Amélioration de la conversion de voix chuchotée enregistrée par capteur NAM vers la voix audible

Viet-Anh Tran, Gérard Bailly, Hélène Lœvenbruck & Christian Jutten

GIPSA-Lab, Département Parole & Cognition, UMR n°5216 CNRS/INPG/UJF/U. Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex, France
{viet-anh.tran, gerard.bailly, helene.loevenbruck}@gipsa-lab.inpg.fr, christian.jutten@lis.inpg.fr
http:// gipsa-lab.inpg.fr

ABSTRACT

The NAM-to-speech conversion proposed by Toda and colleagues, which converts Non-Audible Murmur (NAM) to audible speech by statistical mapping trained using aligned corpora, is a very promising technique, but its performance is still insufficient due to the difficulties in estimating F_0 from unvoiced speech. In this paper, two distinct modifications are proposed, in order to improve the intelligibility and the naturalness of the synthesized speech. In the first change, a simple neural network is used to detect voiced segments in the whisper while a GMM estimates a continuous melodic contour based on these voiced segments. The second change is an attempt to integrate visual information as a complementary input to improve spectral and F_0 estimation, and voicing decision.

Keywords: audiovisual voice conversion, non-audible murmur, whispered speech.

1. INTRODUCTION

La parole contient des informations multiples. Parmi elles, le contenu linguistique du message prononcé est primordial, mais les informations paralinguistiques jouent aussi un rôle crucial dans la communication orale [13]. Toutefois, quand le locuteur murmure ou chuchote, ces informations sont dégradées. Or les avancées technologiques récentes en communication sans fil ont mené à l'utilisation répandue des mobiles pour la communication privée. Parler fort dans un mobile dans les endroits publics peut être une nuisance. La voix chuchotée, cependant, peut être entendue par un nombre limité d'auditeurs entourant le locuteur et utiliser ce mode de parole pourrait donc limiter les nuisances. Dans ce cadre, Nakajima *et al* [11] ont constaté que les vibrations acoustiques dans le conduit vocal, les plus faibles soient-elles, peuvent être capturées avec un dispositif acoustique spécial appelé microphone NAM (pour *Non Audible Murmur*). En utilisant ce microphone NAM pour capturer le murmure non-audible, Toda *et al* [14] ont proposé un système de conversion à partir du murmure vers la voix audible basé sur le modèle de GMM. Il a été montré que ce système est efficace mais sa performance est toujours insuffisante, surtout pour le naturel de la parole convertie en raison des difficultés de l'estimation du F_0 à partir de la voix inaudible. Pour éviter ces difficultés, Nakagiri *et al* [10] ont proposé un système qui convertit la voix inaudible vers la voix chuchotée. Dans ce système, les valeurs de F_0 n'ont pas besoin d'être estimées puisque le chuchotement est comme le NAM un régime de phonation non sonore, mais plus intelligible. Cependant, il est difficile d'employer la voix chuchotée en raison de son peu d'intelligibilité et de familiarité. La conversion du chuchotement vers la voix modale est nécessaire pour le développement de « téléphone silencieux ».

Dans cet article, deux modifications sont proposées pour améliorer la conversion du chuchotement vers la voix modale, basée sur le modèle GMM. Le chuchotement a été utilisé au lieu du NAM en raison de la segmentation phonétique difficile en NAM. La première modification a pour but d'améliorer l'estimation du voisement pour la parole convertie. Seules les trames voisées sont utilisées pour entraîner le modèle de GMM qui convertit les vecteurs spectraux du chuchotement vers les valeurs de F_0 de la parole synthétisée dans la phase d'apprentissage. Dans la phase de conversion, nous utilisons un réseau de neurones pour estimer ces segments voisés dans la phrase chuchotée puis calculer les valeurs de F_0 sur ces segments plutôt que sur toutes les trames. La deuxième modification est un essai préliminaire visant à intégrer les informations visuelles des mouvements faciaux au système original.

2. LA VOIX CHUCHOTÉE

2.1. Les traits acoustiques

Dans la voix modale, les sons voisés impliquent une modulation de la circulation d'air issu des poumons par la vibration des cordes vocales. Cependant, il n'y a aucune vibration des cordes vocales dans la production de voix chuchotée. Pour cette raison, les caractéristiques acoustiques du chuchotement diffèrent de celles de la voix modale. Une étude des propriétés acoustiques des voyelles [8] a montré une augmentation des fréquences de formant pour les voyelles chuchotées comparées à la voix modale. Le décalage est plus grand pour les voyelles à valeurs formantiques peu élevées. Il a également été constaté que les caractéristiques du conduit vocal pour les phonèmes voisés changent plus dans le chuchotement par rapport à la voix modale que celles des consonnes non-voisées. La perception du *pitch* dans la voix modale est principalement liée à la fréquence fondamentale (F_0). Dans le chuchotement, cependant, bien qu'il n'y ait aucune vibration de cordes vocales, une certaine perception de la hauteur peut être possible. Higashikawa *et al.* [5, 6] ont montré que les auditeurs peuvent percevoir le *pitch* dans le chuchotement et que les changements simultanés des formants F1 et F2 pourraient être l'un des indices qui influencent cette perception.

2.2. Microphone NAM

Le son chuchoté, créé par les mouvements coordonnés de la langue, du vélum, des lèvres, etc., peut être capté grâce à la radiation aux lèvres mais aussi par la transmission des chocs entre articulateurs et parois ainsi que la transmission de l'onde de pression par les tissus mous. Le chuchotement peut ainsi être capté par un microphone NAM placé sur la peau, sous l'oreille [11]. Le tissu peaussier et la radiation des lèvres agissent comme un filtre passe-bas et les composantes haute fréquence sont atténuées. Toutefois, les composantes spectrales du chuchotement (et du murmure inaudible) fournissent assez d'information pour identifier les sons [4]. Le capteur NAM enregistre la parole dans une bande de fréquence allant jusqu'à 4kHz, en étant peu sensible au bruit externe.

3. AMELIORATION DE L'ESTIMATION DE F_0 POUR LE WHISPER-TO-SPEECH

Plusieurs approches existent pour convertir la voix inaudible en voix modale. Une première approche consiste à utiliser un pivot phonétique et à combiner reconnaissance de NAM avec synthèse de parole modale [7]. Le *mapping* direct de signal-à-signal en utilisant des corpus alignés est aussi très prometteur : Toda *et al* [14] ont appliqué un *mapping* statistique [9, 13] pour la conversion de NAM vers la parole modale. Toutefois, bien que l'intelligibilité segmentale des signaux synthétiques calculés par *mapping* soit acceptable, les auditeurs ont des difficultés à regrouper les segments pour récupérer des mots. Ce problème est dû en partie à la restauration de la mélodie synthétique. Ainsi, dans notre système (Figure 1), adapté de Toda *et al.*, nous nous sommes concentrés sur l'amélioration de l'estimation de la mélodie et de la détection du voisement.

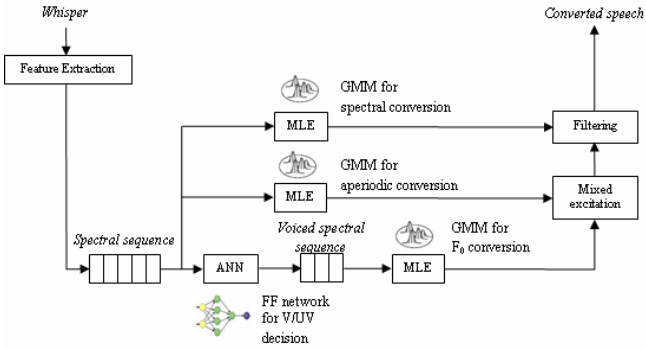


Figure 1 : Système Whisper-to-Speech modifié.

Nous avons choisi un corpus de 270 paires d'énoncés (phrases du journal *Le Monde*) chuchotés et en voix modale, prononcés par un locuteur français entraîné, enregistrés par capteur NAM et par microphone. Les durées sont de 4,9 mn en chuchoté et 4,8 mn en modal. Un microphone de calibration a permis de s'assurer que le mode chuchoté était peu audible à 40cm. Avant la mise en correspondance statistique des trames chuchotées et de parole modale, les phrases en modes chuchoté et audible ont été alignées. Un alignement par segmentation manuelle en phonèmes a été adopté car il est meilleur que l'anamorphose temporelle (DTW) de Toda *et al.* [14].

3.1. Estimation du spectre et de l'excitation

Pour synthétiser la parole, il faut estimer non seulement les traits spectraux mais aussi les traits d'excitation, y compris F_0 et les composants apériodiques [14]. Le trait spectral à chaque trame a été construit en concaténant les vecteurs spectraux de plusieurs trames autour de la trame courante, afin de compenser les caractéristiques perdues sur quelques phonèmes, particulièrement les fricatives, d'énergie élevée sur les bandes à haute fréquence. Trois GMMs ont été utilisés pour convertir les traits spectraux du chuchotement en trois traits de la parole, *i.e.* le spectre, le F_0 et un composant apériodique (qui capture le bruit sur chaque bande de fréquence du signal d'excitation). Pour l'estimation spectrale et celle des composants apériodiques, la méthode du système original a été utilisée. Par contre, pour l'estimation des valeurs de F_0, au lieu de prendre tous les segments, seuls les segments voisés ont été utilisés pour entraîner une mixture de Gaussiennes (GMM), ceci afin d'éviter de perdre des composantes gaussiennes pour représenter les valeurs nulles de F_0 codant les segments non-voisés. De plus, un réseau de neurones (RN) est utilisé pour prédire ces segments. Pour la synthèse, on prédit donc des valeurs de F_0 continues, hachées par le voisement calculé par ce RN.

3.2. Evaluation

Détection du voisement – Dans le système original, Toda *et al* estiment la valeur de F_0 pour toutes les trames en utilisant le modèle GMM. Un seuil de valeur de F_0 est déterminé pour assigner l'étiquette voisée/non-voisée à chaque trame. Dans notre système, nous avons testé la possibilité d'améliorer la détection du voisement en employant un RN *feedforward* simple. De façon empirique, nous avons utilisé 50 neurones d'entrée (*i.e.* autant que la taille des vecteurs d'entrée du système de conversion décrit sur la figure 1), 17 neurones cachés et 1 neurone de sortie. Les vecteurs de paramètres d'entrée du module de conversion spectral – renouvelés toutes les 5ms et issus d'une analyse en composantes principales (ACP) des cepstres de 17 trames de 20ms centrées sur la trame courante - ont été utilisés comme vecteur d'entrée pour ce réseau. Pour l'apprentissage du réseau, le classement en voisé/non-voisé de chaque énoncé chuchoté a été obtenu en l'alignant avec l'énoncé modal correspondant. Le tableau 1 montre l'évaluation des performances de ce RN. Comparativement à l'erreur dans le système original, nous avons une nette amélioration de cette détection (26%).

Table 1 : Err. de détection de voisement par RN et GMM.

	Réseau de neurone(%)	GMM
Err voisé	2.4	3.3
Err non-voisé	4.4	5.9
Total	**6.8**	**9.2**

Evaluation de F_0 – Nous avons aussi comparé les deux systèmes, en fonction du nombre de gaussiennes utilisées, dans l'estimation de F0 sur les corpus d'apprentissage et de test. Le nombre de gaussiennes pour le *mapping* spectral a été fixé à 32. Des matrices de covariance

pleines ont été utilisées pour les GMMs. Le corpus de test était composé de 70 paires d'énoncés non incluses dans le corpus d'apprentissage. Le même corpus a été utilisé pour tester les deux systèmes. L'erreur est calculée comme la différence relative entre le F0 synthétique et le F0 naturel dans les segments voisés bien détectés par les deux systèmes : Err=(F0_synth – F0_naturel)/F0_naturel. La Figure 2 montre que la méthode proposée surpasse l'originale. L'erreur des deux systèmes sur les données d'apprentissage diminue quand le nombre de gaussiennes augmente, mais pas sur les données de test (overtraining).

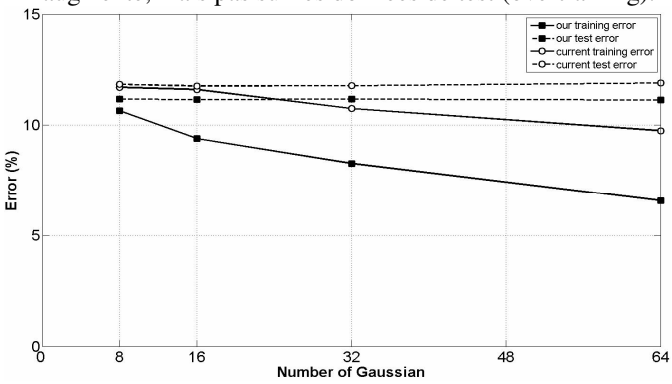


Figure 2 : Taux d'erreur pour les deux systèmes sur les corpus d'apprentissage et de test.

La Figure 3 montre un exemple d'une courbe F_0 cible et des courbes synthétiques produites par les deux systèmes. Notre courbe est plus proche de la F_0 cible que l'originale.

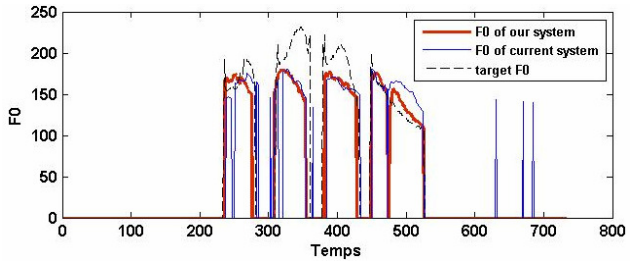
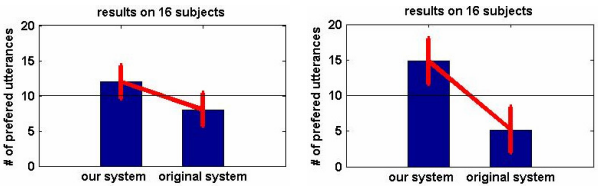


Figure 3 : Courbes de F_0 cible (pointillé) et synthétiques (gras : nouveau système, fin : ancien) pour un énoncé.

Evaluation perceptive – Seize auditeurs français ont participé à nos tests perceptifs sur l'intelligibilité et le naturel de la parole convertie des deux systèmes. 20 phrases, qui n'étaient pas incluses dans le corpus d'apprentissage, ont été utilisées. Chaque auditeur a passé deux tests ABX. Il a entendu une phrase prononcée dans la voix modale (X) et les versions converties à partir du chuchotement par les deux systèmes. Pour chaque phrase, l'auditeur devait choisir laquelle était la plus proche de l'originale (X), en terme d'intelligibilité et de naturel. La Figure 4 fournit les scores moyens d'intelligibilité et de naturel obtenus pour les phrases converties utilisant les systèmes original et modifié, cumulés pour tous les auditeurs. Les scores d'intelligibilité sont significativement plus hauts pour les phrases synthétisées par notre système (F=23.41, p<.001). Ceci est aussi vrai pour le naturel : le système proposé a été encore plus fortement préféré à l'original (F = 74.89, p < .001).



(a) (b)

Figure 4 : Score d'intelligibilité (a) et de naturel (b).

4. CONVERSION AUDIOVISUELLE

Plusieurs études ont décrit la contribution importante des lèvres et des mouvements faciaux pour l'intelligibilité de la parole visuelle humaine et artificielle [2]. Dans le domaine de la communication interpersonnelle ou entre l'Homme et la machine, le signal visuel des lèvres peut être une modalité additionnelle utile, d'autant plus que des caméras de petites tailles peuvent être intégrées dans les téléphones et ordinateurs de nouvelle génération.

4.1. Système de conversion audiovisuelle

Le système de conversion est construit à partir de données audiovisuelles. La base de données comprend cette fois-ci 120 phrases du japonais prononcées par un locuteur natif, en modes chuchoté et modal, enregistrées avec un capteur NAM et un microphone (dont 100 phrases pour le corpus d'apprentissage et 20 phrases pour le corpus de test). Le système capture, à 200Hz, les positions 3D de 142 billes collées sur le visage (Figure 5), en synchronie avec le signal acoustique échantillonné à 16000 Hz.

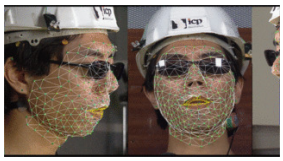


Figure 5 : Points utilisés pour la capture des mouvements.

Un modèle de forme est construit à partir des positions 3D des 142 points caractéristiques augmentés des 30 points de contrôle d'un modèle générique de lèvres ajusté à la main sur un ensemble de visèmes. La méthodologie de clonage développée dans notre département [12] consiste en une ACP itérative appliquée sur des sous-ensembles de points pertinents. Cette analyse guidée extrait 5 paramètres articulatoires de la façon suivante. D'abord, la contribution de la mâchoire est estimée et soustraite des données. Ensuite, l'arrondissement des lèvres est estimé à partir du résidu et soustrait des données. Les mouvements verticaux des lèvres supérieure et inférieure et du larynx sont soustraits dans cet ordre des données résiduelles. De manière analogue à l'audio, chaque trame vidéo interpolée à 200Hz - pour être synchrone avec l'audio - est caractérisée par un vecteur obtenu par ACP des 5 paramètres articulatoires de 17 trames centrées autour de la trame courante. Un vecteur caractéristique audiovisuel est obtenu en combinant caractéristiques audio et visuelle comme pour les AAM (*Active Appearance Models*) de Cootes [3]. Chaque vecteur visuel est multiplié par un poids w avant d'être concaténé avec le vecteur acoustique correspondant. La dimension du vecteur conjoint est ensuite diminuée grâce à une autre ACP.

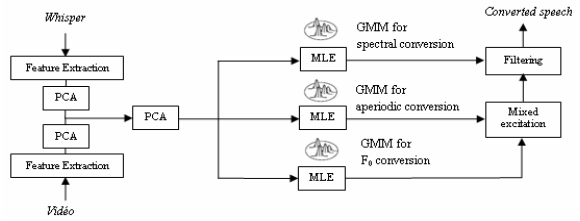


Figure 6 : Conversion utilisant l'information visuelle.

La conversion utilise les vecteurs de projection des trames sur les 40 premiers axes principaux (figure 6). Ici, les nombres de gaussiennes sont fixés à 16 pour l'estimation spectrale et à 8 pour F_0 et les composantes apériodiques.

4.2. Résultats préliminaires

Le tableau 2 montre la contribution positive de l'information visuelle sur la performance du système. Le meilleur résultat est obtenu avec w=1 et une dimension du vecteur visuel de 20. La distorsion spectrale entre paroles convertie et modale diminue alors de 2,3%, l'erreur de voisement de 16,5% et de F_0 de 10,3%. Si on n'utilise que la vidéo, la performance se dégrade nettement.

5. CONCLUSIONS ET PERSPECTIVES

Ce papier décrit nos travaux pour améliorer intelligibilité et naturel de la parole convertie du Whisper-to-speech.

Table 2 : Contribution de l'information visuelle pour :

(a) l'estimation spectrale (distorsion cepstrale par dB)

	Audio	0.25	0.5	0.75	1	1.25	1.5	1.75	2	Vidéo
10		5.656	5.632	5.608	**5.576**	5.595	5.634	5.619	5.654	
20		5.676	5.630	5.596	**5.564**	5.598	5.606	5.598	5.623	
40	5.687	5.676	5.630	5.596	**5.611**	5.568	5.605	5.596	5.618	9.894
50		5.676	5.630	5.596	**5.597**	5.568	5.609	5.592	5.619	

(b) la détection voisée/non-voisée (%)

	Audio	0.25	0.5	0.75	1	1.25	1.5	1.75	2	Vidéo
10		13.786	13.484	12.898	**12.669**	20.707	20.670	20.972	21.219	
20		13.236	13.575	12.559	**12.358**	20.734	20.276	20.569	19.534	
40	14.811	13.236	13.575	12.559	**12.696**	20.450	20.532	20.358	20.258	31.335
50		13.236	13.575	12.559	**13.383**	20.450	20.505	20.743	20.258	

(c) l'erreur de F_0 converti (%)

	Audio	0.25	0.5	0.75	1	1.25	1.5	1.75	2	Vidéo
10		18.39	18.14	17.54	**17.27**	24.58	24.52	24.77	25.53	
20		17.85	18.21	17.14	**17.47**	24.62	24.15	24.51	23.53	
40	19.48	17.85	18.21	17.14	**17.28**	24.39	24.41	24.47	24.21	36.31
50		17.85	18.21	17.14	**17.93**	24.39	24.37	24.63	24.21	

Les résultats préliminaires sur la contribution de l'information visuelle nous permettent de continuer dans cette direction pour un corpus français. Bien que la performance des systèmes modifiés soit améliorée de manière significative par rapport à celle du système original, les variations de F_0 sont sous-estimées. En accord avec un modèle superpositionnel de l'intonation [1], il semble intéressant d'étudier la combinaison de modèles prédictifs opérant à diverses échelles de temps.

Remerciements : Merci à C. Savariaux, C. Vilain & A. Arnal pour l'acquisition de données, à K. Nakamura & T. Toda du NAIST pour la mise à disposition du NAM et à. H. Kawahara de l'univ. de Wakayama pour la permission d'utiliser STRAIGHT.

BIBLIOGRAPHIE

[1] Bailly, G. and B. Holm, SFC: a trainable prosodic model. *Speech Communication*, 2005. **46**(3-4): p. 348-364.

[2] Benoît, C., Intelligibilité audio-visuelle de la parole, pour l'homme et la machine, 1998, INPG.

[3] Cootes, T.F., G.J. Edwards, & C.J. Taylor, Active Appearance Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2001. **23**(6): p. 681-685.

[4] Heracleous, P., *et al*. A tissue-conductive acoustic sensor applied in speech recognition for privacy. in *International Conference on Smart Objects & Ambient Intelligence.* 2005. Grenoble - France. p. 93 - 98.

[5] Higashikawa, M., Nakai, K., Sakakura, A. and Takahashi, H., Perceived Pitch of Whispered Vowels – Relationship with formant frequencies : A preliminary study, *Journal of Voice*, **10**(2), 155-158.

[6] Higashikawa, M., Minifie, F.D., Acoustical perceptual correlates of whispered pitch in synthetically generated vowels, *JSHR,* 42, 583-591, 1999.

[7] Hueber, T., *et al*. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. in *Interspeech*. 2007. Antwerp, Belgium. p. 658-661.

[8] Ito, T.; Takeda, K.; Itakura, F., 2005. Analysis and recognition of whispered speech. In *Speech Communication*. Lisboa. Vol. 45, Issue 2, 139-152.

[9] Kain, A. & M.W. Macon. Spectral voice conversion for text-to-speech synthesis. *ICASSP* 1998. Seattle. 285-288.

[10] Nakagiri, M., et al. Improving body transmitted unvoiced speech with statistical voice conversion. in *InterSpeech*. 2006. Pittsburgh, PE. p. 2270-2273.

[11] Nakajima, Y., et al. Non-Audible murmur recognition. in *EuroSpeech*. 2003. Geneva, Switzerland. p. 2601-2604.

[12] Revéret, L., Bailly G., and Badin P. MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. in *ICSLP*. 2000. Beijing, China. p. 755-758.

[13] Stylianou, Y., O. Cappe, and E. Moulines, Continuous probabilistic transform for voice conversion. *IEEE Trans. on Speech and Audio Processing*, 1998. **6**(2): p. 131-142.

[14] Toda, T. and K. Shikano. NAM-to-Speech Conversion with Gaussian Mixture Models. in *InterSpeech*. 2005. Lisbon - Portugal. p. 1957-1960.