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

Viet-Anh Tran, Gérard Bailly, Hélène Lovenbrück & 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. In this paper, we present our current work to improve the intelligibility and the naturalness of the synthesized speech converted from whispered speech with this technique. The first system is proposed to improve F_0 estimation and voicing decision. A simple neural network is used to detect voiced segments in the whisper while a GMM estimates a continuous melodic contour based on training voiced segments. In the second system, we attempt to integrate visual information for improving both spectral estimation, 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. Pourtant, les informations paralinguistiques comme l'identité et l'humeur du locuteur ou sa position par rapport à qu'il/elle dit jouent aussi un rôle crucial dans la communication orale [10]. Malheureusement, quand le locuteur murmure ou chuchote, ces informations sont dégradées.

Pour résoudre ce problème, Nakajima *et al* [8] ont constaté que les vibrations acoustiques dans le conduit vocal peuvent être capturées à travers les tissus mous de la tête avec un dispositif acoustique spécial appelé un microphone NAM attaché à la surface de la peau, au-dessous de l'oreille. En utilisant ce microphone stéthoscopique pour capturer le murmure non-audible, Toda *et al* [12] ont proposé un système de conversion à partir de cette voix 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* [7] ont proposé un autre 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.

Dans cet article, nous proposons deux versions modifiées du système original. Le premier système améliore l'estimation de la source voisée de 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 et calculer ensuite les valeurs de F_0 sur ces segments plutôt que sur toutes les trames. Le deuxième système est un essai préliminaire d'intégrer les informations visuelles des mouvements faciaux au système original en attendant de valider par un test de perception ses contributions positives sur l'intelligibilité et le naturel de la parole convertie.

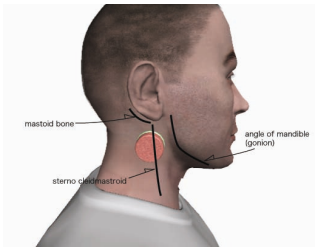


Figure 1 : Position du microphone NAM.

2. MICROPHONE NAM

Le chuchotement est défini comme la production articulée de sons sans vibration des cordes vocales. Le son produit selon les mouvements et les interactions des organes comme la langue, le vélum, les lèvres etc., peut être capté grâce à la radiation aux lèvres mais aussi par la transmission des chocs entre articulateurs et parois du conduit ainsi que la transmission de l'onde de pression par les tissus mous de la tête. Le chuchotement peut ainsi être capturé par un microphone stéthoscopique placé sur la peau, au-dessous de l'oreille (Figure 1). Le tissu peaussier et la radiation des lèvres agissent comme un filtre passe-bas et les composants à haute fréquence sont atténués. Cependant, les composants spectraux du chuchotement (et même du murmure inaudible) fournissent suffisamment d'information pour distinguer et identifier les sons exactement [4]. Actuellement, le microphone NAM peut enregistrer le son de parole avec des composants de fréquence jusqu'à 4 kHz, tout en étant peu sensible au bruit externe.

3. SYSTÈME NAM-TO-SPEECH MODIFIÉ

Plusieurs approches peuvent être considérées pour convertir la voix inaudible en voix modale. Une première approche consiste à utiliser un pivot phonétique et de combiner reconnaissance de NAM avec synthèse de parole modale [voir notamment les travaux de Hueber *et al* dans 5]. Le *mapping* direct de signal-à-signal en utilisant des corpus alignés est aussi très prometteur : Toda *et al* [12] ont appliqué un *mapping* statistique [6, 10, 12] pour la conversion de NAM vers la parole modale. Bien que l'intelligibilité segmentale des signaux synthétiques calculés par *mapping* statistique soit acceptable, les auditeurs ont des difficultés à regrouper ces éléments pour récupérer des mots significatifs. Une grande partie de ce problème est due à la restauration de la mélodie synthétique. Dans notre système (Figure 2), nous nous concentrons sur l'amélioration de l'estimation de la mélodie et la détection voisée/non-voisée de la parole convertie.

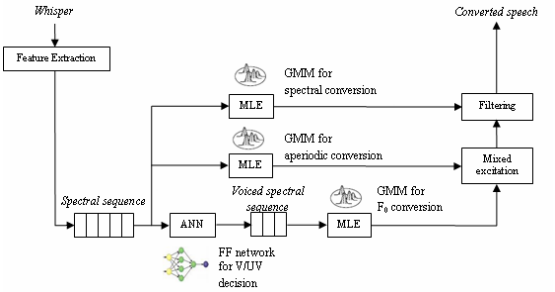


Figure 2 : Système NAM-to-Speech modifié.

Avant la mise en correspondance statistique de trames chuchoté et de trames de parole modale, les paires de phrases prononcées en mode chuchoté et de façon audible par le locuteur cible doivent être alignées. Un alignement utilisant une segmentation manuelle des deux versions en segments phonétiques a été adopté : il produit une erreur de conversion inférieure à la procédure standard itérative proposée par Toda *et al* [12], utilisant anamorphose temporelle (DTW) et conversion .

3.1. Estimation du spectre et de l'excitation

Le module de conversion du système proposé par Toda *et al* [12] se décompose en trois parties principales. Nous utilisons le même schéma pour l'estimation spectrale et celles des composants apériodiques que le système original. Par contre, pour l'estimation des valeurs de F_0, au lieu de prendre tous les segments dans chaque paire d'énoncé, seuls les segments voisés ont été utilisés pour entraîner une mixture de Gausiennes (GMM), ceci afin d'éviter de perdre quelques composants gaussiens pour représenter les valeurs nulles de F_0 pour coder les segments non-voisés. Un réseau de neurones est par contre utilisé pour prédire ces segments. Pour la synthèse, on prédit donc des valeurs de F_0 continues qui sont hachées par le paramètre de voisement calculé par ce réseau.

3.2. Evaluation

Détection voisée/non-voisée

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 utilisons par contre un réseau de neurones « feedforward » avec 50 nœuds d'entrée, 17 nœuds cachés et 1 nœud de sortie. Les vecteurs de paramètres – renouvelés à 200Hz et issus d'une analyse en composantes principales des cepstres de 17 trames de signal de 20ms centrées sur la trame courante - d'entrée du module de conversion spectral ont été utilisés comme vecteur d'entrée pour ce réseau. Le paramètre voisé/non-voisé de sortie de la parole modale cible a été obtenu en alignant les deux énoncés.

Table 1 : Erreur de détection voisée/non-voisée par le réseau de neurones et par le GMM.

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

Le tableau 1 montre l'évaluation des performances de ce réseau. Comparativement à l'erreur dans le système original, nous avons une légère amélioration de cette détection.

Evaluation de F_0

Nous avons également comparé les deux systèmes en fonction du nombre de Gaussiennes utilisé pour estimer F_0 sur le corpus d'apprentissage et le corpus 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. L'erreur est calculée comme la différence relative entre le F_0 synthétique et le F_0 naturel dans les segments voisés bien détectés par les deux systèmes. L'erreur est donnée par la formule suivante : *Err = (F0_synthétique – F0_naturel) / F0_naturel*. La Figure 3 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 ces erreurs sur les données de test sont peu sensibles.

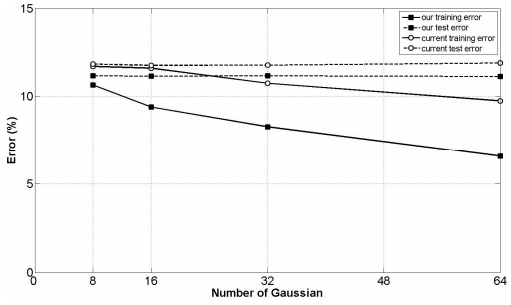


Figure 3 : Taux d'erreur pour les différents systèmes.

La Figure 4 fournit un exemple avec un énoncé chuchoté, la parole convertie par notre système et la parole modale ciblée. Les formants de la parole convertie sont moins modulés que ceux de la parole modale. Les variances globales sont utilisées pour atténuer cette différence [11].

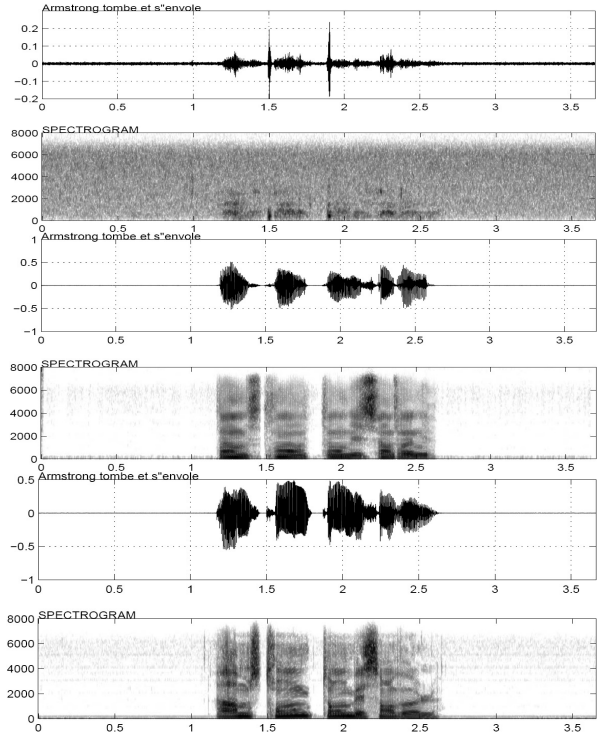


Figure 4 : NAM, parole convertie et parole modale « Armstrong tombe et s'envole ».

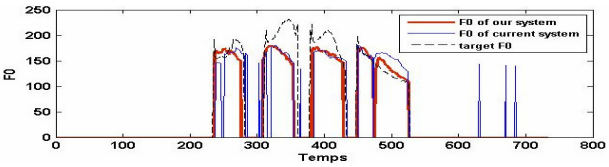


Figure 5 : Courbes de F_0 cible et synthétiques pour la même phrase : « Armstrong tombe et s'envole ».

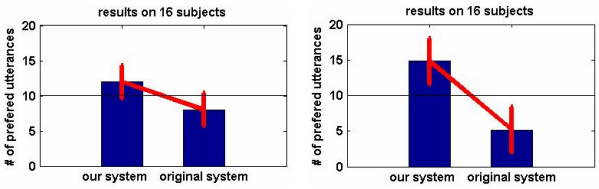
La Figure 5 montre un exemple d'une courbe F_0 cible et les courbes synthétiques produites par les deux systèmes. Notre courbe est plus proche de la courbe F_0 normale que l'originale.

Evaluation perceptive

Seize auditeurs français qui n'avaient jamais écouté de NAM 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 été incluses dans le corpus d'apprentissage ont été utilisées pour ces tests.

Chaque auditeur a passé deux tests ABX. Il a entendu une phrase prononcée dans la voix modale (X) et les deux versions de phrases 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 6 fournit les scores moyens d'intelligibilité et de naturel obtenus pour les phrases converties utilisant le système original et notre système 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 6 : Score d'intelligibilité (a) et de naturel (b).

4. CONVERSION AUDIOVISUELLE

Plusieurs études ont documenté la contribution importante des lèvres et des mouvements faciaux à l'intelligibilité de la parole visuelle humaine et artificielle [2]. Dans le domaine de la communication entre l'homme et machine, le signal visuel des lèvres peuvent être utile comme modalité additionnelle d'entrée et de sortie.

4.1. Système de conversion audiovisuel

Le système de conversion à évaluer est construit à partir de données audiovisuelles. La base de données utilisée comprend cette fois-ci 120 phrases du japonais prononcées par un locuteur natif. Le système capture, à 200Hz, les positions 3D de 142 billes collées sur le visage (Figure 7) de manière synchrone au signal acoustique échantillonné à 16000 Hz.

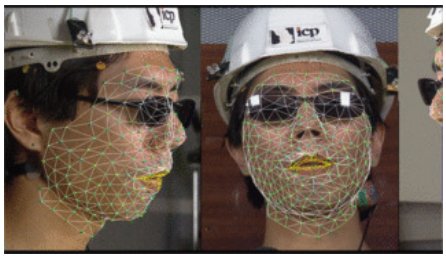


Figure 7 : Disposition des points caractéristiques utilisés pendant la capture des mouvements.

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 à l'ICP [9] consiste en une Analyse itérative en Composants Principaux (ACP) appliquée sur des sous-ensembles pertinents des points caractéristiques. Cette analyse guidée extrait 5 paramètres articulatoires en suivant les étapes : D'abord, les contributions de la rotation de la mâchoire sont estimées et soustraites des données. Ensuite, le mouvement d'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.

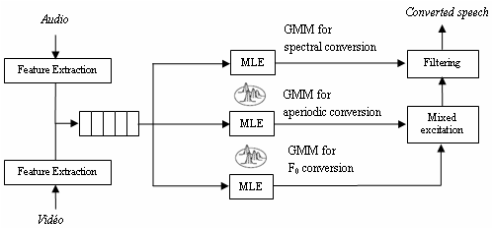


Figure 8 : Schéma du système de conversion utilisant l'information visuelle.

De manière analogue au traitement du signal acoustique, chaque trame vidéo interpolée à 200Hz - de manière à être synchrone au traitement audio - est caractérisée par un vecteur caractéristique obtenu par ACP des 5 paramètres articulatoires de 17 trames centrées autour de la trame courante. Un vecteur caractéristique audiovisuel est finalement obtenu en combinant caractéristiques audio et visuelle de manière identique aux AAM de Cootes [3] : chaque vecteur articulatoire est multiplié avec 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. Le système de conversion utilise alors les vecteurs de projections des trames sur les 40 premiers axes principaux. Le schéma du système est montré dans figure 8. Dans cette évaluation, les nombres de Gaussiennes sont fixés à 16 pour l'estimation spectrale et à 8 pour l'estimation de F_0 et des composantes apériodiques.

Table 2 : Estimation de caractéristiques de la parole modale à partir de diverses caractéristiques du NAM.

	Audio	Vidéo	Audiovisuel
Distortion spectrale	5.69	9.89	5.56
Détection V/UV	14.81	31.34	12.36
Erreur de F_0	19.48	36.31	17.47

4.2. Résultats préliminaires

Le tableau 2 nous montre la contribution positive de l'information visuelle sur la performance du système. Le meilleur résultat est obtenu avec w=1 et la dimension du vecteur visuel à 20. La distorsion spectrale entre la parole convertie et la parole modale diminue alors de 2.3 %, la détection voisée/non-voisée de 16,5 % et l'erreur de F_0 de 10,3%. Si on n'utilise que les informations visuelles, la performance du système se dégrade significativement.

5. CONCLUSIONS ET PERSPECTIVES

Cet article décrit nos travaux pour améliorer l'intelligibilité et le naturel de la parole convertie du système NAM-to-speech basé sur le modèle GMM. 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 à cet égard intéressant d'étudier la combinaison de modèles prédictifs opérant à diverses échelles de temps.

Remerciements : Nous sommes reconnaissants à C. Savariaux, C. Vilain & A. Arnal de leur aide 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'université de Wakayama pour la permission d'utiliser le système 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, Institut National Polytechnique: Grenoble.

[3] Cootes, T.F., G.J. Edwards, and C.J. Taylor, *Active Appearance Models.* IEEE Transactions 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] 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.

[6] Kain, A. and M.W. Macon. *Spectral voice conversion for text-to-speech synthesis*. in *ICASSP*. 1998. Seattle, WA. p. 285-288.

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

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

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

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

[11] Toda, T., A.W. Black, and K. Tokuda. *Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis*. in *International Speech Synthesis Workshop*. 2004. Pittsburgh, PA. p. 26-31.

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