Influence de l'expressivité sur le degré d'articulation

Grégory Beller

IRCAM - Institut de Recherche et Coordination Acoustique Musique
1 place Igor Stravinsky, 75004 Paris
beller@ircam.fr

ABSTRACT

In this paper, we present a study on the influence of expressivity on the articulation degree. It is a part of a project that aims to transform expressivity of an utterance for artistic purposes such as cinema, theater and contemporary music. It involves an expressive French database. An algorithm for formant frequency estimation is presented. A measure of the articulation degree is proposed. It involves a joint statistical analysis of the degree of expressive intensity, of the vocalic triangle area and of the speech rate, and shows that the articulation degree depends on expressivity. Using Lindblom's theoretical framework, it is shown how this measure could be related to the activation degree of a dimensional representation of emotions.

1. INTRODUCTION

Dans nos précédents travaux, nous avons élaboré un système de gestion de bases de données de parole permettant la manipulation de grands corpus pour différents objectifs artistiques. Le premier est l'analyse statistique de variables acoustiques selon le contexte (prosodique, accentuel, phonétique). Ce type d'analyse a permis, entre autres, d'observer l'influence de l'expressivité sur le débit de parole [Bel06]. Le second objectif réside dans la synthèse vocale de haute qualité à partir du texte. L'un des buts d'un tel synthétiseur est la reconstruction de la voix d'un locuteur spécifique, celle d'une célébrité défunte, par exemple. Une extension a permis la synthèse musicale et la synthèse hybride parole/musique [Bel05] pour des compositeurs de musique contemporaine. Enfin, la dernière utilisation de ce système est la transformation dépendante du contexte de l'expressivité dans la parole pour des metteurs en scène de théâtre et des studios de doublage de cinéma. C'est dans ce cadre que s'inscrit cette étude.

Nous avons enregistré un acteur français simulant un ensemble d'expressivités. Ce terme désigne l'ensemble des émotions simulées ou non, ainsi que des attitudes et des modes de jeu d'acteurs. Ces données ont été segmentées phonétiquement et analysées sous plusieurs angles : La hauteur, l'intensité, le débit de parole, la qualité vocale et le degré d'articulation. L'étude de ces descripteurs acoustiques relatifs aux cinq dimensions de la prosodie [Pfi06] permet d'observer des tendances pour chacune des expressivités. Ces tendances sont ensuite utilisées pour transformer l'expressivité d'une phrase neutre grâce à des algorithmes de transformation du signal de parole. Cet étude concerne spécifiquement l'analyse du degré d'articulation pour différentes expressivités.

Certaines théories sur les émotions [Sch06] représentent celles-ci dans un espace à trois dimensions dont les axes sont la valence (évaluation positif-négatif), l'intensité (peu ou beaucoup) et l'*activation*. Le degré d'activation d'une émotion relate si le locuteur est amené à agir ou à rester passif lorsqu'il est dans un état émotif. Ainsi certaines émotions simulées par les acteurs les orientent vers l'introversion ou l'extraversion. Afin de mesurer le degré d'activation des phrases expressives prononcées par les acteurs, nous relions celui-ci au degré d'articulation de la théorie de Lindblom [lin83]. La théorie "H and H" propose deux degrés d'articulation de la parole : la parole *Hyper* qui s'oriente vers une clarté maximale du signal produit et la parole *Hypo* qui a comme objectif de produire le signal la plus économique possible. Le degré d'articulation renseigne ainsi sur la motivation/personnalité du locuteur vis à vis de ses interlocuteurs et sur son intoversion/extraversion en situation de communication parlée. Cette position peut provenir de plusieurs facteurs contextuels dont l'état émotionnel du locuteur, ou l'expressivité avec laquelle celui-ci s'exprime. En conséquence, nous proposons qu'une mesure du degré d'articulation permet de quantifier le degré d'activation pour différentes expressivités.

Cet article présente les données employées, la méthode développée pour estimer les trajectoires des formants, une mesure des degrés d'articulation d'un corpus expressif et son interprétation vis à vis de la représentation dimensionnelle des émotions.

2. CORPUS EXPRESSIF

Les données employées dans cette étude proviennent d'un corpus français de parole expressive d'environ 1H30. Il est constitué de l'enregistrement d'un comédien français d'une quarantaine d'année, en conditions professionnelles. L'acteur a répété un texte phonétiquement équilibré composé de dix phrases [Com81]. Ce texte est sémantiquement neutre vis à vis de l'expressivité, c'est à dire que chaque phrase possède une signification quelque soit l'expressivité choisie pour la prononcer. Les expressivités retenues pour cette expérience sont : *Neutre, colère (extravertie), joie (introvertie, douce), peur (introvertie, tétanisante), tristesse (extravertie, larmoyante), ennui (introverti)*, dégoût, indignation, surprise positive et surprise négative. Les expressivités en italique ont été répétées trois fois avec un degré d'intensité expressive croissant. 550 phrases environ ont été manuellement segmentées et étiquetées phonétiquement. Ces unités ont permis la mesure du degré d'articulation.

3. MESURE DU DEGRÉ D'ARTICULATION

La mesure du degré d'articulation utilisée est différente de celle employée classiquement [Wou01]. Le degré d'articulation est influencé par : le contexte phonétique, le débit de parole et la dynamique spectrale (qui correspond à la vitesse du changement de configuration du conduit vocal). Ainsi la mesure traditionnelle du degré d'articulation consiste à définir des cibles formantiques pour chaque phonème, en tenant compte de la coarticulation, et à étudier les différences entre les réalisations et les cibles par rapport au débit de parole. Compte tenu de la difficulté à définir les cibles, nous avons opté pour une mesure statistique du degré d'articulation.

La mesure du degré d'articulation proposée nécessite préalablement trois types d'analyse du signal de parole. Tout d'abord, ce dernier doit être segmenté phonétiquement afin de connaître à quelle catégorie phonétique appartient chaque portion (trame) de signal. Puis, une segmentation syllabique permet la mesure dynamique du débit local de la parole [Bel06]. Enfin, l'estimation des trajectoires de formant permet la mesure de l'aire du triangle vocalique (voir partie 4.3). La mesure du degré d'articulation d'une expressivité provient de l'observation conjointe des évolutions de l'aire du triangle vocalique et du débit de parole en fonction de l'intensité de l'expressivité (voir figure 3). La mesure de l'aire du triangle vocalique nécessite une segmentation phonétique ainsi que l'estimation de la fréquence des formants.

4. ESTIMATION DE LA FRÉQUENCE DES FORMANTS

Il existe de nombreux outils permettant l'estimation de la fréquence des formants. La majorité de ces outils modélisent l'enveloppe spectrale du signal découpé en N trames temporelles et fenêtré, par un système autorégressif dont les pôles P correspondent aux résonances du conduit vocal. En filtrant ces pôles selon un ordre d'importance, ils sont capables de définir pour chaque trame (n), un ensemble restreint de pôles candidats $P_{can}(n)$ parmi lesquels certains correspondent aux formants. Mais ces outils n'attribuent pas à ces pôles, un index de formant. C'est à dire qu'ils donnent un ensemble de candidats possibles mais qu'ils n'affectent pas ces candidats à un formant particulier. Or, l'étape d'attribution des pôles aux formants est nécessaire à l'observation du triangle vocalique puisque celle-ci requiert la connaissance de F_1 et F_2.

4.1. Attribution des pôles à des formants

En pratique, il se peut qu'un ensemble de pôles candidats $P_{can}(n)$ soient très différent de celui le précédant $P_{can}(n-1)$. Même le nombre de candidats peut changer d'une trame à l'autre. La figure 1 présente les ensembles de pôles estimés par Wavesurfer [Sjö00] et PRAAT [Boe01]. Si l'on attribueš d'emblée ces ensembles rangés par ordre de fréquence croissante aux formants, on observe des sauts de fréquence importants en ce qui concerne leurs trajectoires. Il suffit que le pôle de plus basse fréquence disparaisse pour que tous les autres se voient affecter à un formant de rang supérieur (exemple sur la figure 1, à la 0,98 seconde). Traditionnellement, l'affectation des ensembles $P_{can}(n)$ aux formants est à la charge de l'utilisateur. Celui-ci trie les pôles selon leurs régions fréquentielles. Par exemple, pour une voix d'homme, un pôle dont la fréquence est située entre 800Hz et 1500Hz sera souvent nommé deuxième formant. Mais cet a priori peut biaiser les résultats si plusieurs formants sont présents dans une région fréquentielle ou si la fréquence d'un formant excède ces limites, ce qui est parfois le cas dans la parole expressive. De plus, la quantité de données utilisées nécessitent une automatisation de l'attribution des ensembles de pôles candidats $P_{can}(n)$ aux formants.

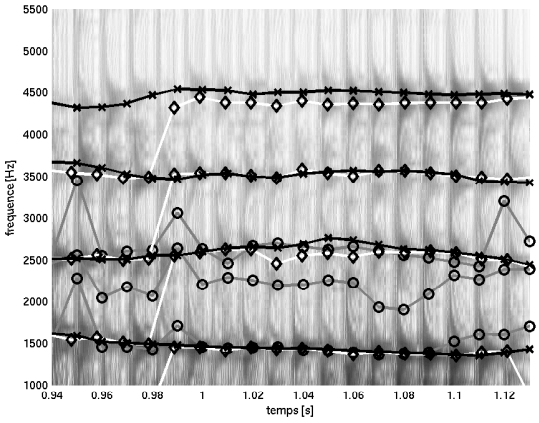

FIG. 1: Trajectoires de formants estimées par Praat gris-cercle), par Wavesurfer (blanc-losange) et par la méthode formant-Viterbi proposée (noir-croix), tracées sur le spectrogramme d'une voyelle [A].

4.2. Algorithme formant-Viterbi

Nous présentons un nouvel algorithme baptisée formant-Viterbi qui attribue à chaque formant, un des pôles candidats, sans aucun à priori sur sa région fréquentielle, et qui prend simultanément en compte une contrainte de continuité sur la trajectoire du formant.

Hypothèses Cet algorithme repose sur trois hypothèses :
- Hyp_1 : Les formants correspondent à des pôles proéminents de la modélisation de l'enveloppe spectrale par un système autorégressif.
- Hyp_3 : Ces pôles peuvent être classés selon des règles de proéminence et selon leurs places respectives les uns par rapport aux autres.
- Hyp_2 : La trajectoire d'un formant possèdent une certaine continuité dans le plan temps-fréquence.

La contrainte de continuité des trajectoires permet en pratique, de diminuer le bruit de l'estimation trame à trame.

Appartenance d'un pôle à un formant La première étape est une quasi-dérivation du signal découpée en N trames et fenêtré. Puis une analyse linéaire prédictive (LP) 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 d'une trame n, on mesure :
- F(p,n) : la fréquence correspondante (angle du pôle)
- Q(p,n) : la largeur de bande (proximité du pôle au cercle unité)
- Gd(p,n) : le délai de groupe du polynôme LPC à la fréquence du pôle [Mur89]
- A(p,n) : l'amplitude du polynôme LPC à la fréquence du pôle.

Les trois dernières grandeurs caractéristiques des pôles sont normalisées par rapport à l'horizon temporel correspondant à la phrase. Un poids (entre 0 et 1) est attribué à chacune de ces grandeurs caractéristiques. Pour la trame n, la probabilité d'appartenance d'un pôle p à un formant P(p,n), découle de la somme pondérée de ses caractéristiques (Hyp_1). La matrice P(p,n) représente une phrase entière dans le plan temps-fréquence. à un instant n donné, P(p,n) renseigne sur la probabilité que la trajectoire d'un formant passe par une fréquence angle(p) donnée.

Trajectoires de formant La contrainte de continuité de la trajectoire spectro-temporelle d'un formant (Hyp_3) est représentée par une matrice de probabilité de transition T(p, n), de Toeplitz (symétrique et circulaire). Les trajectoires des formants sont "décodées" récursivement, une après l'autre, par un algorithme de Viterbi qui prend en compte les N trames de la phrase. La programmation dynamique permet de tracer une trajectoire de formant sur la matrice P(p,n) tout en respectant la continuité T(p,n) à chaque trame. La trajectoire du premier formant est estimée en initialisant sa fréquence à 0Hz à la première trame (t = 0). Les pôles correspondant à ce premier formant sont ensuite éliminés de la matrice de probabilité d'appartenance des pôles au formant P(p,n) (Hyp_2). Puis la trajectoire du second formant est évaluée de la même façon et ainsi de suite (voir figure 1). La tentative d'estimer la densité de probabilité conjointe de tous les formants en même temps a échoué à cause de la complexité à définir la matrice de transition T(p,n) de tous les formants en même temps.

Fréquence des formants Une de ces trajectoires estimées permet de connaître à chaque trame temporelle, la fréquence du formant correspondant qui évolue pendant la durée d'un phone. Afin de minimiser les erreurs d'estimation et d'obtenir une seule valeur représentative par phone appelée *fréquence caractéristique*, une mesure globale est effectuée sur toutes les trames temporelles de chaque phone, en utilisant la segmentation phonétique. Un polynôme de Legendre d'ordre 2 modélise l'évolution temporelle de la trajectoire d'un formant sur le phone. Si l'évolution de la fréquence est linéaire, la fréquence caractéristique correspond à la valeur médiane des fréquences prises sur quelques valeurs avoisinant le milieu du phone. Si l'évolution de la fréquence est parabolique, ce qui montre que la fréquence d'un formant a atteint une cible puis s'en est écarté (coarticulation), la fréquence caractéristique correspond à la valeur médiane des fréquences prises sur quelques valeurs avoisinant l'instant où la fréquence a atteint sa cible (instant ou la dérivée du polynôme d'interpolation du 2^{nd} ordre s'annule). Cette mesure reflète mieux 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 du phone.

4.3. Triangle vocalique

Le *triangle vocalique* est le nom donné à la figure géométrique que forment les voyelles /a/, /i/ et /u/ (voir figure 2) lorsqu'elles sont placés dans un espace bi-dimensionnel appelé *espace cardinal*, dont les axes sont les fréquences du 1^{er} (F_1) et du 2^{eme} (F_2) formant. Chacune de ces voyelles limitrophes, à expressivité donnée et à une intensité donnée, se voit attribuer les moyennes statistiques des fréquences caractéristiques de tous les phones de leurs classes phonétiques. Pour dix phrases par expressivité et par intensité, chacune de ces moyennes impliquent une vingtaine d'individus. C'est pourquoi nous avons aussi représenter les variances des ces mesures. Enfin, tous les triangles vocaliques sont constitués à partir de voyelles situées dans les mêmes contextes phonétiques, afin d'éliminer une possible variabilité due à la coarticulation.

5. INFLUENCE DE L'EXPRESSIVITÉ

Une étude concernant l'influence du débit sur le triangle vocalique en parole neutre [Gen04], montre que les formants tendent vers une voyelle centrale pour les segments de courte durée. Une différence majeure existe entre le cas neutre et les autres expressivités : Le degré d'articulation n'est plus uniquement dépendant de la variable débit. Dans le cas neutre, une accélération et une décélération correspondent respectivement à une réduction et à une expansion du triangle vocalique. Il semble que cette tendance naturelle ne soit pas préservée dans le cas de certaines expressivités.

5.1. Sur le triangle vocalique

En effet, La figure 2 présente quatre triangles vocaliques superposés et mesurés dans le cas neutre (le plus petit) et dans le cas de la colère extravertie, pour trois degrés d'intensité différents (colère faible, moyenne et forte). Les voyelles y sont représentées par des ellipses dont les coordonnées du centre et les largeurs sont définies respectivement par les moyennes et les variances des fréquences caractéristiques du 2^{nd} et du 1^{er} formant (voir partie 4.3). Cette figure montre une expansion du triangle vocalique au fur et à mesure que l'intensité augmente. Ce qu'elle ne montre pas, c'est que le débit syllabique augmente selon l'intensité. Cette tendance dans le cas de la colère extravertie est donc inverse à la tendance du cas neutre qui proposerait une réduction du triangle vocalique au fur et à mesure que le débit s'accélère.

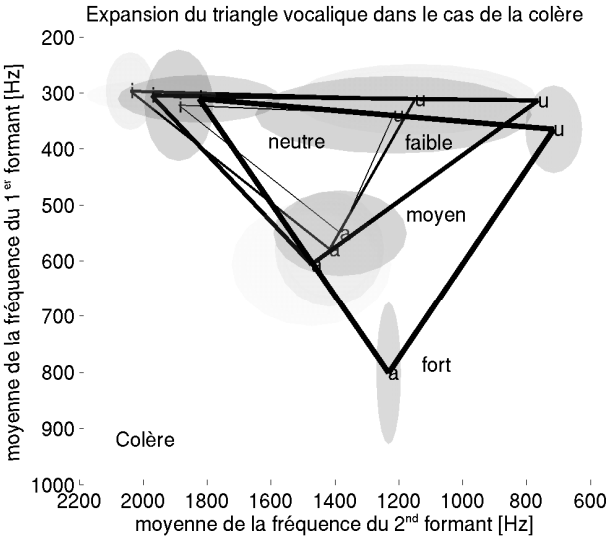

FIG. 2: Triangle vocalique neutre et selon trois niveaux d'intensité de la colère extravertie

5.2. Sur le degré d'articulation

Ce phénomène peut être observé pour d'autres expressivités. Ainsi la tristesse et l'ennui (avec moins d'ampleur) sont deux expressivités dont le débit est plus lent que dans le cas neutre et qui montrent une réduction du triangle vocalique d'autant plus forte qu'elles sont exprimées intensément. Ceci est visible sur la figure 3, 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. Pour une expressivité donnée, la mesure du degré d'articulation est explicitement reliée à sa position dans cet espace par rapport au neutre (référence).

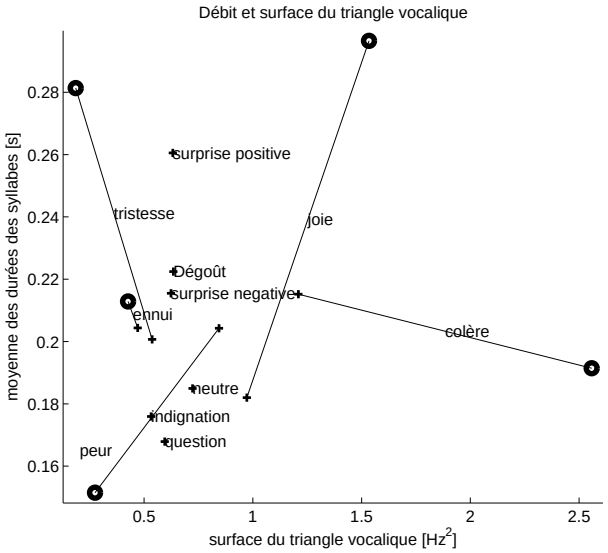

FIG. 3: Représentation des expressivités dans un espace dont l'abscisse est l'aire couverte par le triangle vocalique $[10^4 Hz^2]$ et l'ordonnée est la durée moyenne des syllabes [s]

5.3. Sur le degré d'activation

Les expressivités utilisées possèdent des degrés d'activation différents. Le degré d'activation distingue les émotions nous rendant passifs de celles nous rendant actifs. En pratique, les émotions de degré d'activation négatif (passivité) et positif (activité) se manifestent respectivement par l'introversion et l'extraversion du locuteur qui s'exprime par une parole hypo- et hyper-articulée. Par exemple, l'ennui et la tristesse introvertie dont les degrés d'activation sont négatifs, présentent une réduction du triangle vocalique malgré une diminution de débit. Au contraire, une expansion du triangle vocalique malgré une accélération du débit est visible pour les expressivités jugées à activation positive comme la colère extravertie.

6. CONCLUSION

Dans cet article nous avons présenté nos motivations pour l'analyse de la parole expressive. Après la présentation du corpus utilisé, un nouvel algorithme d'estimation des trajectoires de formant a été décrit est utilisé afin de mesurer l'aire du triangle vocalique. Une étude conjointe de l'évolution de l'aire du triangle vocalique et de l'évolution du débit de parole, en fonction de l'intensité expressive, définit une mesure du degré d'articulation. L'application de cette mesure au corpus expressif met en évidence l'influence de l'expressivité sur le degré d'articulation. Certaines interprétations concernant la joie, la colère et d'autres expressivités permettent de penser que cette mesure est corrélée au degré d'activation d'une représentation dimensionnelle des émotions. Cette étude propose donc une nouvelle méthode de mesure objective et automatique de la dimension passif/actif de l'expressivité grâce à l'étude de l'articulation de la parole.

RÉFÉRENCES

[Bel05] Beller, G., Schwarz, D., Hueber, T. and Rodet, X. (2005), "Hybrid concatenative synthesis in the intersection of speech and music", *JIM*, vol. 12, pp. 41–45.

[Bel06] Beller, G., Schwarz, D., Hueber, T. and Rodet, X. (2006), "Speech rates in french expressive speech", in *Speech Prosody*, SproSig, Dresden : ISCA.

[Boe01] Boersma, P. (2001), "Praat, a system for doing phonetics by computer", in *Glot international, 10*, vol. 5, pp. 341–345.

[Com81] Combescure, P. (1981), "20 listes de dix phrases phonétiquement équilibrées", *Revue d'Acoustique*, vol. 56, pp. 34–38.

[Gen04] Gendrot, C. and Adda-Decker, M. (2004), "Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande", in *MIDL*.

[lin83] lindblom, B. (1983), *Economy of Speech Gestures*, vol. The Production of Speech, Spinger-Verlag, New-York.

[Mur89] Murthy, H., Murthy, K.M. and Yegnanarayana, B. (1989), "Formant extraction from phase using weighted group delay function", in *Electronics Letters*, IEE, vol. 25, pp. 1609–1611.

[Pfi06] Pfitzinger, H. (2006), "Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction", in *Speech Prosody*, H. Hoffmann R. ; Mixdorff, ed., Dresden, no. 40 in Abstract Book, pp. 6–9.

[Sch06] Schröder, M. (2006), "Expressing degree of activation in synthetic speech", in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1128–1136.

[Sjö00] Sjölander, K. and Beskow, J. (2000), "Wavesurfer - an open source speech tool", in *International Conference on Spoken Language Processing*.

[Wou01] Wouters, J. and Macon, M. (2001), "Control of spectral dynamics in concatenative speech synthesis", in *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 30–38.