Analyse et Modification de la Qualité Vocale basée sur l'Excitation

Thomas Drugman, Baris Bozkurt, Thierry Dutoit

TCTS Lab - Faculté Polytechnique - Université de Mons
31, Boulevard Dolez - 7000 Mons - Belgique

ABSTRACT

This paper investigates the differences occuring in the excitation for different voice qualities. Its goal is two-fold. First a large corpus containing three voice qualities (modal, soft and loud) uttered by the same speaker is analyzed and significant differences in characteristics extracted from the excitation are observed. Secondly rules of modification derived from the analysis are used to build a voice quality transformation system applied as a post-process to HMM-based speech synthesis. The system is shown to effectively achieve the transformations while maintaining the delivered quality.
Keywords : Speech Analysis, Speech Synthesis, Voice Quality, Glottal Source, Voice Modification

1. INTRODUCTION

Depuis les débuts de la recherche en synthèse de parole, l'analyse et la modification de qualité vocale (ou timbre vocal perçu) ont attiré un intérêt scientifique particulier [9]. L'analyse de qualité vocale trouve des applications dans divers domaines du Traitement de la Parole, tels que la synthèse paramétrique de parole de haute qualité, la synthèse de parole expressive/émotionnelle, l'identification du locuteur, la reconnaissance d'émotions, l'analyse de prosodie, etc...En raison de la disponibilité de revues tels que [4] et de la limitation d'espace, une présentation des méthodes d'analyse de la qualité vocale ne sera pas détaillée dans cet article.

Pour analyser la qualité vocale sur un corpus de parole, il est d'usage d'estimer des paramètres spectraux directement à partir du signal de parole, comme par exemple les amplitudes relatives d'harmoniques, ou encore le rapport signal-sur-bruit (HNR). Bien que les variations de qualité vocale soient principalement considérées comme étant contrôlées par la source glottique, l'estimation de cette dernière est bien souvent considérée comme étant problématique et donc évitée dans les procédures d'estimation de paramètres sur d'importants corpus. Dans ce travail, nous adoptons une approche essentiellement orthogonale en étudiant les différences présentes dans la source glottique estimée via un algorithme automatique quand un locuteur donné produit différentes qualités vocales. En se basant sur une analyse paramétrique de la contribution glottique (Section 2), nous examinons l'utilisation de l'information extraite d'un important corpus pour modifier, dans un synthétiseur de parole basé sur des HMMs (Section 3), la qualité vocale d'autres bases de données.

2. ANALYSE DE LA QUALITÉ VOCALE BASÉE SUR L'EXCITATION

Le but de cette partie est de mettre en exergue les différences présentes dans l'excitation quand un locuteur donné produit différentes qualités vocales. La base de données De7 utilisée dans cette étude a initialement été développée par Marc Schroeder dans un but de synthèse de parole expressive par diphones [13]. Cette base de données contient 3 qualités vocales (modale, douce et tendue) prononcées par une locutrice allemande, avec environ 50 minutes de parole par qualité vocale. Dans la Section 2.1, les méthodes d'estimation du flux glottique et de paramétrisation de celui-ci sont présentées. L'harmonicité de la parole est étudiée via la fréquence maximale de voisement en Section 2.2. En tant que caractéristique perceptuelle importante, le tilt spectral est analysé en Section 2.3. La Section 2.4 compare les *résidus propres* [8] estimés pour différentes qualités vocales. Finalement la Section 2.5 quantifie la séparabilité entre les 3 qualités vocales pour les attributs d'excitation ainsi extraits.

2.1. Source Glottique

Nous avons récemment montré que le cepstre complexe permet d'estimer efficacement le flux glottique [6]. Cette méthode a pour but de séparer les composantes à minimum et à maximum de phase du signal de parole. En effet il a été montré précédemment [5] que la parole est un signal à phase mixte où la contribution à maximum de phase (c-à-d anticausale) correspond à la phase ouverte glottique, alors que la composante à minimum de phase est liée à la transmittance du conduit vocal. Isoler la composante à maximum de phase permet donc une estimation fiable de la source glottique, ce qui peut être réalisé par le cepstre complexe. La phase ouverte de la glotte est ensuite paramétrisée par 3 attributs : la fréquence du formant glottique (F_g), le Quotient d'Amplitude Normalisé (NAQ, [2]) et le Quotient de Quasi-Ouverture (QOQ, [1]).

La fréquence du formant glottique est extraite par la méthode décrite dans [3]. La Figure 1(a) montre les histogrammes de F_g/F_0 pour les 3 qualités vocales. Des différences significatives entre les distributions sont observées. Entre autres, il s'avère qu'une voix plus tendue (douce) est traduite par une fréquence de formant glottique plus élevée (basse). La présence de deux modes pour les voix modale et tendue peut être également remarquée sur cette figure. Ceci peut être expliqué par le fait que la source glottique estimée peut comprendre une *ondulation résiduelle* dans les domaines temporel et fréquentiel. Cette

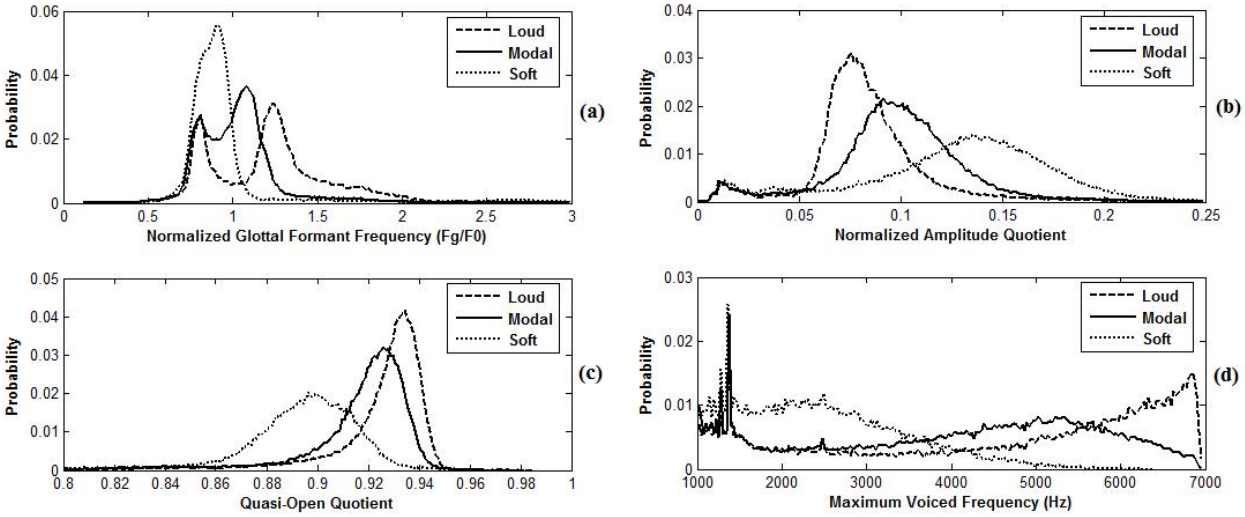

FIGURE 1: Histogrammes, pour les 3 qualités vocales, de *(a) :* la fréquence du formant glottique normalisée F_g/F_0, *(b) :* le Quotient d'Amplitude Normalisé NAQ, *(c) :* the Quotient de Quasi-Ouverture QOQ, et *(d) :* la fréquence maximale de voisement F_m.

dernière peut avoir 2 causes possibles : une séparation incomplète entre F_g et le premier formant F_1 [3], et/ou une interaction non-linéaire entre le conduit vocal et la glotte [12]. Cette ondulation résiduelle peut alors perturber la détection du formant glottique et de cette façon expliquer le pic parasite dans l'histogramme de F_g/F_0 (pic pour les valeurs de F_g/F_0 inférieures à 1).

Dans des travaux précédents [2], [1], Alku et al. ont proposé le Quotient d'Amplitude Normalisé et le Quotient de Quasi-Ouverture comme 2 paramètres temporels utiles pour caractériser respectivement les phases de fermeture et ouverte du flux glottique. Ces paramètres sont ici extraits via la librairie Aparat [11], librement accessible, à partir de la source glottique estimée par le cepstre complexe. Les Figures 1(b) et 1(c) illustrent les histogrammes de ces 2 attributs pour les 3 qualités vocales. Des différences notables entre ces distributions peuvent être observées.

2.2. *Fréquence Maximale de Voisement*

Certaines approches, telles que le Modèle Harmonique plus Bruit (HNM, [14]), considèrent que la parole peut être modélisée par une composante non-périodique au-delà d'une fréquence donnée. Dans le cas du HNM, cette *fréquence maximale de voisement* (F_m) démarque la frontière entre 2 bandes spectrales distinctes, où respectivement des modélisations harmonique et stochastique sont supposées être valides. Plus F_m est élevée, plus l'harmonicité est forte, plus la présence de bruit dans la voix sera faible. Dans ce travail, F_m est estimée par l'algorithme décrit dans [14]. La Figure 1(d) montre les histogrammes de F_m pour les 3 qualités vocales. Il peut être noté qu'en général la voix douce a une fréquence maximale de voisement basse (résultant de sa nature *soufflante*), et que plus l'effort vocal est marqué, plus la parole est harmonique, et par conséquence plus F_m est grande.

2.3. *Tilt Spectral*

Le tilt spectral de la parole est connu pour jouer un rôle important dans la perception de la qualité vocale [16]. Pour capturer cet attribut essentiel, un spectre moyenné est calculé sur le corpus entier (pour une qualité de voix donnée) par un processus indépendant de la prosodie et des variations du conduit vocal. Pour cela, les trames de parole voisée sont extraites par un fenêtrage de Hanning long de 2 périodes de pitch et centré sur un Instant de Fermeture Glottique (GCI). Les positions des GCIs sont déterminées par la technique décrite dans [7]. Ces trames sont ensuite ré-échantillonnées sur un nombre fixé de points et normalisées en énergie. Le spectre moyen est finalement obtenu en moyennant les spectres des trames ainsi normalisées. Ce spectre moyen d'amplitude contient donc un mélange des contributions moyennes de la glotte et du conduit vocal. Le spectre moyenné est illustré en Figure 2 pour les 3 qualités vocales. Puisque ces spectres moyens ont été calculés pour le même locuteur et que l'ensemble d'enregistrements utilisé est phonétiquement équilibré (les formants du conduit vocal auront donc tendance à se contrebalancer), il est raisonnable de penser que les principales différences entre eux sont liées au tilt spectral de la qualité vocale considérée. Entre autres il peut être remarqué que plus l'effort vocal est important, plus le contenu spectral dans la bande [1kHz-5kHz] est riche.

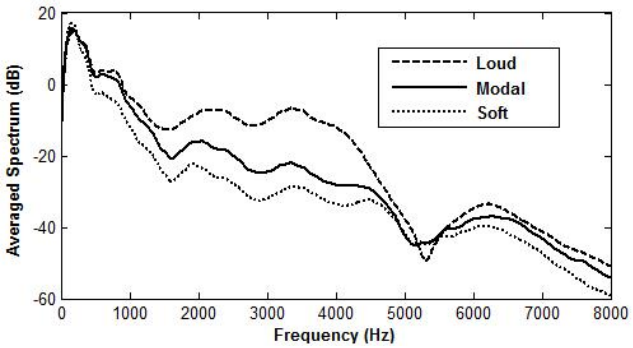

FIGURE 2: Spectre moyenné pour les 3 qualités vocales.

2.4. *Résidus propres*

Nous avons proposé dans [8] de modéliser le signal résidu (obtenu par filtrage LPC inverse) en décomposant des

trames de résidu pitch-synchrones sur une base orthonormée. Il a été également montré que le premier vecteur propre ainsi obtenu (appelé *résidu propre*) permet d'améliorer le naturel en synthèse paramétrique de parole. Comme les résidus propres seront utilisés dans l'application de modification de qualité vocale (Section 3), la Figure 3 illustre la forme d'onde de ce signal selon la qualité vocale produite. On peut noter que les conclusions tirées en Section 2.1 à propos de la phase ouverte glottique sont corroborées. On peut en effet observer que plus l'effort vocal est important, plus la réponse de la phase ouverte du résidu propre est rapide.

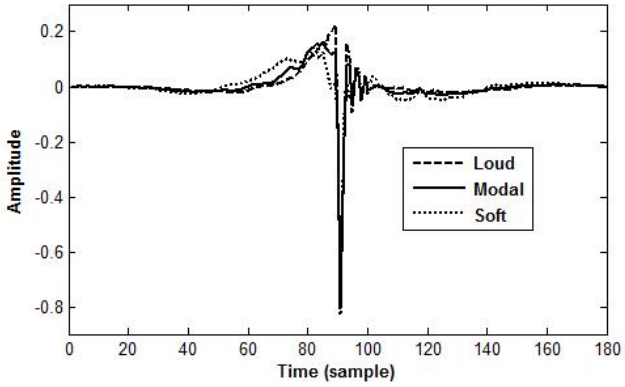

FIGURE 3: Premier résidu propre pour les 3 qualités vocales.

2.5. Séparabilité entre Distributions

Des différences importantes entre les distributions d'attributs ont été présentées dans les sections précédentes. Celles-ci sont d'ailleurs en accord avec les conclusions exposées dans d'autres études [9], [2], [16]. Dans cette section, nous quantifions combien ces différences entre les qualités vocales sont importantes. Pour cela, la divergence de Kullback-Leibler (KL) est connue pour mesure la séparabilité entre 2 fonctions de densité discrètes A et B [10]. Mais puisque cette mesure est non-symétrique (et n'est donc pas une vraie distance), sa version symétrisée, appelée divergence de Jensen-Shannon [10], est généralement préférée. Elle consiste en la somme de 2 mesures KL :

$$D_{JS}(A,B) = \frac{1}{2}(\sum_i A(i)\log_2\frac{A(i)}{M(i)} + \sum_i B(i)\log_2\frac{B(i)}{M(i)}) \quad (1)$$

où M est la moyenne des 2 distributions ($M = 0.5 * (A + B)$). Le Tableau 1 montre les résultats pour les 4 attributs présentés précédemment. Entre autres, on peut remarquer que les voix tendues et douces sont fortement séparables, alors que la qualité tendue est plus proche de la voix modale que la qualité douce ne l'est. On voit également que F_g et NAQ sont hautement informatives pour l'annotation automatique de qualités vocales.

3. MODIFICATION DE QUALITÉ VOCALE

Nous avons proposé dans un travail précédent [8] le Modèle Déterministe plus Stochastique (DSM) du signal résidu. Selon cette approche, l'excitation est divisée en 2 bandes spectrales distinctes délimitées par la fréquence maximale de voisement F_m. La partie déterministe concerne le contenu basses-fréquences et est modélisée par le résidu propre tel que décrit en Section 2.4. Quant à la composante stochastique, il s'agit d'un bruit hautes-fréquences filtré similairement à ce qui est fait dans le modèle HNM [14]. Le signal résidu est ensuite passé dans un filtre de type LPC pour obtenir la parole synthétique.

	F_g	NAQ	QOQ	F_m
$D_{JS}(T,M)$	0.196	0.118	0.035	0.076
$D_{JS}(T,D)$	0.353	0.371	0.279	0.297
$D_{JS}(M,D)$	0.175	0.194	0.175	0.215

TABLE 1: Divergence de Jensen-Shannon entre les 3 qualités vocales (M = Modal, T = Tendu, D= Doux) pour les 4 caractéristiques de l'excitation extraites.

Cette section a pour but d'appliquer les modifications de qualité vocale comme un post-traitement à la synthèse de parole basé sur des HMMs [15], tout en utilisant la modélisation DSM du signal résidu. Plus précisément, un synthétiseur basé sur des HMMs est entraîné sur un corpus de voix modale pour un locuteur donné. Le but est ensuite de transformer la voix de synthèse afin qu'elle soit perçue comme étant douce ou tendue, tout en évitant une dégradation de qualité globale dans la parole produite.

Puisqu'aucun enregistrement de voix expressive n'est disponible pour le locuteur considéré, les modifications sont extrapolées à partir des prototypes décrits pour le locuteur De7 en Section 2, en faisant l'hypothèse que d'autres locuteurs modifient leur qualité vocale de la même façon. Trois principales transformations sont ici considérées :

- Les résidus propres présentés en Section 2.4 sont utilisés pour la partie déterministe du modèle DSM. Ces formes d'onde véhiculent implicitement les modifications de la phase ouverte glottique qui ont été soulignées en Section 2.1.
- La fréquence maximale de voisement F_m est, pour une qualité de voix donnée, fixée selon la Section 2.2 en prenant sa valeur moyenne : 4600 Hz pour la voix tendue, 3990 Hz pour la modale (ce qui confirme les 4 kHz que nous avons utilisé dans [8]), et 2460 Hz pour la qualité douce.
- Le tilt spectral est modifié en utilisant l'inverse du processus décrit en Section 2.3. Pour cela, le spectre des segments voisés est transformé, dans le domaine pitch-normalisé, par un filtre exprimé comme le ratio entre les modélisations auto-régressives des spectres moyennés des qualités vocales source et cible (voir Figure 2). A la synthèse, les trames de résidu sont ensuite ré-échantillonnées à la fréquence fondamentale cible. Cette dernière transformation est donc pitch-dépendante.

Pour évaluer la technique, 10 personnes ont participé à un test subjectif. Le test consistait en 27 phrases générées par notre système pour 3 locuteurs (2 masculins et 1 féminin). Un tiers de ces phrases a été converti vers une voix plus douce, et un autre tiers vers une voix plus tendue. Pour chaque phrase, il a été demandé aux participants d'évaluer l'effort vocal perçu (0 = très doux, 100 = très tendu), et de donner un score MOS selon leur appréciation de la qualité générale. Les résultats sont détaillés dans le Tableau

2 avec leur intervalles de confiance à 95%. De manière intéressante, il peut être noté que les modifications de qualité vocale sont perçues comme souhaité, alors que la qualité globale n'est sensiblement pas altérée (bien que les sujets aient une tendance naturelle à préférer les voix douces).

	Effort vocal	Scores MOS
Modal vers Doux	36.11 ± 2.60	3.189 ± 0.145
Modal	52.89 ± 2.82	3.017 ± 0.147
Modal vers Tendu	72.11 ± 2.60	2.606 ± 0.146

TABLE 2: Evaluation de l'effort vocal perçu (0 = voix très douce, 100 = voix très tendue) et scores MOS pour les 3 versions, ainsi que leurs intervalles de confiance à 95%.

4. CONCLUSION

Nous avons montré dans cette étude qu'un algorithme d'estimation du flux glottique [6] peut être utilisé de manière efficace pour l'analyse de la qualité vocale sur un important corpus de parole, où la plupart de la littérature sur l'estimation du flux glottique se base sur des tests avec des voyelle soutenues. Nous avons étudié les variations de paramètres d'excitation pour différentes qualités vocales et conclu que les 2 attributs F_g et NAQ caractérisant le flux glottique sont hautement informatifs pour l'annotation de qualité vocale. De plus, nous avons montré que l'information extraite d'une base de données peut être appliquée à d'autres corpus de parole dans un but de modification de qualité vocale, tout en maintenant dans une application de synthèse paramétrique de parole une qualité globale relativement élevée.

5. REMERCIEMENTS

Thomas Drugman est supporté par le Fonds National de la Recherche Scientifique (FNRS). Les auteurs aimeraient également remercier M. Schroeder pour la base de données De7, ainsi que Y. Stylianou pour nous avoir fourni l'algorithme d'extraction de la fréquence maximale de voisement.

RÉFÉRENCES

[1] M. Airas and P. Alku. Comparison of multiple voice source parameters in different phonation types. In *Proc. Interspeech*, 2007.

[2] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. In *JASA*, volume 112, pages 701–710, 2002.

[3] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. A method for glottal formant frequency estimation. In *Proc. Interspeech*, 2004.

[4] C. D'Alessandro. Voice source parameters and prosodic analysis. In *Method in empirical prosody research*, pages 63–87, 2006.

[5] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *Proc. ISCA ITRW VOQUAL03*, pages 15–19, 2003.

[6] T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. Interspeech*, 2009.

[7] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[8] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech*, 2009.

[9] D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. In *JASA*, volume 87, pages 820–857, 1990.

[10] J. Lin. Divergence measures based on the Shannon entropy. In *IEEE Trans. on Information Theory*, volume 37, pages 145–151, 1991.

[11] [Online]. TKK Aparat main page. In *http ://aparat.sourceforge.net/*.

[12] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. In *IEEE Trans. on Speech and Audio Processing*, volume 7, pages 569–586, 1999.

[13] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In *Proc. 15th International Conference of Phonetic Sciences*, pages 2589–2592, 2003.

[14] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. In *IEEE Trans. Speech and Audio Processing*, volume 9, pages 21–29, 2001.

[15] K. Tokuda, H. Zen, and A. Black. An HMM-based speech synthesis system applied to english. In *Proc. IEEE Workshop on Speech Synthesis*, pages 227–230, 2002.

[16] O. Turk, M. Schroeder, B. Bozkurt, and L. Arslan. Voice quality interpolation for emotional text-to-speech synthesis. In *Proc. Interspeech*, 2005.