Paramétrisation de la qualité de voix : EGG vs. filtrage inverse

Nicolas Audibert, Solange Rossato & Véronique Aubergé

Institut de la Communication Parlée
Université Stendhal/INPG/CNRS, Grenoble, France
Mél: {audibert, rossato, auberge}@icp.inpg.fr - http://www.icp.inpg.fr/EMOTION

ABSTRACT

This paper aims at testing on an authentic expressive speech corpus the consistency for characterizing emotional expressions in voice of the Normalized Amplitude Quotient (NAQ) parameter, proposed as the 4th prosodic dimension vs. the Open Quotient (OQ) estimated from inverse filtering vs. the direct electroglottographic (EGG) measurement of glottal parameters. The phonemic influence of the NAQ parameter was first evaluated by matching measure locations with an expert phonetic labeling. Results show a speaker-dependent phoneme effect on NAQ, and seem moreover to indicate a systematic overestimation of NAQ on [n] segments. Moreover F0 measurements used for the calculation of amplitude-based parameters reveal underestimated when compared to EGG-estimated F0 values. No correlation could be found between OQ values extracted from EGG signals and amplitude-based parameters.

1. INTRODUCTION

Selon des critères aussi bien objectifs que subjectifs, la qualité de voix a été rattachée à l'expression vocale des affects, au delà d'informations extralinguistiques telles que l'âge ou le sexe du locuteur. Des études en psychologie, essentiellement basées sur de la parole émotionnelle actée, ont en outre intégré les expressions vocales, notamment la qualité de voix, dans un modèle global de la production des émotions. Scherer et al. [14] affirment ainsi que l'état général de tension des muscles du larynx dépend directement de la réponse émotionnelle, et donc que la qualité de voix participe à l'expression de l'émotion ; de plus, ce modèle prédit l'évolution de paramètres spectraux relatifs à la qualité de voix tels que la pente spectrale. Une augmentation de la pente spectrale est ainsi prédite pour la tristesse, ainsi qu'une diminution pour la colère. Des tests ne faisant varier que la qualité de voix de stimuli synthétisés ont de plus montré l'effet perceptif de la modification des paramètres de la source glottique sur la parole attitudinale et émotionnelle [8].

Laver [12] propose une description globale associant les mouvements des muscles du larynx aux qualités de voix résultantes, décrites en termes subjectifs, et suggère qu'en anglais la *breathy voice* est liée à l'intimité, la *whispery voice* avec la confidentialité et la *harsh voice* avec la colère. Campbell [5] met l'accent sur la corrélation entre le continuum *pressed-breathy* et le degré d'« attention » apporté à la voix, décrit par le logarithme du Quotient d'Amplitude Normalisée (NAQ) proposé par Alku [1].

L'objectif de ce papier est de tester la pertinence du paramètre NAQ pour la caractérisation des expressions émotionnelles : un algorithme de calcul de NAQ, développé par Mokhtari [13], a été appliqué à un corpus phonétiquement équilibré, exprimant diverses expressions émotionnelles authentiques de deux locuteurs, afin de vérifier la robustesse phonémique de ce paramètre de qualité de voix. L'estimation de F0 à partir du signal acoustique, utilisée dans le calcul des paramètres basés sur l'amplitude tels que NAQ, a été comparée à la valeur de référence extraite du signal EGG. Nous nous sommes également intéressés au quotient ouvert (OQ, [12]), qui est calculé de deux façons: (1) par filtrage inverse du signal acoustique, dans le même paradigme d'inversion que pour l'estimation de NAQ, pour extraire une valeur OQ_A [9] ; (2) à partir du signal enregistré par l'électroglottographe (EGG) pour extraire OQ_{EGG}. Cela nous permet de comparer OQ_A et OQ_{EGG} afin d'évaluer les artefacts du paradigme d'inversion.

2. CHOIX DU CORPUS

Plusieurs raisons ont motivé le choix d'un corpus de parole expressive authentique enregistré en laboratoire plutôt que d'un corpus acté. Tout d'abord, la neurophysiologie a montré que les émotions actées *vs.* non actées sont régies par des mécanismes neuraux distincts [6], les émotions actées n'étant pas liées à des changements physiologiques. De plus comme l'ont montré Aubergé et Cathiard [2], l'amusement acté *vs.* non acté peut être discriminé avec un effet inter-juge important. Il est donc impossible de mettre au point une méthode objective d'évaluation de la capacité d'un acteur à simuler fidèlement des expressions émotionnelles authentiques.

D'autre part, certaines analyses acoustiques nécessitent un enregistrement de haute-fidélité qui ne peut être réalisé qu'en conditions de laboratoire [4], ce qui impose de développer des protocoles pour l'induction d'états émotionnels. De plus, le choix d'une telle méthode permet de contrôler le contenu linguistique et phonétique des énoncés grâce à l'usage d'un langage de commandes qui contraint l'expression vocale des sujets. Enfin, cela permet de recueillir sur des énoncés identiques des états émotionnels variés, impliquant ainsi des qualités de voix variées.

Les stimuli utilisés pour cette étude ont donc été extraits d'un corpus de parole expressive authentique mais

contrôlée, composé entre autres d'énoncés monosyllabiques. Les états émotionnels ont été induits chez les sujets grâce à un scénario de Magicien d'Oz, Sound Teacher, implémenté sur une plate-forme dédiée à la mise en place de scénarios d'induction émotionnelle (E-Wiz) [3]. Sound Teacher imite un logiciel à commandes vocales qui propose à l'utilisateur d'apprendre implicitement les voyelles de langues étrangères. Le but est en réalité d'induire des états émotionnels positifs puis négatifs chez les sujets en manipulant leurs performances. Le corpus recueilli consiste en des énoncés monosyllabiques correspondant à des couleurs ([ʀuʒ], [ʒon], [sabl], [vɛʀ], [bʀik]), choisis pour la répartition de leurs voyelles dans l'espace phonologique, ainsi qu'en des occurrences de [paʒ sqivãt].

3. MESURES

Deux locuteurs ont été sélectionnés sur la base de productions émotionnelles claires et comparables, pour un corpus de 373 stimuli et d'une durée utile totale de 204 secondes, enregistré en chambre sourde sur DAT Sony à l'aide d'un micro C1000S AKG. Après extraction du corpus brut des stimuli pertinents, un étiquetage phonétique expert a été effectué. Nombre des productions traitées ont révélé la présence d'un chwa non attendu en fin d'énoncés supposés monosyllabiques (par exemple [ʒonə] au lieu de [ʒon]), les rendant ainsi bisyllabiques. Les chwas ont donc été inclus dans les analyses, au même titre que les autres voyelles. Le signal EGG utilisé a été enregistré à l'aide du laryngographe portable Laryngograph Processor développé par Laryngograph Ltd., relié à la plate-forme d'expérimentation EVA2. La synchronisation des signaux acoustique et EGG a été réalisée à l'aide de bips enregistrés simultanément sur les deux canaux.

3.1. Electroglottographie

L'électroglottographie est une mesure d'impédance qui fournit des informations sur la région de contact des cordes vocales. $F0_{EGG}$ peut être estimée fidèlement à partir du signal EGG. Cette valeur a été calculée sur des trames d'environ 4 périodes par la méthode d'autocorrélation. De plus la durée de la phase ouverte de la pulsation glottique $T1_{EGG}$ est estimée par la méthode d'intercorrélation [10] entre le signal EGG et sa dérivée. L'estimation de cette valeur permet ainsi de calculer le quotient ouvert $OQ_{EGG}=T1_{EGG}.OQ_{EGG}$.

3.2. Analyse acoustique

Les paramètres basés sur l'amplitude de l'onde de débit glottique ont été proposés comme une méthode de caractérisation de la qualité de voix plus robuste qu'à partir de paramètres temporels. En particulier, le Quotient d'Amplitude Normalisé (NAQ) proposé par Alku [1], peut être considéré comme une normalisation du « temps de déclinaison » défini par Fant [7] et s'exprime comme le rapport de l'amplitude crête à crête de l'onde de débit glottique (UP) et du pic négatif maximal de sa dérivée (EE), normalisé par la période fondamentale. Le calcul automatique de NAQ s'effectue grâce à un algorithme développé aux ATR, Japon, dans le cadre du projet du JST/CREST *Expressive Speech Project*. Cet algorithme procède au calcul de NAQ à partir du signal acoustique, sur des segments identifiés automatiquement comme centres de confiance [13]. Ceci permet d'extraire de façon entièrement automatique une mesure de la qualité de voix à partir de parole spontanée non étiquetée.

Gobl et Ní Chasaide [9] ont proposé d'étendre les paramètres basés sur l'amplitude à l'estimation d'autres paramètres temporels. La phase ouverte de la pulsation glottique peut ainsi être estimée par $T_{1A}=\frac{\pi}{2}\frac{UP}{EI}+\frac{UP}{EE}$, où EI est la valeur du pic positif maximum de la dérivée. OQ est alors estimé par T_{1A}.F0. Le calcul de ce quotient ouvert issu de l'amplitude, noté OQ_A, a également été effectué en même temps que celui de NAQ. De plus les estimations de F0 par sommation subharmonique [11] réalisées par cet algorithme à chaque centre de confiance détecté ont été extraites afin de pouvoir être comparées aux mesures de F0 effectuées à partir du signal EGG et par détection de cycles sur le signal acoustique.

4. RESULTATS

4.1. Influence du phonème sur NAQ

Lorsqu'il est calculé à partir de parole continue non étiquetée, NAQ ne peut être obtenu qu'au niveau des centres de confiance, *i.e.* des vocoïdes comme définies par Mokhtari [13]. La localisation de ces centres de confiance a donc également été extraite, et mise en correspondance avec l'étiquetage phonétique du corpus, afin de s'assurer de leur statut de vocoïdes. 68% des centres de confiance sont trouvés dans des voyelles, contre 15% dans des consonnes sonorantes qui satisfont aux critères d'énergie des vocoïdes et 17 % dans d'autres consonnes. La consonne nasale [n], fréquemment détectée comme centre de confiance, a également été prise en compte dans la suite des analyses. Il ressort de la table 1, qui présente la répartition des centres de confiance en fonction des étiquettes phonétiques, que les distributions des phonèmes étudiés sont comparables, à l'exception du chwa non systématiquement réalisé.

Table 1: Répartition (%) des centres de confiance en fonction des étiquettes phonétiques.

i	ɛ	a	o	u	ə	ɑ̃	n	Autres
9.4	11.6	14.7	7.3	8.8	3.0	13.2	8.3	23.7

La figure 1 présente les valeurs moyennes et l'intervalle de confiance de NAQ par phonème. Les valeurs de NAQ sont comprises entre 0,07 et 0,32 ce qui, en comparant aux valeurs obtenues par Alku et al. [1] pour cinq locuteurs masculins, signifie que les stimuli analysés se répartissent sur l'ensemble du continuum *pressed-breathy*. Les valeurs moyennes de NAQ semblent plus élevées pour les voyelles orales hautes, bien que cette tendance ne soit pas significative. Le phonème [ə] présente en outre un NAQ moyen plus élevé, mais présente une répartition clairement bimodale des valeurs de NAQ. Le locuteur 1 ajoute [ə] sur

les fins de mots avec des valeurs de NAQ (0,28) élevées, correspondant à une voix *breathy*. Le locuteur 2 ajoute [ə] avec une voix modale (valeurs de NAQ autour de 0,12 de même que pour [ɛ]). Le choix de l'ajout ou non d'un chwa final semble relever de stratégies relatives aux valeurs expressives des actes de langage. A noter que les deux locuteurs présentent des fréquences d'ajout de [ə] voisines: 36,8% des stimuli de locuteur 1, contre 42,9% chez le locuteur 2. Tandis que la voyelle nasale [ɑ̃] présente des valeurs de NAQ similaires à celles des voyelles hautes, la consonne nasale [n] a des valeurs de NAQ correspondant à une voix *breathy* (0,19). Toutes les différences sont significatives à l'exception de celle entre [n] et [ə]. Il paraît irréaliste que le phonème [n] de [ʒon] soit systématiquement *breathy* alors que [o] ne l'est pas. On pourrait arguer que cela est dû à sa position finale, mais ceci reste observable lorsque [ə] est ajouté. Une explication possible est que la nasalité produit essentiellement des basses fréquences, qui augmentent la pente spectrale en atténuant les hautes fréquences. En effet les mouvements supra-laryngés dans le cas de la nasalité, et laryngés pour le caractère *breathy* produisent le même effet acoustique, à savoir une augmentation de la pente spectrale. On a donc dans le cas du [n] une mauvaise interprétation : un effet supra-laryngé est attribué à une voix *breathy*.

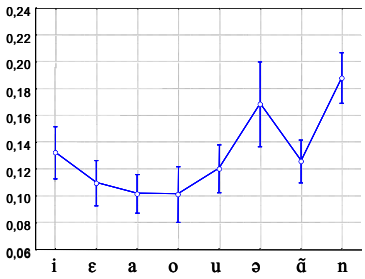

Figure 1 : Valeur moyenne et intervalle de confiance p<0,01 de NAQ pour chaque phonème.

4.2. Estimations de F0

La plupart des paramètres basés sur l'amplitude étant normalisés par la fréquence fondamentale, cela implique que les erreurs dans son estimation sont répercutées sur l'estimation de tous les autres paramètres. $F0_A$, estimée par l'algorithme de calcul des paramètres basés sur l'amplitude, a donc été comparée à $F0_{EGG}$, extraite par autocorrélation du signal EGG. Ces deux valeurs ont été calculées sur les mêmes portions de signal, centrées sur les centres de confiance détectés. La corrélation entre ces deux mesures est de r^2=0,64. Ceci doit être comparé aux valeurs de F0 obtenues à l'aide de l'éditeur EdiProso développé à l'ICP (basé sur la détection par seuil des points d'annulation du signal) pour lesquelles la corrélation avec $F0_{EGG}$ atteint une valeur de 0,79. Il ressort en outre de cette comparaison que les valeurs de fréquence fondamentale utilisées pour normaliser les paramètres basés sur l'amplitude tendent à être sous-estimées, ce qui implique que les valeurs de ces paramètres devraient également être sous-estimées. L'utilisation de deux méthodes distinctes pour l'estimation de F0, ainsi que le préfiltrage réalisé par l'EGG, peuvent participer à cette sous-estimation.

Dans notre corpus, pour les deux locuteurs masculins sélectionnés, F0 présente des valeurs significativement supérieures pour [ə] réalisé par le locuteur 1, accompagné de valeurs élevées de NAQ.

4.3.OQ_A vs. OQ_{EGG}

OQ représente la durée la phase ouverte, i.e. la somme de la phase d'ouverture et de la phase de fermeture. Son estimation basée sur l'amplitude OQ_A devrait donc être partiellement corrélée à NAQ, lié à la phase de fermeture [1]. Dans notre corpus, cette corrélation est élevée (r^2=0,93) ce qui tend à prouver que la phase de fermeture explique l'essentiel de la variance du quotient ouvert, l'asymétrie entre les phases d'ouverture et de fermeture de la glotte étant moins importante.

La corrélation entre OQ_A et $F0_{EGG}$ est de r^2=0,28. La fréquence fondamentale ne peut donc expliquer la variation de durée de la phase ouverte, qui semble clairement indépendante des autres paramètres prosodiques. Les valeurs de quotient ouvert extraites du signal EGG, OQ_{EGG}, ne présentent pas de corrélation avec F0. Ces résultats doivent être comparés à ceux obtenus par Henrich [10] en voix chantée, qui a trouvé une corrélation entre F0 et OQ chez les chanteurs utilisant le mécanisme laryngé II, mais pas pour le mécanisme I qui est le plus fréquemment utilisé par les sujets masculins en voix parlée.

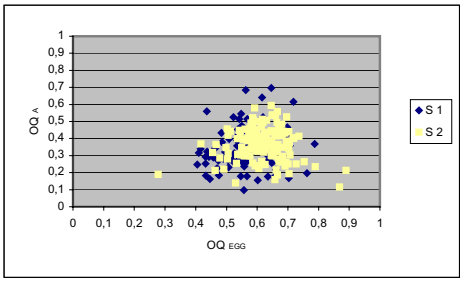

Figure 2 : Répartitions relatives de OQ_A et OQ_{EGG}.

La figure 2 montre la répartition des valeurs de OQ_A par rapport à celles de OQ_{EGG}. Les valeurs de OQ_A sont moins élevées, ce qui s'explique en partie par la sous-estimation de $F0_A$. Toutefois on observe une répartition similaire entre $T1_A$ et $T1_{EGG}$, bien que F0 n'intervienne pas dans leur calcul. En effet, les valeurs de $T1_A$ sont toujours plus faibles et on n'observe pas plus de corrélation entre $T1_A$ et $T1_{EGG}$ qu'entre OQ_A et OQ_{EGG}, quand bien même on considère chaque phonème séparément.

5. DISCUSSION

Bien qu'ils soient fortement corrélés dans notre corpus, il convient de souligner que les quotients NAQ et OQ_A décrivent des phénomènes bien distincts, respectivement la part dans la durée totale du cycle glottique de la phase de fermeture et de la phase ouverte de la glotte. L'énergie de la

source glottique est produite lorsque les cordes vocales sont en contact (phase de fermeture), plus que lorsque la glotte est ouverte. Les estimations de NAQ sont donc vraisemblablement plus fiables que celles de OQ_A. De plus le calcul de OQ_A requière l'estimation d'un paramètre de plus que celui de NAQ, à savoir EI, ce qui introduit une source d'erreur supplémentaire.

A la lumière des résultats de Gobl et Ní Chasaide [9], il est toutefois surprenant que OQ_A et OQ_{EGG} soient si faiblement corrélés. Une explication pourrait être une inadéquation du filtre inverse utilisé pour l'estimation de l'onde de débit glottique. En effet nous avons calculé automatiquement les paramètres basés sur l'amplitude, sans adaptation au locuteur particulière, tandis que les résultats de Gobl et Ní Chasaide ont été obtenus après une détection de formants réalisée par un expert. Etant donné qu'aucune méthode ne donne de mesure directe du débit glottique, la meilleure solution pour assurer un filtrage inverse adéquat semble être la supervision par un expert.

En dépit de nos tentatives, nous sommes dans l'incapacité de lier les mesures articulatoires issues du signal EGG aux estimations basées sur l'amplitude du débit obtenu par filtrage inverse du signal acoustique. Il est cependant indubitable que les caractéristiques de l'onde de débit glottique influencent le jugement perceptif émotionnel [9], et que l'on peut lier NAQ au degré « d'attention » porté à la voix, comme l'a montré Campbell [5]. Ainsi NAQ apparaît clairement comme un paramètre extrait du signal acoustique qui est porteur d'informations sur la qualité de voix.

6. CONCLUSION

A partir d'un corpus de parole émotionnelle authentique enregistré en laboratoire [3], nous avons comparé des paramètres obtenus par filtrage inverse du signal acoustique à ceux extraits directement du signal EGG. Les paramètres issus de l'amplitude du signal obtenu par filtrage inverse ont été calculés grâce à un algorithme de calcul automatique de NAQ sur les vocoïdes d'un signal de parole non étiqueté [12], appliqué également au calcul d'une estimation du quotient ouvert, OQ_A [9].

Les résultats ont montré un effet du phonème sur NAQ avec une distribution différente pour les deux locuteurs. Cet effet implique la nécessité de normaliser NAQ par des facteurs phonémiques préalablement à son utilisation comme paramètre prosodique. De plus, les valeurs de NAQ sur le segment nasal [n], fréquemment détecté comme vocoïde, se sont avérés surestimés. Ceci peut être interprété comme un problème d'inversion, à savoir la similitude des effets acoustiques induits par la nasalité *vs.* le contrôle de la voix *breathy*. Si ces résultats peuvent mettre en cause la validité de la mesure dynamique directe de NAQ, cela n'affecte pas la pertinence des estimations globales et statiques de NAQ calculé sur des corpus équilibrés et de grande taille (par exemple [5]).

La comparaison de F0 estimé par l'algorithme de calcul de NAQ, $F0_A$, et extrait directement du signal EGG, $F0_{EGG}$, a montré une sous-estimation de $F0_A$, automatiquement répercutée sur les paramètres normalisés NAQ et OQ_A.

Enfin la comparaison des valeurs de OQ_A et OQ_{EGG} n'a montré aucune corrélation entre ces paramètres supposés estimer la même quantité. Ceci laisse supposer une plus grande sensibilité du quotient ouvert au filtrage inverse utilisé pour estimer de l'onde de débit glottique. Une perspective intéressante serait donc d'effectuer une adaptation experte au locuteur avant l'estimation de l'onde de débit.

7. REMERCIEMENTS

Ce travail s'inscrit dans le projet Expressive Speech Project du JST/CREST, dirigé par N . Campbell. Nous remercions chaleureusement N. Campbell et P. Mokhtari pour leurs solutions techniques et leurs conseils judicieux.

BIBLIOGRAPHIE

[1] P. Alku, T. Bäckström & E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustic Society of America*, 112 (2), 701-710, 2002.

[2] V. Aubergé & M. Cathiard. Can we hear the prosody of smile ? Special issue *Emotional Speech*, *Speech Communication Review* 40, 2003.

[3] V. Aubergé, N. Audibert & A. Rilliard. Why and how to control emotional speech corpora. *8th European Conference on Speech Communication and Technology*, 185-188, 2003.

[4] N. Campbell. Databases of Emotional Speech. *ISCA Workshop on Speech and Emotions*, Newcastle, Northern Ireland, 34-38, 2000.

[5] N. Campbell & P. Mokhtari. Voice Quality: the 4th Prosodic Dimension. *15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2417-2420, 2003.

[6] A. R. Damasio. *Descartes' error. Emotion, reason, and the human brain.* A. Grosset/ Putnam Books, 1994.

[7] G. Fant. The voice source in connected speech. *Speech Communication Review* 22, 125-139, 1997.

[8] C. Gobl & A. Ní Chasaide. Testing affective correlates of voice quality through analysis and resynthesis. *ISCA Workshop on Speech and Emotions*, Newcastle, Northern Ireland, 178-183, 2000.

[9] C. Gobl & A. Ní Chasaide. Amplitude-based source parameters for measuring voice quality. *ISCA Workshop on Voice Quality VOQUAL'03*, 151-156, 2003.

[10] N. Henrich, C. d'Alessandro, M. Castellengo & B. Doval. Mesures électroglottographiques de quotient d'ouverture en voix parlée et chantée. *XXIIIèmes Journées d'Etude sur la Parole*, Aussois, France, 2000.

[11] D. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustic Soc. of America*, 83 (1), 257-264, 1988.

[12] J. Laver. *The phonetic description of voice quality.* Cambridge University Press, Cambridge, 1980.

[13] P. Mokhtari & N. Campbell. Automatic Detection of Acoustic Centres of Reliability for Tagging Paralinguistic Information in Expressive Speech. *3rd International Conference on Language Evaluation and Resources*, Las Palmas, Spain, 2015-2018, 2002.

[14] K. R. Scherer, T. Johnstone & G. Klasmeyer. Vocal Expression of Emotion. In R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds). *Handbook of Affective Sciences*, 433-456, 2003.