mean**	**68 (17)**

Table 68. Average /ɛ/ rime duration (T1-3), according to S2 onset type and speaker group.

/ɛ/ rime	*Young female*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
lab. stop	158 (40)	183 (24)	175 (30)	194 (28)	189 (41)	207 (32)
den. stop	163 (35)	178 (20)	178 (20)	210 (36)	214 (37)	219 (23)
lab. fric	140 (26)	151 (18)	177 (26)	187 (27)	188 (48)	222 (35)
den. fric	133 (20)	169 (33)	171 (24)	189 (32)	193 (17)	216 (30)
Mean	**149 (32)**	**170 (26)**	**175 (24)**	**195 (31)**	**196 (37)**	**216 (29)**
/ɛ/ rime	*Young male*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
lab. stop	144 (20)	151 (22)	160 (18)	174 (23)	185 (19)	191 (36)
den. stop	134 (24)	162 (19)	160 (14)	185 (28)	180 (38)	198 (24)
lab. fric	152 (23)	139 (20)	161 (18)	174 (33)	175 (25)	193 (18)
den. fric	144 (41)	130 (29)	158 (15)	178 (31)	186 (30)	196 (33)
Mean	**151 (21)**	**146 (25)**	**160 (16)**	**178 (27)**	**182 (27)**	**194 (27)**
/ɛ/ rime	*Elderly female*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
lab. stop	182 (47)	174 (18)	186 (43)	208 (41)	218 (46)	223 (41)
den. stop	179 (27)	199 (35)	201 (41)	223 (51)	207 (37)	244 (41)
lab. fric	181 (25)	182 (47)	193 (48)	213 (41)	226 (36)	229 (51)
den. fric	193 (52)	180 (30)	179 (44)	213 (52)	238 (64)	230 (56)
Mean	**184 (38)**	**184 (33)**	**190 (42)**	**214 (44)**	**222 (46)**	**232 (45)**
/ɛ/ rime	*Elderly male*					
	Post T1			Post T2		
	T1	T2	T3	T1	T2	T3
lab. stop	151 (52)	137 (24)	161 (31)	165 (33)	194 (35)	195 (38)
den. stop	150 (13)	169 (31)	166 (38)	190 (36)	179 (20)	210 (39)
lab. fric	169 (26)	174 (38)	161 (14)	172 (33)	188 (40)	203 (47)
den. fric	162 (43)	153 (36)	166 (29)	180 (28)	213 (56)	187 (30)
Mean	**158 (34)**	**158 (33)**	**163 (26)**	**177 (31)**	**193 (38)**	**199 (36)**

Table 69. Average /aʔ/ rime duration (T4-5), according to S2 onset type and speaker group.

/aʔ/ rime	*Young female*			
	Post T1		Post T2	
	T4	T5	T4	T5
lab. stop	81 (29)	127 (33)	104 (41)	123 (23)
den. stop	73 (26)	121 (23)	112 (29)	144 (25)
lab. fric	99 (24)	103 (31)	107 (37)	143 (25)
den. fric	93 (28)	132 (29)	80 (32)	148 (36)
Mean	**87 (27)**	**121 (30)**	**101 (35)**	**139 (27)**
/aʔ/ rime	*Young male*			
	Post T1		Post T2	
	T4	T5	T4	T5
lab. stop	107 (15)	102 (19)	96 (17)	105 (21)
den. stop	97 (17)	101 (19)	101 (10)	125 (9)
lab. fric	99 (15)	117 (17)	106 (10)	127 (15)
den. fric	95 (12)	119 (22)	94 (16)	134 (10)
Mean	**99 (15)**	**110 (20)**	**99 (14)**	**123 (17)**
/aʔ/ rime	*Elderly female*			
	Post T1		Post T2	
	T4	T5	T4	T5
lab. stop	86 (22)	121 (32)	106 (20)	117 (21)
den. stop	105 (21)	112 (26)	114 (31)	140 (27)
lab. fric	111 (19)	109 (24)	118 (23)	127 (22)
den. fric	115 (25)	122 (40)	94 (25)	133 (22)
Mean	**104 (23)**	**116 (30)**	**108 (25)**	**129 (23)**
/aʔ/ rime	*Elderly male*			
	Post T1		Post T2	
	T4	T5	T4	T5
lab. stop	81 (16)	86 (17)	96 (16)	107 (9)
den. stop	70 (11)	93 (23)	94 (10)	116 (22)
lab. fric	85 (10)	87 (14)	87 (7)	108 (17)
den. fric	80 (14)	106 (23)	96 (16)	114 (11)
Mean	**79 (13)**	**93 (19)**	**96 (15)**	**111 (14)**

Appendix 3. Statistical results

1. Results of Linear discriminant analyses (LDA) of Experiment 1 (§4.1.6.3):

Table 70. Results of the linear discriminant analyses for (1) unchecked and (2) checked syllables: (A) young female, (B) young male, (C) elderly female, and (D) elderly male speakers. Asterisked acoustic measures are significant according to stepwise discriminant analyses. Shaded cells indicate the most important measure revealed by the coefficients of discriminants functions or the partial Wilks' Lambda.

(1A) Young female

Acoustic measure	Breathy (T3) vs. Modal (T1-2)				T2 vs. T3			
	Wilks' lambda	Coeff.	*F* value	*p* value	Wilks' lambda	Coeff.	*F* value	*p* value
* H1-H2		-0.094	9.5	<.005		0.130	7.9	<.01
H1-A1		-0.047	—	n.s.		0.140	—	n.s.
* H1-A2	0.88	-0.121	—	n.s.	0.78	-0.191	4.1	<.05
* CPP		-0.230	—	n.s.		-0.643	9.2	<.005
* F1		-0.005	9.4	<.005		0.005	4.9	<.05

(1B) Young male

Acoustic measure	Breathy (T3) vs. Modal (T1-2)				T2 vs. T3			
	Wilks' lambda	Coeff.	*F* value	*p* value	Wilks' lambda	Coeff.	*F* value	*p* value
* H1-H2		-0.276	35.5	<.0001		0.194	—	n.s.
* H1-A1		-0.095	13.1	<.0005		0.029	—	n.s.
H1-A2	0.74	-0.094	—	n.s.	0.73	0.096	—	n.s.
* CPP		-0.114	—	n.s.		-0.262	6.6	<.05
* F1		-0.007	7.5	<.01		0.015	—	n.s.

(1C) Elderly female

Acoustic measure	Breathy (T3) vs. Modal (T1-2)				T2 vs. T3			
	Wilks' lambda	Coeff.	*F* value	*p* value	Wilks' lambda	Coeff.	*F* value	*p* value
* H1-H2		-0.468	92.8	<.0001		0.274	21.2	<.0001
H1-A1		0.118		n.s.		-0.010	—	n.s.
* H1-A2	0.59	-0.114	7.2	<.01	0.31	0.015	—	n.s.
* CPP		-0.068		n.s.		-0.448	113.8	<.0001
(*) F1		-0.004	4.8	<.05		0.007	32.2	<.0001

NB: F1 is overall higher for modal than breathy vowels for elderly females, contrary to the predicted result.

(1D) Elderly male

Acoustic measure	Breathy (T3) vs. Modal (T1-2)				T2 vs. T3			
	Wilks' lambda	Coeff.	*F* value	*p* value	Wilks' lambda	Coeff.	*F* value	*p* value
* H1-H2		-0.504	212.2	<.0001		0.490	210.9	<.0001
H1-A1		-0.129	—	n.s.		0.150	—	n.s.
H1-A2	0.23	-0.023	—	n.s.	0.23	0.021	—	n.s.
* CPP		0.165	6.6	<.05		-0.157	6.3	<.05
F1		0.003	—	n.s.		-0.003	—	n.s.

(2A) Young female

Acoustic measure	Modal (T4) vs. Breathy (T5)			
	Wilks' lambda	Coeff.	*F* value	*p* value
* H1-H2		0.361	27.2	<.0001
H1-A1		0.078	—	n.s.
H1-A2	0.72	0.004	—	n.s.
CPP		0.252	—	n.s.
* F1		-0.003	4.3	<.05

(2B) Young male

Acoustic measure	Modal (T4) vs. Breathy (T5)			
	Wilks' lambda	**Coeff.**	***F* value**	***p* value**
* H1-H2		0.261	10.6	<.005
* H1-A1		0.088	18.2	<.0001
H1-A2	0.68	0.085	—	n.s.
* CPP		0.413	17.6	<.0001
F1		0.000	—	n.s.

(2C) Elderly female

Acoustic measure	Modal (T4) vs. Breathy (T5)			
	Wilks' lambda	**Coeff.**	***F* value**	***p* value**
* H1-H2		0.008	51.2	<.0001
H1-A1		0.053	—	n.s.
H1-A2	0.58	0.097	—	n.s.
CPP		-0.198	—	n.s.
F1		0.006	—	n.s.

(2D) Elderly male

Acoustic measure	Modal (T4) vs. Breathy (T5)			
	Wilks' lambda	**Coeff.**	***F* value**	***p* value**
* H1-H2		0.456	75.6	<.0001
* H1-A1		0.371	18.5	<.0001
* H1-A2	0.33	-0.135	5.2	<.05
CPP		-0.081	—	n.s.
F1		-0.001	—	n.s.

2. Results of generalized linear models (GLM) of Experiment 3 (§5.1):

(1) Congruence predictor

Models:

m2: response ~ age + gender + place + manner + itone + vowel + (1 | subject)

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m2	9	2027.2	2084.7	-1004.61			
m1	10	1949.1	2013.0	-964.6	80.085	1	< 2.2e-16 ***

(2) Place predictor

Models:

m3: response ~ change + age + gender + manner + itone + (1 | subject)

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m3	9	2018.0	2075.5	-999.98			
m1	10	1949.1	2013.0	-964.6	70.822	1	< 2.2e-16 ***

(3) Manner predictor

Models:

m4: response ~ change + age + gender + place + itone + (1 | subject)

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m4	9	1980.0	2037.5	-980.98			
m1	10	1949.1	2013.0	-964.6	32.8	1	1.005e-08 ***

(4) Imposed tone predictor

Models:

m5: response ~ change + age + gender + place + manner + itone + (1 | subject)

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m4	8	2003.7	2054.8	-993.84			
m1	10	1949.1	2013.0	-964.6	58.6	2	1.932e-13 ***

(5) Imposed tone predictor

Models:

m6: response ~ change + age + gender + place + manner + vowel + (1 | subject)

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m6	9	1989.0	2046.5	-985.48			
m1	10	1949.1	2013.0	-964.6	41.8	1	9.955e-11 ***

(6) Congruence × Place predictor

Models:

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

m7: response ~ change * place + age + gender + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	10	1949.1	2013.0	-964.57			
m7	11	1929.9	2000.2	-954.0	21.237	1	4.058e-06 ***

(7) Congruence × Manner predictor

Models:

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

m8: response ~ change * manner + age + gender + place + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	9	1967.2	2024.7	-974.58			
m8	10	1949.1	2013.0	-964.6	20.02	1	7.673e-06 ***

(8) Congruence × Vowel predictor

Models:

m1: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

m9: response ~ change * vowel + age + gender + place + itone + manner + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	10	1949.1	2013.0	-964.57			
m9	12	1934.4	2011.0	-955.2	18.75	2	8.48e-05 ***

(9) Congruence predictor for data subset without /f,v/ onsets

Models:

m2b: response ~ age + gender + place + manner + vowel + itone + (1 | subject)

m1b: response ~ change + age + gender + place + manner + itone + vowel + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m2b	9	1129.0	1183.9	-555.50			
m1b	10	1129.6	1190.6	-554.8	1.4168	1	0.2339

(10) Congruence predictor for /f,v/ onsets data subset

Models:

m2c: response ~ age + gender + vowel + itone + (1 | subject)

m1c: response ~ change + age + gender + vowel + itone + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m2c	7	793.1	828.03	-389.53			
m1c	8	634.5	674.4	-309.2	160.6	1	< 2.2e-16 ***

3. Results of generalized linear models (GLM) of Experiment 4 (§5.1):

(1) Duration predictor

Models:

m2: response ~ onset + rime + otone + step + (1 | subject)

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m2	12	2994.8	3076.4	-1485.4			
m1	13	2937.4	3025.8	-1455.7	59.347	1	1.322e-14 ***

(2) Onset predictor

Models:

m3: response ~ duration + rime + otone + step + (1 | subject)

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m3	12	3182.2	3263.8	-1579.1			
m1	13	2937.4	3025.8	-1455.7	246.74	1	< 2.2e-16 ***

(3) Rime predictor

Models:

m4: response ~ duration + onset + otone + step + (1 | subject)

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m4	12	3044.3	3125.9	-1510.1			
m1	13	2937.4	3025.8	-1455.7	108.84	1	< 2.2e-16 ***

(4) Original tone predictor

Models:

m5: response ~ duration + onset + rime + step + (1 | subject)

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m5	12	2977.8	3059.4	-1476.9			

m1 13 2937.4 3025.8 -1455.7 42.373 1 7.543e-11 ***

(5) Duration × Onset predictor

Models:

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

m6: response ~ duration * onset + rime + otone + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	13	2937.4	3025.8	-1455.7			
m6	14	2918.5	3013.7	-1445.3	20.926	1	4.775e-06 ***

(6) Duration × Rime predictor

Models:

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

m7: response ~ duration * rime + onset + otone + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	13	2937.4	3025.8	-1455.7			
m7	14	2928.6	3023.8	-1450.3	10.806	1	0.001012 **

(7) Duration × Original tone predictor

Models:

m1: response ~ duration + onset + rime + otone + step + (1 | subject)

m8: response ~ duration * otone + onset + rime + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	13	2937.4	3025.8	-1455.7			
m8	14	2924.5	3019.7	-1448.2	14.973	1	0.0001091 ***

4. Results of generalized linear models (GLM) of Experiment 5 (§5.1):

A. overall *yang* response rate for synthesized stimuli

(1) Voice quality predictor

Models:

m2: response ~ gender + onset + step + (1 | subject)

m1: response ~ quality + gender + onset + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m2	14	1574.3	1656.1	-773.1			
m1	15	1539.9	1627.6	-755.0	36.4	1	1.647e-09 ***

(2) Onset predictor

Models:

m3: response ~ quality + gender + step + (1 | subject)

m1: response ~ quality + gender + onset + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m3	11	2040.8	2105.1	-1009.4			
m1	15	1539.9	1627.6	-755.0	508.9	4	< 2.2e-16 ***

(3) Step predictor

Models:

m4: response ~ quality + gender + step + (1 | subject)

m1: response ~ quality + gender + onset + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m4	11	2040.8	2105.1	-1009.4			
m1	15	1539.9	1627.6	-755.0	508.9	4	< 2.2e-16 ***

(4) Voice quality × Onset predictor

Models:

m1: response ~ quality + gender + onset + step + (1 | subject)

m5: response ~ quality * onset + gender + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
ms1	15	1539.9	1627.6	-755.0			
ms5	19	1547.0	1658.0	-754.5	0.9515	4	0.917

B. overall *yang* response rate for natural stimuli

(1) Voice quality predictor

Models:

m2: response ~ gender + onset + step + (1 | subject)

m1: response ~ quality + gender + onset + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m2	15	1776.2	1864.6	-873.1			
m1	16	1688.8	1783.19	-828.4	89.3	1	< 2.2e-16 ***

(2) Onset predictor

Models:

m3: response ~ quality + gender + step + (1 | subject)

m1: response ~ quality + gender + onset + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m3	11	1734.8	1799.7	-856.4			
m1	16	1688.9	1783.2	-828.4	56.0	5	8.281e-11 ***

(3) Step predictor

Models:

m4: response ~ quality + gender + onset + (1 | subject)

m1: response ~ quality + gender + onset + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m4	9	3547.0	3600.1	-1764.5			
m1	16	1688.9	1783.2	-828.4	1872.2	7	< 2.2e-16 ***

(4) Voice quality × Onset predictor

Models:

m1: response ~ quality + gender + onset + step + (1 | subject)

m5: response ~ quality * onset + gender + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m1	16	1688.9	1783.2	-828.4			
m5	21	1669.3	1793.1	-813.7	29.6	5	1.809e-05 ***

(5) Voice quality predictor for /m/ onset data subset

Models:

m2: response ~ gender + step + (1 | subject)

m1: response ~ quality + gender + step + (1 | subject)

	Df	AIC	BIC	logLik	Chisq	Chi	Df	Pr(>Chisq)
m2	10	323.67	364.68	-151.84				
m1	11	325.67	370.77	-151.83	0.0046		1	0.946

LIST OF TABLES

LIST OF FIGURES

Interdépendance entre tons, segments et types de phonation en shanghaïen: acoustique, articulation, perception et évolution

Résumé

Cette étude porte sur les corrélats phonétiques des registres tonals *yin* vs. *yang* du shanghaïen parlé dans la région urbaine de Shanghai. Nos investigations acoustique, articulatoire et perceptive ont montré qu'en dehors du F0, des indices multi-dimensionnels comme le voisement (voisé pour *yang* et non-voisé pour *yin*), le pattern de durée (ratio C/V bas pour *yang* et élevé pour *yin*), et le type de phonation (soufflé pour *yang* et modal pour *yin*) participent tous à la définition du registre tonal. Parmi tous ces indices, nous tâchons de distinguer les traits redondants liés aux effets coarticulatoires des survivances de changements diachroniques. En particulier, la voix soufflée qui accompagne les tons *yang* est un trait redondant, issu d'une évolution tonale qui est la transphonologisation de distinction de voisement vers la distinction de registre tonal, ou « bipartition tonale ». Nous proposons que la perte d'un trait redondant issu d'un changement diachronique peut être très lente si ce trait ne contrarie pas les effets coarticulatoires et/ou si le trait a une fonction perceptive.

En nous basant sur les données synchroniques des locuteurs de deux générations (20-30 ans vs. 60-80 ans), nous constatons une tendance vers la disparition de cette phonation soufflée. Nous constatons également une évolution plus avancée chez les femmes que les hommes de leur âge. Dans notre étude, nous essayons d'expliquer ce changement tant par des causes internes que par des causes externes.

Mots clés : Shanghaïen, ton, voix soufflée, durée, voisement, transphonologisation

Interdependence between Tones, Segments, and Phonation types in Shanghai Chinese: acoustics, articulation, perception, and evolution

Abstract

This study bears on the phonetic correlates of the *yin* vs. *yang* tone registers of Shanghai Chinese as spoken in Shanghai urban area. Our acoustic, articulatory, and perceptual investigations showed that beside F0, multidimensional cues, such as voicing (voiced for *yang* vs. voiceless for *yin*), duration pattern (low C/V ratio for *yang* vs. high C/V ratio for *yin*), and phonation type (breathy for *yang* vs. modal for *yin*) enter in the specification of tone register. Among all these cues, we attempt to distinguish the redundant features related to coarticulatory effects from those that are remnants of diachronic changes. In particular, the breathy voice accompanying *yang* tones, which is a redundant feature, arose from a tonal evolution, namely the transphonologization of a voicing contrast into a tone register contrast, that is, the "tone split." We propose that the loss of a redundant feature arisen from a diachronic change may be very slow if that feature does not conflict with coarticulatory effects and/or if that feature has a perceptual function.

Based on the synchronic data from the speakers of two generations (20-30 years vs. 60-80 years), we find a trend toward the loss of this breathy phonation. We also find that this evolution is more advanced in women than men of the same age. In our study, we try to explain this change by internal factors as well as by external factors.

Keywords : Shanghai Chinese, tone, breathy voice, duration, voicing, transphonologization

UNIVERSITE SORBONNE NOUVELLE - PARIS 3
ED 268 « Langage et langues : description, théorisation, transmission »
Laboratoire de Phonétique et Phonologie (CNRS-Paris 3)
1 rue Censier, 75005 Paris