Influence de l'expressivité sur le triangle vocalique

Grégory Beller

IRCAM (Institut de Recherche et Coordination Acoustique Musique
1 place Igor Stravinsky, 75004 Paris
beller@ircam.fr
http ://www.ircam.fr/anasyn/beller

ABSTRACT

In this paper, we present a study on the influence of expressivity on the vocalic triangle. The system used in this study has been designed for artistic purposes such as cinema, theater and contemporary music. It involves a relational database containing expressive and neutral French utterances. A formant estimation shows that vocalic triangle surface is influenced by expressivity. A new effect is introduced called articulation effort which seems to be related to the activation degree of a multi dimensional representation of emotions. It is shown that this effect could supplant the high speech rate effect on target undershoot of neutral vowels if the speaker expresses a very active expressivity.

1. INTRODUCTION

Dans nos précédents travaux, un synthétiseur musical par concaténation d'unités a été élaboré et nommé *Caterpillar* [18]. Cette plateforme de travail a été élargie pour devenir un synthétiseur vocal de haute qualité baptisé *Talkapillar* [3]. L'un des buts de ce synthétiseur est la reconstruction de la voix d'un locuteur spécifique, celle d'une célébrité défunte, par exemple. *Talkapillar* doit prononcer de nouveaux textes comme s'ils avaient été prononcés par un locuteur cible. Ce système est donc tourné vers la manipulation et l'analyse de la voix à des fins artistiques. Par exemple, des compositeurs de musique contemporaine sont intéressés par l'influence des émotions sur la voix et aimeraient aisément explorer, à l'instar des phonéticiens, de grandes bases de données de parole expressive. Un studio de doublage souhaite utiliser une synthétiseur expressif pour le cinéma. Des metteurs en scène désirent transformer et synthétiser des voix dans des pièces de théâtre afin de jouer sur l'expressivité et l'identité de la voix.

Dans cette étude, nous avons enregistré un acteur français afin de construire une base de données de parole expressive. Ces données ont été analysées sous plusieurs angles, dont l'un a consisté en une estimation de formants. Après l'explication de ce qui a conduit nos travaux vers de telles analyses, cet article présente la constitution d'une base de données, l'algorithme utilisé pour l'estimation du triangle vocalique, et des résultats concernant une éventuelle corrélation entre ce dernier et l'expressivité.

La synthèse par concaténation d'unités provenant d'une large base de données, également nommée synthèse par corpus [13], est aujourd'hui employée par de nombreux synthétiseur Texte-Parole (TTS : Text to Speech Synthesis) [16]. Récemment, le développement de ces méthodes allié à la taille grandissante des bases de données a permis la synthèse de parole expressive (ESS : Expressive Speech Synthesis). Black [4] a enregistré plusieurs corpus d'expressivité différente afin d'utiliser les méthodes TTS classiques sur des corpus séparés. Bulut [6] a essayé de grouper plusieurs corpus expressifs dans la même base de données sans séparation formelle. Ce genre d'approche nécessite une modélisation de l'expressivité par des descripteurs acoustiques fiables semblables à ceux utilisés par les systèmes de reconnaissance des émotions.

Une première phase dans ce type de caractérisation a été réalisée et a consisté en l'observation des modifications prosodiques induites par les changements d'expressivité [2, 15]. Les variations significatives de la fréquence fondamentale, de l'énergie et du débit syllabique ont été analysées, modélisées et utilisées en synthèse et en transformation. Cette première phase a montré qu'une modification de la qualité vocale était nécessaire afin d'obtenir un résultat plus réaliste. Une seconde phase complémentaire a donc été initiée et vise à décrire les variations de la qualité vocale, induites par différentes expressivités. La majorité des études analysant la qualité vocale s'accorde sur la nécessité d'une séparation source filtre du signal de parole afin d'extraire des mesures (quotient ouvert, asymétrie) de la forme d'onde de la dérivée du débit glottique [7, 12, 11]. Cette approche nécessite des estimations conjointes ou successives des caractéristiques du conduit vocal et de celles de la source glottique. Nous avons débuté par la modélisation du filtre constitué par le conduit vocal par l'estimation des formants.

Les résultats de ce type d'analyse sont exposés dans cet article et concernent l'influence de l'expressivité sur le triangle vocalique et sur le débit syllabique. Une différence majeure existe entre le cas neutre et les autres expressivités puisque la surface du triangle vocalique n'est plus seulement fonction de la seule variable débit, mais aussi de ce que nous nommons "l'effort d'articulation". Nous tentons finalement de relier cet effort d'articulation au degré d'activation (passif/actif) d'une représentation multi dimensionnelle des émotions.

2. BASE DE DONNÉES

Pour cette étude, nous avons constitué une base de données de parole expressive d'environ 1H30. Pour cela nous avons enregistré un comédien français d'une quarantaine d'année dans une chambre anéchoïque. Le matériau est composé de 26 phrases de tailles variables répétées chacune avec les expressivités suivantes : *Neutre, colère, joie, peur, tristesse, ennui*, dégoût, indignation, surprise

positive, surprise négative et interrogation neutre. Les expressivités en italique ont été répétées trois fois avec un degré d'intensité différent : *faible*, *moyen* et *fort*. 539 phrases ont été retenues par l'acteur après une étape de post-sélection et ont été segmentées en unités labelisées [1] : Semiphones, phones, diphones, syllabes, groupes prosodiques et phrases. Les enregistrements de qualité CD ont été ensuite analysés : f0, débit syllabique, énergie, apériodicité et formants (voir section 3). Toutes ces informations ont été synchronisées et centralisées dans une base de données relationnelle accessible via une interface graphique Matlab [2].

3. ESTIMATION DES FORMANTS

3.1. Hypothèses

L'algorithme employé afin d'estimer les caractéristiques des formants est dérivé des travaux de Murthy [14], ainsi que d'autres [5, 19, 8] s'appuyant sur le délai de groupe . Trois hypothèses majeures ont été faites :

– Hyp_1 : Les formants correspondent à des pôles d'importance si l'on modélise l'enveloppe spectrale par un système AR.
– Hyp_3 : Ces pôles peuvent être classés selon des régions fréquentielles à priori et selon leurs places respectives les uns par rapport aux autres.
– Hyp_2 : Les trajectoires de formants possèdent une certaine continuité dans le plan temps-fréquence.

3.2. Appartenance d'un pôle à un formant

La première étape est une quasi-dérivation du signal découpée en N trame. Le but de cette pré-accentuation est de tenter d'éliminer les effets spectraux dus aux pentes de la source (-12 dB/octave) et du rayonnement aux lèvres (+ 6 dB/octave) [11]. Puis une analyse linéaire prédictive (LP) d'ordre P (P = 80) du signal filtré est effectuée. On évalue les racines de ce polynôme, constituant les P pôles de l'enveloppe spectrale pour chaque trame n. Pour chaque pôle p de la trame n, on mesure :

– F(p) : la fréquence correspondante (angle du pôle)
– Q(p) : la largeur de bande (proximité du pôle au cercle unité)
– Gd(p) : le délai de groupe du polynôme LPC à la fréquence du pôle
– A(p) : l'amplitude du polynôme LPC à la fréquence du pôle.

Ces grandeurs caractéristiques des pôles sont normalisées par rapport à l'horizon temporel correspondant à la phrase entière. Un poids (entre 0 et 1) est attribué à chacune de ces grandeurs caractéristiques : W_A, W_{Gd} et W_Q.

La probabilité d'observation d'un pôle p à la trame n est la somme pondérée de ses caractéristiques (Hyp_1) : L'information de continuité de la trajectoire temporelle d'un formant est réalisée grâce à une matrice de probabilité de transition (Hyp_3), symétrique et circulaire (de Toeplitz) décrite dans [10].

3.3. Trajectoire d'un formant

Les trajectoires des formants sont décodées une à une grâce à un algorithme de Viterbi récursif qui prend en compte les N trames de la phrase. La trajectoire du premier formant est estimée dans une première zone fixée à priori (Hyp_2). Les pôles correspondant à ce premier formant sont ensuite éliminés de la matrice d'observation. Puis la trajectoire du second formant est évaluée dans une seconde zone fixée à priori et ainsi de suite (voir figure 1). Une tentative d'estimer la densité de probabilité conjointe de tous les formants a échoué à cause de la complexité à définir la matrice de transition.

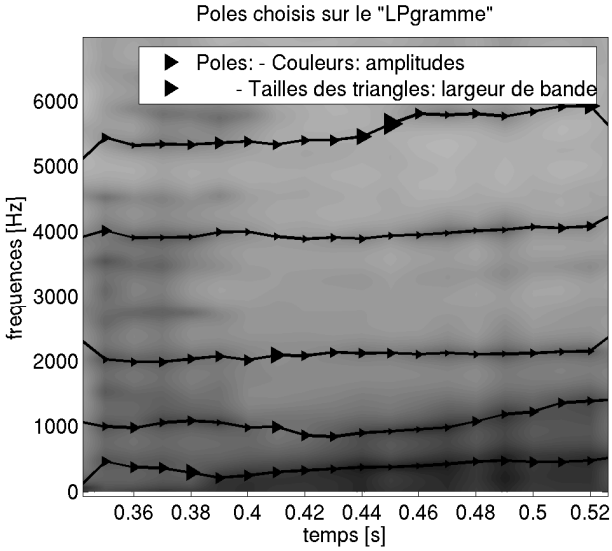

FIG. 1: Exemple d'estimation des trajectoires des cinq premiers formants du diphone [Oa] (X-Sampa)

3.4. post-filtrage des données

Toutes les analyses importées dans la base de données sont modélisées temporellement pour chaque unité par des valeurs caractéristiques :

– Moyennes arithmétiques et géométriques, écart-types
– Valeurs minimum, maximum, initiale, finale et écarts absolus
– Valeurs des trois premiers coefficients de Legendre de l'interpolation polynômiale du 2^{nd} ordre
– Centre de gravité temporel donnant l'instant de la plus importante dépression ou élévation de la courbe
– Spectre normalisé dans cinq bandes ainsi que les quatre premiers moments du spectre (utilisés pour mesurer les indices Jitter et Shimmer)

Afin de tenir compte des phénomènes de coarticulation, la valeur retenue du formant pour une voyelle correspond à une moyenne locale calculée autour de l'instant ou la dérivée du polynôme d'interpolation du 2^{nd} ordre s'annule. Cette mesure est censée mieux refléter la cible visée par le locuteur lors de la prononciation de la voyelle et possède une variance inférieure à celle de la moyenne calculée sur tout l'horizon temporel de l'unité "phone".

4. DISCUSSION

4.1. Généralité des résultats

Les résultats que nous allons exposés sont bien sûr relatifs à nos données, elles mêmes influencées par les stratégies employées par l'acteur pour simuler les expressivités. Ces stratégies se manifestent parfois par des traits acoustiques très différents : Alors que la littérature confère à la tristesse une moyenne de la fréquence fondamentale (f0) basse, l'acteur l'exprime en prenant une voix de tête dont la

moyenne de la f0 dépasse l'octave par rapport au neutre. Ces résultats ne sont donc pas généraux puisqu'ils sont issus d'une étude sur un seul acteur.

4.2. Influence du débit sur le triangle vocalique dans le cas neutre

L'étude de Gendrot et al. [9] sur l'influence du débit sur le triangle vocalique montre que les formants tendent vers une voyelle centrale pour les segments de courte durée. Ceci suggère que la réduction n'est pas un phénomène exclusivement linguistique, mais admet aussi une cause d'ordre physique ou physiologique. Or les émotions sont liées à des modifications sur les plans physique et physiologique. C'est pourquoi elles aussi, peuvent influencer ce phénomène de réduction/expansion du triangle vocalique.

5. RÉSULTATS

5.1. Présentation

Les figures 2 et 3 montrent les triangles vocaliques dans le plan [fréquence du 2^{nd} formant/fréquence du 1^{er} formant] pour différentes intensités de l'expressivité concernée. Les voyelles y sont représentées par des ellipses dont les coordonnées du centre sont définies par la moyenne des moyennes locales (voir partie 3.4) et dont les largeurs X et Y représentent les variances respectives de ces mesures.

5.2. Réduction/expansion du triangle vocalique

Influence de la joie : La figure 2 présente quatre triangles vocaliques superposés et mesurés dans le cas neutre (le plus petit) et dans le cas de la joie, pour ses trois degrés d'intensité différents (joie faible, moyenne et forte). Elle montre que le triangle vocalique a tendance à s'élargir au fur et à mesure que la joie est simulée de manière intense. On y observe aussi que la fréquence du 1^{er} formant augmente au fur et à mesure que l'intensité augmente. Ce phénomène peut être relié à l'augmentation simultanée de la f0 non représentée sur la figure (d'un peu plus d'une octave pour la joie forte). Une autre information manquante est que le débit de parole ralentit au fur et à mesure que l'intensité est grande. La corrélation entre débit et taille du triangle vocalique semble donc être respectée dans le cas de la joie puisque cette dernière est plus grande pour un débit plus lent.

Influence de la colère : En revanche, cette corrélation n'est plus respectée dans le cas de la colère. La figure 3 est l'équivalent de la figure 2 dans le cas de la colère. Elle montre une même expansion du triangle vocalique au fur et à mesure que l'intensité augmente. Cependant et contrairement à la joie, le débit syllabique a tendance à accélérer en fonction de l'intensité. Si cela va à l'encontre du phénomène explicité dans la partie 4.2 dans le cas de l'expressivité neutre, c'est pour une motivation extérieure induite par la stratégie de l'acteur afin d'exprimer la colère (voir section 5.4).

5.3. Triangle vocalique et débit syllabique

Ce phénomène peut être observé pour d'autres expressivités. Ainsi la tristesse et l'ennui (avec moins d'ampleur), deux expressivités dont le débit est plus lent que le cas neutre, montrent une réduction du triangle vocalique d'autant plus forte qu'elles sont exprimées intensément. Ceci est visible sur la figure 4, dans laquelle ont été représentées toutes les expressivités enregistrées, en fonction de l'aire couverte par le triangle vocalique (en abscisse) et de la moyenne du débit syllabique (en ordonnée). La peur, la colère, la joie, l'ennui et la tristesse y sont représentées par des droites reliant les états de faible intensité (petites croix) aux états de forte intensité (grand cercle). L'accélération du débit dans le cas de la peur produit une réduction du triangle vocalique accentuée par rapport à l'accélération du débit dans le cas du neutre [9].

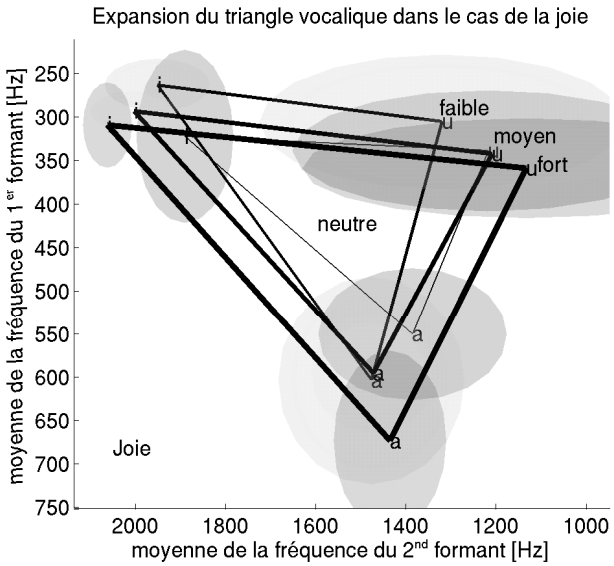

FIG. 2: Triangle vocalique neutre et selon trois niveaux d'intensité de la joie

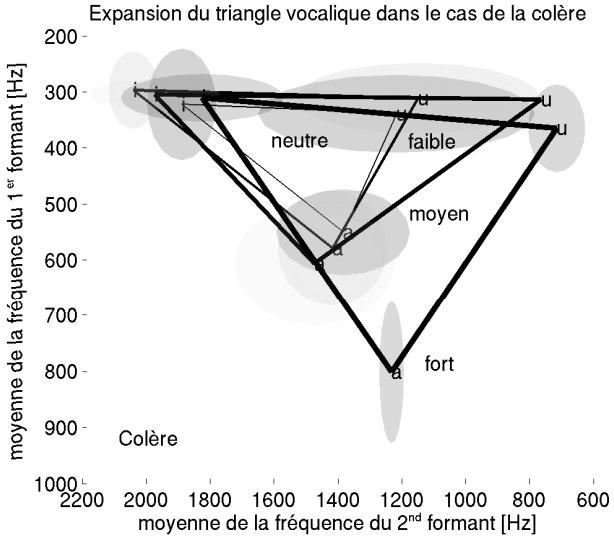

FIG. 3: Triangle vocalique neutre et selon trois niveaux d'intensité de la colère

5.4. Lien entre degré d'activation et effort d'articulation

La réduction/expansion du triangle vocalique n'est plus fonction unique du débit lorsque l'on sort de l'expressivité neutre. Nous pensons qu'un facteur supplémentaire changeant selon l'expressivité s'ajoute à cette dépendance et l'appelons "effort d'articulation" réalisé par l'acteur. Il est

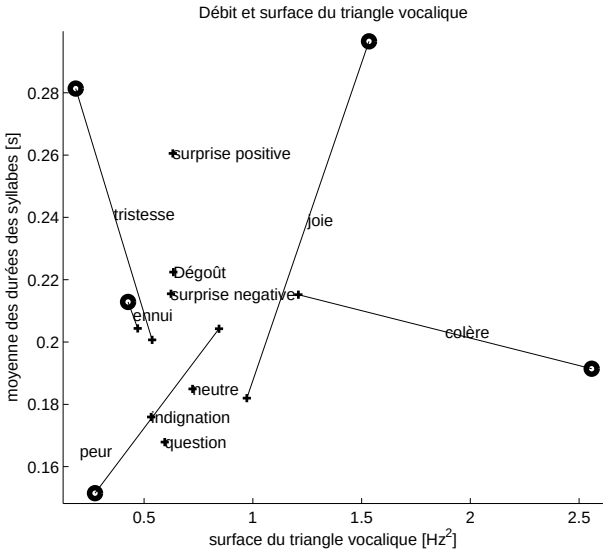

FIG. 4: Triangle vocalique neutre et selon trois niveaux d'intensité de la joie

intéressant de relier cet effort d'articulation au degré d'activation d'une représentation dimensionnelle des émotions décrite par Schröder [17]. En effet, une réduction du triangle vocalique pour l'ennui et la tristesse malgré une diminution de débit peut-être due au fait que ces deux émotions sont jugées à activation négative (ce qui traduit la passivité). Au contraire, une expansion du triangle vocalique malgré une accélération du débit serait visible pour les expressivités jugées à activation positive (traduisant l'action) comme la colère. Ces hypothèses demandent à être validées par un test d'écoute impliquant l'enregistrement de plusieurs acteurs.

6. REMERCIEMENTS

Nous tenons à remercier singulièrement l'acteur français Jacques Combe pour sa performance.

7. CONCLUSION

Dans cet article nous avons présenté nos motivations pour l'analyse de la voix parlée expressive. Ces motivations nous ont amené à constituer une base de données de parole expressive. Un algorithme d'estimation de formants a été employé afin d'analyser l'influence des expressivités sur le triangle vocalique. Plusieurs résultats concernant la joie, la colère et d'autres expressivités ont été présentés et ont permis une constatation globale : Elles se différencient par un effet visible de réduction/expansion du triangle vocalique. Cet effet est partiellement corrélé à la variation de débit mais semble dépendre d'un facteur supplémentaire appelé effort d'articulation.

RÉFÉRENCES

[1] Michel Bagein, Thierry Dutoit, Nawfal Tounsi, Fabrice Malfrère, Alain Ruelle, and Dominique Wynsberghe. Le projet EULER, Vers une synthèse de parole générique et multilingue. *Traitement automatique des langues*, 42(1), 2001.

[2] Grégory Beller. Etude et modèle génératif de l'expressivité dans la parole. Master-2 sar-atiam, Paris 6, IRCAM, Paris, 2005.

[3] Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet. Hybrid concatenative synthesis in the intersection of speech and music. *JIM*, 12 :41–45, 2005.

[4] A.W. Black. Unit selection and emotional speech. *Eurospeech*, 2003.

[5] Baris Bozkurt and Laurent Couvreur. On the use of phase information for speech recognition. In *EUSIPCO*, 2005.

[6] M. Bulut, S. Shrikanth, S.S. Narayanan, and A. K. Syrdal. Expressive speech synthesis using a concatenative synthesizer. In *ICSLP*, ATT Labs-Research, Florham Park, NJ, 2002.

[7] Boris Doval, Nicolas d'Alessandro, and Nathalie Henrich. The voice source as a causal/anticausal linear filter. In *VOQUAL*, August 2003.

[8] G. Duncan, B. Yegnanarayanan, and Hema A. Murthy. A non parametric method of formant estimation using group delay spectra. In *IEEE*, 1989.

[9] Cédric Gendrot and Martine Adda-Decker. Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande. In *MIDL*, 2004.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[11] Nathalie Henrich. *Etude de la source glottique en voix parlée et chantée*. PhD thesis, Université Paris 6, Paris, France, nov 2001.

[12] Lu Hui-Ling. *Toward a High Quality Singing Synthesizer with Vocal Texture Control*. PhD thesis, Stanford University, Jul 2002.

[13] Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, Atlanta, GA, May 1996.

[14] Hema A Murthy, K. V. Mahu Murthy, and B. Yegnanarayana. Formant extraction from phase using weigthed group delay function. In *IEEE*, volume 25, pages 1609–1611, July 1989.

[15] C. Pereira and C. Watson. Some acoustic characteristics of emotion. In *Fifth International Conference on Spoken Language Processing, Sydney*, 1998.

[16] Romain Prudon and Christophe d'Alessandro. A selection/concatenation TTS synthesis system : Databases developement, system design, comparative evaluation. In *4th Speech Synthesis Workshop*, Pitlochry, Scotland, 2001.

[17] Marc Schröder. *Speech and Emotion Research : An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. PhD thesis, University of Saarland, 2003.

[18] Diemo Schwarz. New Developments in Data-Driven Concatenative Sound Synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 443–446, Singapore, October 2003.

[19] Donglai Zhu and Kuldip K. Paliwal. Product of power spectrum and group delay function for speech recognition. In *ICASSP*, 2004.