# Etude acoustique et perceptive des hésitations autonomes multilingues

*Jeanne Clerc-Renaud*[1], *Ioana Vasilescu*[2], *Maria Candea*[3], *Martine Adda-Decker*[1]

[1]LIMSI-CNRS, bat. 508, BP 133, F-91403 Orsay cedex, [2]LTCI-ENST, 46, rue Barrault, 75634 Paris cedex 13, [3]Paris 3 - EA 1483, bur.431, 13 rue Santeuil, 75005 Paris

Mél : jeanne.clerc-renaud@isep.fr, vasilesc@tsi.enst.fr, candea@ext.jussieu.fr, madda@limsi.fr

## ABSTRACT

This paper deals with the analysis of autonomous filled pauses in 8 languages. The general question addressed here concerns whether filled pauses carry language-specific information. They occur frequently in spontaneous speech and represent an interesting topic for improving language-specific models for automatic language identification. Most of the current studies focus on few languages as English and French. Our study aims to describe the acoustic peculiarities of autonomous filled pauses in French, English, Spanish, Italian, Portuguese, German, Mandarin Chinese and Arabic. Thus we propose an acoustic typology based on the vocalic characteristics of the filled pauses. We also present a perceptual experiment concerning French/L2 discrimination based on filled pauses stimuli.

## 1. INTRODUCTION

La transcription automatique de la parole non lue pose de nouveaux défis aux systèmes de reconnaissance de la parole. Qu'elle soit préparée ou spontanée, les mots sont souvent moins bien articulés qu'en parole lue, et une part non négligeable du signal acoustique correspond à des « disfluences » dont les phénomènes dits d'hésitation font partie [6].

Parmi les différents phénomènes d'hésitation généralement répertoriés, il y en a un qui se manifeste dans de nombreuses langues : la possibilité d'insérer pratiquement à tout moment dans la parole une unité généralement vocalique et allongée, dont le seul sens est de marquer une recherche de formulation ou une rupture inattendue dans la parole. Ce phénomène est à distinguer de l'allongement d'une voyelle appartenant à un mot précis du lexique. La plupart des études sur de grands corpus ayant été faite sur du français ou de l'anglais [1], [2], [3], [4], [5], [6] il a été constaté que la voyelle autonome d'hésitation est une voyelle centrale fortement allongée. On peut se demander si toutes les langues utilisent plus ou moins le même type de voyelle (central par rapport au système vocalique) pour produire des hésitations autonomes.

Dans la perspective d'une modélisation économique et efficace pour l'identification automatique des langues se pose la question du type de modèle pour les hésitations : peut-on envisager un modèle commun à toutes les langues ou faut-il plutôt construire des modèles dépendants de la langue ? Y a-t-il suffisamment de différences entre les hésitations autonomes des différentes langues pour justifier la mise en place de modèles différents ?

Ce travail exploratoire s'inscrit dans le projet MIDL (Modélisations pour l'identification des langues) faisant partie du programme interdisciplinaire du CNRS STIC-SHS et réunissant les laboratoires LIMSI-CNRS, LTCI-ENST, CTA/DGA, ILPGA Paris et EA 1483 - Paris 3. Le but du projet est de mettre ensemble des compétences pluridisciplinaires linguistiques et informatiques afin d'augmenter nos connaissances sur l'identification des langues et éventuellement contribuer à améliorer la modélisation pour l'identification automatique des langues. Le présent travail a un double objectif. Il s'agit d'une part de caractériser les voyelles autonomes dites d'hésitations dans un corpus de 8 langues du point de vue de leurs particularités acoustiques et d'autre part de tester si l'hésitation porte une information permettant à l'humain d'identifier la langue.

L'étude des hésitations dans un cadre multilingue est une perspective relativement nouvelle. A notre connaissance il n'y a pas d'autres travaux sur ce sujet.

Nous présenterons dans la première partie le corpus multilingue sur lequel nous avons travaillé et la méthode d'extraction des hésitations autonomes que nous recherchions. La partie suivante présente les caractéristiques acoustiques des phénomènes analysés et les différences constatées entre les langues. La troisième partie présente une étude perceptive préliminaire portant sur la discrimination des langues à partir d'hésitations. Nous présenterons enfin les conclusions de ce travail et nos perspectives.

## 2. CORPUS ET MÉTHODE

Les hésitations sont issues d'un corpus multilingue contenant 3 heures d'enregistrement par langue. Le corpus multilingue est représenté par des émissions journalistiques en Français, Espagnol, Italien, Portugais, Anglais, Allemand, Arabe et Chinois mandarin. Les corpus en Français et Arabe représentent des ressources DGA en partie disponibles chez ELDA. Les corpus en Anglais, Espagnol et Chinois mandarin représentent des extraits des corpus LDC Hub4, tandis

que ceux en Allemand, Portugais et Italien représentent des ressources diverses acquises via des projets européens FP5 (OLIVE, ALERT) ou chez ELDA et disponibles au LIMSI. De cet ensemble initial, un sous corpus d'hésitations a été extrait de façon semi-automatique selon des critères de durée et d'autonomie. Ainsi, une durée minimale de 200ms a été choisie comme seuil inférieur et uniquement des items isolés du contexte par des silences ont été sélectionnés afin d'éviter au mieux des mots allongés. De ce fait, 30 à 200 occurrences prononcées par des locuteurs hommes et femmes ont été relevées par langue. Le nombre d'échantillons obtenus n'est pas le même pour toutes les langues, néanmoins ce corpus nous permet d'explorer les hypothèses mentionnées ci-dessus.

Le logiciel PRAAT[1] a été utilisé pour extraire les paramètres acoustiques fréquence fondamentale (F0) et les premiers deux formants (F1, F2). Les tests perceptifs ont été présentés aux auditeurs via le logiciel E-PRIME[2].

## 3. ANALYSE ACOUSTIQUE

Trois paramètres ont été pris en compte pour l'analyse acoustique : la distribution de la voyelle d'hésitation dans l'espace F1/F2 (le timbre de la voyelle d'appui), la hauteur globale de l'hésitation (F0) et la durée (ms). Parmi ces trois paramètres, le premier permettrait d'identifier la langue, tandis que les deux derniers serviraient plutôt à localiser l'hésitation dans la parole.

### *3.1. Fréquence fondamentale et durée des hésitations vocaliques autonomes*

En ce qui concerne les valeurs absolues en Hz de la F0 moyenne des hésitations, aucune différence significative n'a été constatée entre les langues. Nous estimons que cette piste pourrait sans doute donner des résultats intéressants sur un nombre plus important de données et si l'on prend en compte des mesures plus précises portant sur les différences de hauteur entre les hésitations et leur contexte et selon des variables telles que la langue, le genre, le type de corpus utilisé.

Les conclusions sur la durée vont dans le même sens : il n'y a aucune différence significative entre les langues, les durées confirment la tendance qui a été observée dans les études précédentes sur le français ou l'anglais [3], [5], [6] à savoir que la durée, avec une moyenne supérieure à 400ms, est aberrante par rapport à la durée habituelle des voyelles et ce, quelle que soit la langue. La procédure, basée sur la durée, utilisée par Shriberg pour reconnaître automatiquement les hésitations en anglais pourrait donc fonctionner de manière satisfaisante pour d'autres langues.

### *3.2. Premier et deuxième formant des hésitations vocaliques*

L'analyse acoustique de la distribution des voyelles dans un espace bidimensionnel dessiné par F1/F2 s'est révélée bien plus intéressante. Les mesures ont été faites sur la voyelle principale de chaque hésitation. Compte tenu de la variabilité inter et intra langues de la structure des hésitations autonomes, pouvant aller d'une seule voyelle jusqu'à des structures plus complexes (diphtongaison, segments consonantiques adjacents – ex. *hum* en anglais, etc.), nos mesures ont pris en compte uniquement la voyelle la plus longue et la plus stable de chaque occurrence. Ce segment a été considéré comme élément principal d'une hésitation, i.e. sa *voyelle d'appui*. Nous avons ainsi pris en compte la valeur moyenne des deux premiers formants de la voyelle d'appui de chaque hésitation.

La Figure 1 illustre la dispersion des voyelles d'appui pour les 8 langues du corpus et pour tous les échantillons sélectionnés pour notre analyse.

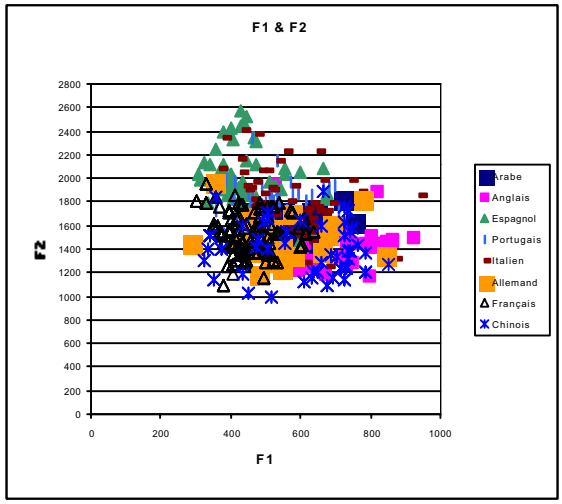


**Figure 1 :** Distribution des voyelles d'appui des hésitations dans l'espace F1/F2 : toutes langues confondues, locuteurs hommes/femmes.

La dispersion des valeurs inter-langues est plus importante que la dispersion intra-langue. Le nuage permet de remarquer que les langues se regroupent autour de centres de concentration différents et aussi que les voyelles utilisées sont loin d'être toujours des voyelles centrales mi-fermées.

Afin d'éliminer les différences dues à la variable « genre », nous avons séparé les données selon que les locuteurs étaient des hommes ou des femmes. Les figures 2 et 3 présentent les données moyennes par genre.

[1] Voir références sur www.praat.org.

[2] Voir références sur www.pstnet.com/e-prime.

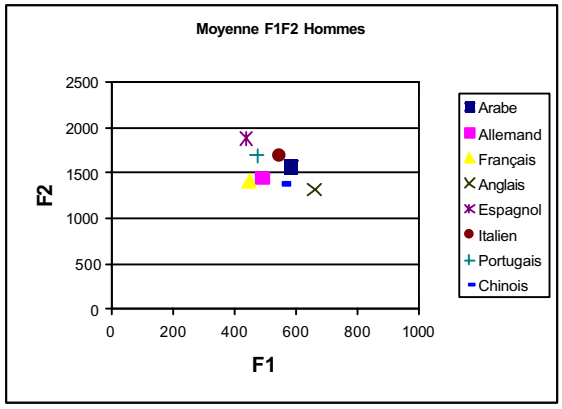


**Figure 2 :** Distribution des voyelles d'appui des hésitations dans l'espace F1/F2 : moyennes, hommes.

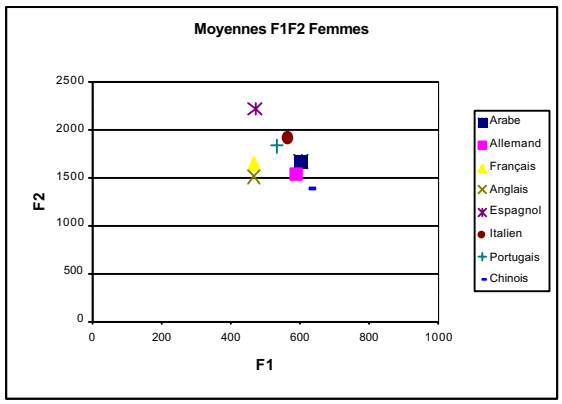


**Figure 3 :** Distribution des voyelles d'appui des hésitations dans l'espace F1/F2 : moyennes, femmes.

Les graphiques montrent que les voyelles utilisées par les langues étudiées sont parfois très différentes et qu'elles se placent sur deux axes, central et antérieur, allant de [æ] à [e] en passant par le [ɐ]. Le français et l'anglais utilisent des voyelles de timbre très proche (pour les femmes) et le français et l'allemand sont très proches pour les hommes. Le timbre utilisé par l'anglais oscille entre le [a], [æ] et [ə], ce qui confirme les observations de Shriberg [6]. Le français en revanche est très stable autour du [ɐ], tout comme l'espagnol autour du [e].

Nous constatons par conséquent de fortes différences entre les langues et nous pouvons faire l'hypothèse que la voyelle utilisée pour les hésitations autonomes est dépendante du système vocalique de chaque langue et ne représente pas une hypothétique « position de repos » qui serait plus ou moins la même pour tous les humains. Néanmoins les données préliminaires obtenues semblent écarter la possibilité d'une voyelle d'hésitation isolée très fermée ou très postérieure (comme [i] ou [u]). Cette observation privilégierait l'hypothèse d'une préférence des lieux d'articulations proches du centre du triangle vocalique. Le nombre de langues ainsi que celui d'échantillons par langue utilisés dans cette étude préliminaire ne nous permettent pas de l'affirmer avec certitude.

En dehors des différences acoustiques relevées en prenant en compte la voyelle d'appui, nous avons également constaté des différences de structure : par exemple, en anglais, la voyelle d'appui est souvent diphtonguée ou suivie par un [m] allongé (ce qui corrobore les remarques de Shriberg [6]), en portugais elle est souvent diphtonguée.

A la suite de cette analyse multilingue exploratoire, nous avons fait l'hypothèse qu'il existe des différences acoustiques significatives entre les hésitations vocaliques autonomes des 8 langues sélectionnées et que ces différences, une fois modélisées, pourraient servir d'indice d'identification de la langue.

## 4. ETUDE PERCEPTIVE

Un test perceptif a été mené parallèlement afin de vérifier si les humains sont capables de discriminer les langues à partir d'hésitations sans support lexical, hors contexte.

### *4.1. Expérience et sujets*

Un sous-ensemble d'hésitations sélectionnées dans les huit langues du corpus a été extrait du corpus initial. Le test a demandé à des auditeurs français de faire la distinction entre les hésitations de leur langue maternelle et celles de sept langues restantes.

Plus précisément, sept sous-tests appariant le français à chacune des autres langues du corpus multilingue (L2) ont été présentés à 20 sujets. Dans une première phase, une brève familiarisation avec le type d'enregistrement a été proposée aux auditeurs. Elle consistait en une dizaine de secondes de parole de journaux télévisés par langue, incluant quelques occurrences d'hésitations et prononcées par deux locuteurs par langue, homme et femme. Ensuite, dans la phase de test, les sujets devaient identifier la langue en écoutant une hésitation présentée isolément. Plus précisément, ils devaient décider s'il s'agissait du français ou de L2 après écoute de chacune des 24 hésitations (12 en français et 12 en L2) présentées en ordre aléatoire. Les hésitations étaient prononcées par des locuteurs hommes et femmes autres que ceux de la phase de familiarisation. De plus, les hésitations en français variaient d'un test à l'autre afin d'éviter des effets d'apprentissage.

### *4.2. Résultats*

Les réponses ont été analysées en termes de pourcentages de réussite. Ces résultats pour chaque sous-test sont présentés dans le graphique suivant. Ils sont significativement au dessus du hasard, les scores d'identification correcte allant de 75% (Français/Allemand) à 96% (Français/Anglais).

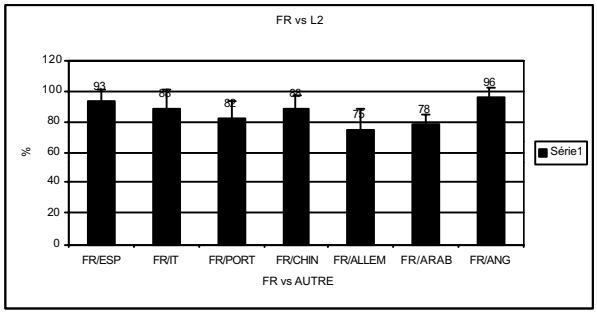


**Figure 4 :** Réussite en discrimination des langues à partir d'hésitations pour les 7 tests Français vs L2.

Les résultats semblent corroborer l'hypothèse selon laquelle les différences acoustiques de timbre vocalique sont responsables d'une réussite dans la reconnaissance d'une langue grâce aux hésitations. Ainsi, les meilleurs résultats sont obtenus pour le test Français/Anglais (96%) opposant des stimuli très différents ([ə] vs [æ]/[a] suivi ou non de segments adjacents tel [m]). Ils sont suivis par les résultats pour le sous-test Français/Espagnol (93%) présentant des hésitations avec des voyelles d'appui aux timbres également différents ([ə] vs [e]/[ɛ]). En revanche, le résultat du sous-test Français/Allemand, bien que significativement au dessus du hasard (75%), pourrait s'expliquer par les proximités acoustiques des timbres des voyelles d'appui dans les deux langues. Il en serait de même pour Français/Arabe (78%) et Français/Portugais (82%). Cependant, étant donné que le Français était systématiquement présent dans les tests et que les auditeurs sont tous natifs de cette langue, ils est difficile d'évaluer dans quelle mesure il s'agit d'une réussite due à une discrimination acoustique aisée des deux langues ou d'une performance liée aux compétences dans la langue maternelle des sujets. Nous envisageons donc de continuer cette étude en variant à la fois la variable population et les langues testées afin de mieux circonscrire le poids des paramètres acoustiques. Une sélection des échantillons d'hésitations selon des critères acoustiques permettrait également de mieux contrôler les effets de timbre.

## 5. CONCLUSIONS ET PERSPECTIVES

Ce travail a permis de répondre, au moins partiellement, à la question de départ : il semblerait que les hésitations autonomes soient porteuses d'information spécifique à la langue. Les performances moyennes de discrimination entre deux langues sont de 85% (significativement au dessus du hasard). Nous avons pu observer que les langues les mieux discriminées ont en général les voyelles les plus éloignées dans le plan F1/F2. D'autres facteurs interviennent probablement : diphtongaison, coda nasale.

Nos résultats plaident pour l'intérêt de construire des modèles d'hésitations autonomes différents. Cette hypothèse est intéressante car, si elle se vérifiait, elle permettrait d'utiliser des modèles d'hésitation vocalique différents en fonction des langues pour améliorer la performance des systèmes automatiques dans une tâche d'identification de la langue sur un corpus spontané non transcrit.

Dans la suite de nos travaux, nous nous proposons d'affiner nos hypothèses en étudiant le lien entre la voyelle d'appui de l'hésitation autonome et le système vocalique de chaque langue. Il est important de recourir à l'étude d'un corpus plus important, incluant plus de langues. Enfin nous comptons poursuivre l'étude perceptive en variant la construction des tests pour mieux contrôler d'une part l'effet de la langue maternelle des sujets sur leurs performances et d'autre part le rôle des particularités acoustiques des stimuli vocaliques utilisés.

**Remerciements** :

Nous remercions Cédric Gendrot (ILPGA, Paris 3) pour son aide dans l'analyse acoustique de notre corpus.

## BIBLIOGRAPHIE

[1] M. Adda-Decker, B. Habert, C. Barras, G. Adda, P. Boula de Mareuil, P. Paroubek, A Disfluency study for cleaning spontaneous automatic transcripts and improving speech language models, Proc.DISS'03, Göteborg, Sweden (Papers in Theoretical Linguistics 90, pp. 67-70), 2003.

[2] J.Bear, J.Dowding & E.E.Shriberg, Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. Proc. Annual Meeting of the Association for Computational Linguistics, pp. 56-63, Newark, Delaware, 1992.

[3] M. Candea, Contribution à l'étude des pauses silencieuses et phénomènes dits "d'hésitation" en français oral spontané. Etude sur un corpus de récits en classe de français. [Thèse de doctorat, Univ. Paris 3], 2000.

[4] I. Guaïtella, Hésitations vocales en parole spontanée : réalisations acoustiques et fonctions rythmiques, *Travaux de l'Institut de Phonétique d'Aix*, vol.14, pp. 113-130, 1991.

[5] E. Shriberg, Phonetic consequences of speech disfluency, ICPhS'99, San Francisco, 1999.

[6] E. Shriberg, To 'errrr' is human: ecology and acoustics of speech disfluencies, *Journal of the International Phonetic Association*, 31/1, 2001.