Transformations de la voix pour l'évaluation de systèmes de vérification du locuteur

J.Ph. Goldman * **, G. Chollet * ***

* IDIAP, CP 609, 1920 Martigny, Suisse
** ESIEE, BP 99, 93162 Noisy-le-Grand Cedex, France
*** TELECOM-Paris, CNRS URA-820, 46 r. Barrault, 75634 Paris, France
e-mail: goldman@idiap.ch chollet@idiap.ch

RESUME

Cet article compare les performances de deux systèmes de vérification du locuteur à l'aide d'une nouvelle méthodologie d'évaluation.

Nous décrivons en premier lieu plusieurs modules de transformation de la voix, modifiant la fréquence fondamentale, le débit d'élocution ainsi que la longueur du conduit vocal tout en conservant le naturel et l'intelligibilité de la voix. Les deux caractéristiques prosodiques sont transformées par la technique TD-PSOLA. Grâce à ces transformations nous simulons une partie de la variabilité intra-locuteur et nous évaluons les performances relatives de deux systèmes de vérification du locuteur : SAMREC_1 et SPREC0[1].

ABSTRACT

This paper tries to compare the performance of two speaker verification systems with a new methodology of assessment. We first describe a succession of simple voice transformations that alter the fundamental frequency, the speech duration and the vocal tract length while keeping naturalness and intelligibility. The two prosodic features are modified by the time-domain waveform technique TD-PSOLA (Time-Domain-Pitch-Synchronous-OverLap-and-Add). The vocal tract length is used for the spectral modifications. With these transformations, we compare two speaker verification systems, SAMREC_1 and SPREC_0[1], and compute the evolution of their performance with the simulated intra-speaker variability.

INTRODUCTION

Alors que de plus en plus de systèmes de traitement automatique de la parole sont mis au point et proposés sur le marché, leur utilisation dans le grand public exige fiabilité et précision. Par exemple, les systèmes de vérification du locuteur utilisés pour les transactions bancaires doivent être performants même à long terme pour assurer un niveau de sécurité correct.
Or chaque fabricant emploie sa propre technique pour définir les performances de son système (différentes bases de données de test, taux d'erreur de types différents), ce qui signifie que comparer deux systèmes de vérification ou d'identification du locuteur à l'aide des seules données fournies par le fabricant, se révèle difficile. D'où l'interêt de créer une méthodologie d'évaluation de système de reconnaissance, et c'est le but que s'est fixé la tâche 2500 du projet Esprit 6819 SAM-A ("Assessment Methodology for Speaker Verification Systems").
La première partie de cet exposé décrit brièvement les techniques employées pour l'évaluation des performances de systèmes de vérification d'identité. L'étude de la variabilité intra- et inter-locuteur se révèle essentielle pour l'élaboration de systèmes automatiques de reconnaissance de la parole et du locuteur.

De nombreuses bases de données contenant des locuteurs variés ont été créées, mais peu d'entre elles représentent la variabilité intra-locuteur. C'est pourquoi des systèmes de transformations peuvent être utile pour simuler la variabilité de la parole afin de tester leur robustesse à cette variabilité. Ces modifications peuvent être produites par un environement bruyant, l'effet Lombard, ou la qualité de l'enregistrement, dégradation due au microphone ou au canal de transmission. Les techniques de transformation de la voix ont été utilisées en adaptation au locuteur, ou, plus récemment, dans le domaine de la synthèse pour la personnalisation de la voix. Dans le cas présent, ces techniques peuvent être utiles pour simuler de légères variations des caractéristiques prosodiques et spectrales. On peut trouver une liste de quelques unes de ces techniques dans la seconde partie. Enfin nous comparerons les deux systèmes avec le même banc de test.

1 EVALUATION DE SYSTEMES DE VERIFICATION

Les mesures usuelles en vérification du locuteur représentent les deux types d'erreur possibles. Un faux rejet (FR) a lieu quand un bon utilisateur n'est pas reconnu par le système alors que la fausse acceptation (FA) est la validation d'un imposteur. Ces deux taux d'erreur ne sont pas indépendants:

[1]Subventionné par l'Office Féderal de l'Education et de la Science (OFES) de la Confédération suisse sous contrat no. E3320.

le taux de FR peut être diminué si l'on augmente le seuil de décision car un locuteur aura plus de chance d'être accepté. Malheureusement si ce locuteur est un imposteur, le taux de FA augmente aussi, puisque que la comparaison avec l'élocution de référence est moins stricte.
Le point où le pourcentage de FA est le même que le pourcentage de FR, est appelé Equal Error Rate (EER). En théorie, on s'efforce de réduire l'EER pour améliorer les performances du système de vérification, mais cette méthode d'évaluation comporte certains inconvénients. Si seul l'EER est donné, on ignore l'évolution relative de FR et de FA en fonction du seuil de décision.
D'autre part, un seuil optimal peut être défini pour chaque locuteur, ce qui améliore considérablement les performances globales du système. Finalement, chaque banc de test doit être validé statistiquement par une estimation de l'intervalle de confiance.

Même si en général toutes ces mesures sont acceptées pour l'évaluation de systèmes de reconnaissance, beaucoup de facteurs doivent être pris en compte quand on donne les performances d'un système. Etant donné que chaque fabricant possède ses propres bases de données (qui ne sont pas forcément représentatives de la population, ou assez grande pour avoir un bon intervalle de confiance sur les résultats statistiques), sa manière de l'enregistrer (une ou plusieurs sessions,état émotionnel des locuteurs, qualité de l'enregistrement), sa propre manière de calculer l'EER (un seuil général au système ou un seuil optimal par locuteur), un méthodologie d'évaluation devrait être trouvée afin de normaliser les prétendues performances des systèmes proposés.

2 TRANSFORMATIONS VOCALES

Nous avons implémenté des modules de transformation de la voix permettant de modifier les trois paramètres suivant: fréquence fondamentale, durée d'élocution et longueur de conduit vocal. Ils tentent de simuler la variabilité intra-locuteur. Ce modèle n'est pas parfait mais traduit quand même certaines variations réalistes chez un même locuteur. Par exemple, son émotion et son état de fatigue peuvent modifier les propriétés statistiques du pitch ou de la durée d'elocution (dynamique de F0 moins grande, débit ralenti). Nous avons privilégié le naturel de la voix.
Une comparaison des différentes méthodes de codage pour la synthèse de la parole[2] montre qu'un système LPC à excitation résiduelle, avec la technique TD-PSOLA[3] pour transformer le signal glottique, est une technique par laquelle le naturel et l'intelligibilité de la parole sont bien conservés. Cette technique permet une modification naturelle et très précise des paramètres prosodiques[4].

Les transformations spectrales sont faites en modifiant la longueur du conduit vocal.

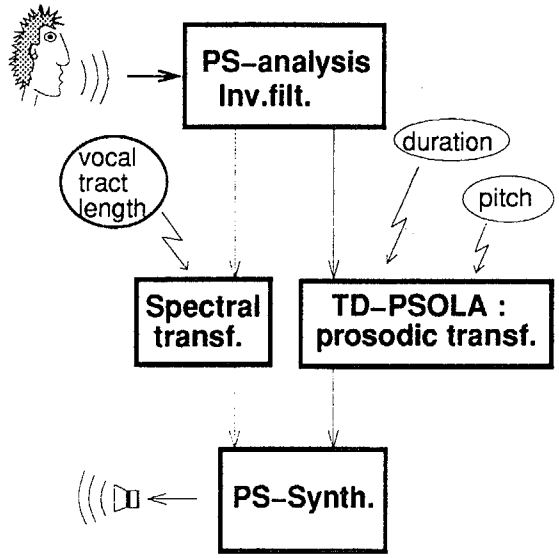

Figure 1. *Transformations de la voix.*

Transformations prosodiques

L'analyse est faite de manière pitch-synchrone : on définit un filtre tout-pôle d'ordre élevé à chaque marque de périodicité (ou "pitch-marque") pour obtenir par filtrage inverse le signal résiduel. A ce stade, les modifications prosodiques et spectrales peuvent être effectuées de manière independante puisque l'excitation glottique et l'influence du conduit vocal ont été déconvoluées. Les transformations prosodiques sont effectuées directement sur le signal résiduel.
La première étape consiste à extraire des signaux à court-terme par fenêtrage et de manière synchrone à la fréquence fondamentale. On utilise une fenêtre de type Hamming d'une longueur proportionnelle à la période fondamentale locale.
Puis de nouvelles pitch-marques sont calculées suivant les modifications désirées. Des transformations linéaires simples de la fréquence fondamentale et du débit d'élocution peuvent être effectuées. Le signal résiduel de synthèse est obtenu par supperposition et addition de la séquence de signaux à court-terme, tout en respectant la nouvelle synchronisation $\hat{t}_n$ et la normalisation $x(n)$:

$$y(m) = \sum_{n=-\infty}^{+\infty} f_n(m-\hat{t}_n).\alpha_n.h_n(\hat{t}_n-m).x(t_n-\hat{t}_n+m) \quad (1)$$

où, la fonction de pondération f_n est calculée à partir des marques de synthèse et de la fonction de fenêtrage h_n:

$$f_n(m) = \frac{1}{\sum_{n=-\infty}^{+\infty} h_n(\hat{t}_n - m)} \quad (2)$$

Une fois le nouveau signal d'excitation obtenu, il faut resynthétiser le signal de parole: ceci est fait grâce aux filtres de prédiction linéaires résultants

de l'analyse LPC. Ceux-ci sont resynchronisés suivant $\hat{t}_n$ et modifiés par une procédure d'interpolation pour améliorer la qualité du signal de synthèse.

Transformations spectrales

Les transformations spectrales sont basées sur les travaux de Fant sur la technique de normalisation formantique et de Wakita qui a proposé une méthode d'estimation de la longueur du conduit vocal[5]. Les deux études ont permi de conclure qu'une estimation linéaire était suffisement précise pour être utilisée en adaption au locuteur dans le cadre de la reconnaissance de la parole. Le $k^{ème}$ pôle d'un filtre tout-pôle estimé par l'analyse LPC est donné par :

$$z_k = e^{(\frac{2l}{Mc}(\pi B_k + j2\pi F_k))} \quad (3)$$

où M est l'ordre d'analyse, c est la vitesse du son, et l est la longueur du conduit vocal.

D'après cette équation on peut voir que la longueur de conduit vocal est proportionnelle à la phase de chaque pôle. Si on modifie la phase de chaque pôle tout en compensant l'energie (donné par la largeur de bande B_k), cela équivaut à changer la longueur de conduit vocal.

Cette transformation est également utilisée dans nos expériences pour simuler une légère variabilité de la longueur du conduit vocal.

Le schéma général (figure.1) montre que les transformations prosodiques et spectrales peuvent être calculées de manière indépendante selon les trois facteurs de modifications (la valeur 1.0 signifie qu'aucun changement n'est effectué). Les modifications de la longueur du conduit vocal et de la fréquence fondamentale sont réalisées uniquement sur les parties voisées du signal de parole alors que le changement de débit d'élocution est fait sur tout le signal.

3 EXPERIENCES

Le but de ces expériences est de faire une comparaison réaliste de deux systèmes de vérification du locuteur, évalués sur la même base de données et dans les mêmes conditions.

Le premier, SAMREC_1, est un système de reconnaissance s'appuyant sur les coefficients MFCC (Mel-Frenquency-Ceptral-Coefficent). Il a été développé durant le projet Esprit SAM 2589. Le second, SPREC0, utilise les coefficients LPCC (Linear Predictive Cepstrum Coefficient). Les deux systèmes sont dépendant du texte et utilise l'algorithme d'alignement dynamique temporel (DTW). Une version modifiée de SAMREC_1 est aussi testée. Dans cette version, la distance calculée pour la vérification est pondérée par l'energie à court-terme. Les partiés voisées du signal, dans lesquelles les paramètres pertinents comme les formants, sont alors privilégiées par rapport portions non-voisées. Leur performances originales sont comparées dans la figure 2, où le taux de fausse acceptation est représenté en fonction du taux de faux rejet.

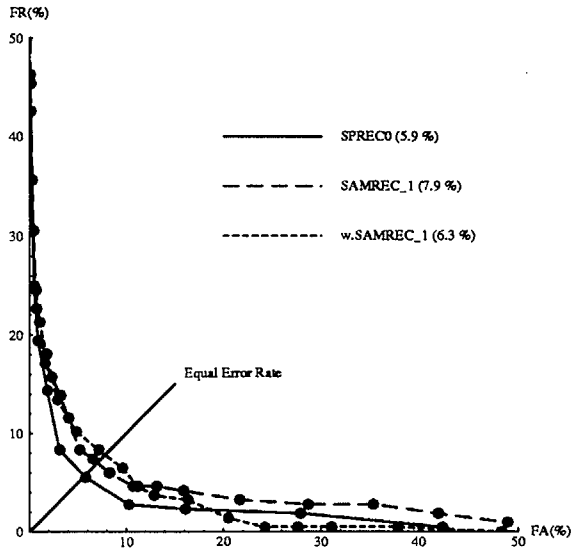

Figure 2. *Comparaison de SPREC0, de SAMREC_1 (avec et sans pondération par l'énergie).*

La base de donnée utilisée provient de BD-SONS (réalisée en 1990 par GDR-Communication-Homme-Machine, France) et contient 6 élocutions du même texte, par 27 locuteurs differents. Les deux premières élocutions représentent la partie référence, les quatres autres la partie test. Cette dernière est transformée selon les méthodes décrites précédement, pour recalculer les nouvelles performances des systèmes étudiés.

La longueur de conduit vocal a été modifiée de -15% à +15%. Dans la figure 3, on voit la dégradation des performances de SPREC0 alors que la longueur du conduit vocal est progressivement diminuée par pas de 2,5%. Pour un raccoursissement de 15%, l'EER est égal à 19%.

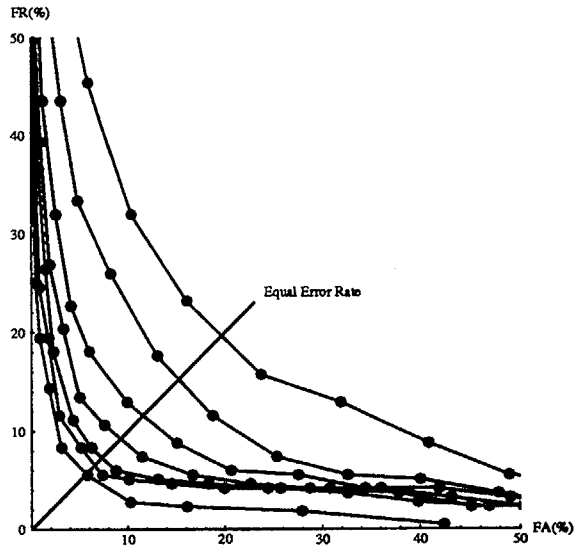

Figure 3. *Evolution des performances de SPREC0 lorsqu'une diminution de la longueur du conduit vocal est simulée (de 0 à 15% par pas de 2.5%).*

La figure 4 montre les performances relatives des trois systèmes en fonction du taux de modification de la longueur du conduit vocal. SPREC0, pour lequel les performances originales paraissaient être les meilleures, est beaucoup plus sensible aux variations spectrales de ce type que SAMREC_1.

D'un autre côté, comme la fonction de pondération utilisée dans la version modifiée de SAMREC_1, donne plus d'importance aux portions de signal qui ont été transformées, les performances se dégradent beaucoup plus vite que dans la version originale.

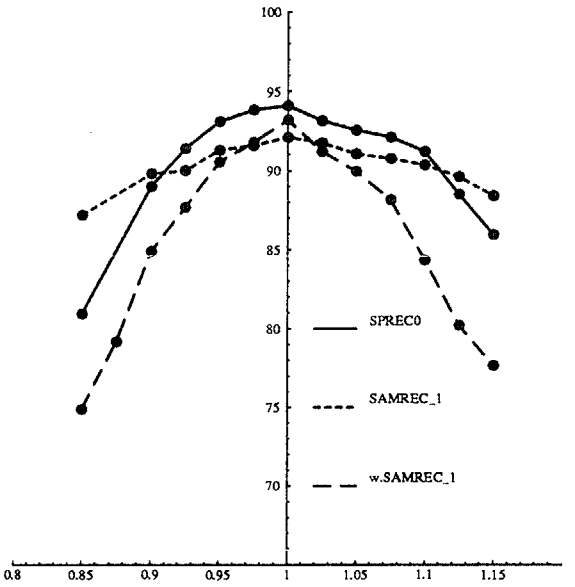

Figure 4. *Performances of SPREC0, SAMREC_1 and weighted SAMREC_1 as a function of the vocal-tract-length modification rate.*

Des expériences du même type sont réalisées pour la fréquence fondamentale. Mais cette modification prosodique ne change pas les performances des systèmes de manière significative pour une transformation allant jusqu'à $\pm 25\%$, bien que la procédure d'interpolation linéaire effectuée sur les paramètres spectraux des filtres modifie l'information spectrale.

Les tests effectués sur le débit d'élocution sont appropriés car les systèmes considérés utilisent l'algorithme d'alignement temporel. La figure 5 montre que les performances de SPREC0 et SAMREC_1 ont un comportement similaire pour des transformations temporelles. Mais la version modifiée de SAMREC_1 semble plus robuste à ce type de variabilité intra-locuteur.

CONCLUSION

Nous avons proposé de nouveaux tests pour l'évaluation de systèmes de vérification du locuteur. Ces tests tentent de simuler une partie de la variabilité intra-locuteur en considérant trois paramètres de transformations: la fréquence fondamentale, le débit d'élocution et la longueur du conduit vocal.

Nous avons fait plusieurs types comparaisons sur deux systèmes de vérification du locuteur et décrit leurs performances suivant le type de transformations effectuées. Nous avons vu que certains systèmes paraissent plus robustes que d'autres suivant le type de test effectué.

Il serait alors intéressant de créer d'autres types de modification, ou de combiner plusieurs de ces transformations pour modéliser la variabilité intra-locuteur réaliste.

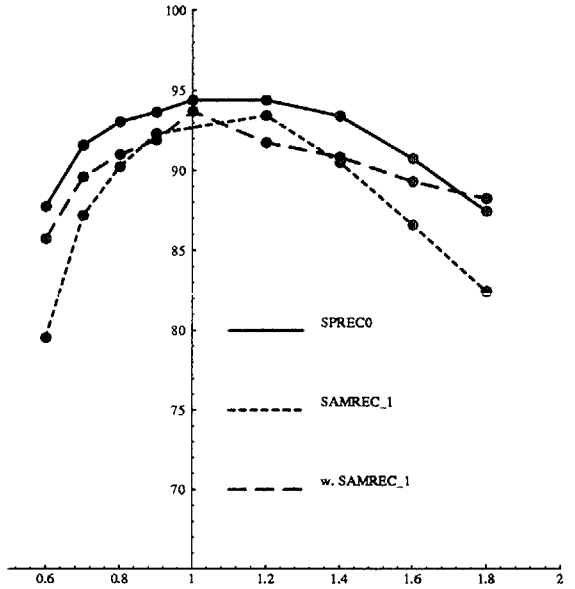

Figure 5. *Performances of SPREC0, SAMREC_1 and weighted SAMREC_1 as a function of the speech duration modification rate.*

Références

[1] M.M.Homayoupour. Vérification vocale d'indentité. *E.S.I.E.E. I.L.P.G.A.*, 1992.

[2] M.J.Macchi, M.J.Altom, D.Kahn, S.Singhal, and M.F.Spiegel. Intelligibility as a function of speech coding method for template-based speech synthesis. *IEEE Workshop on Interactive Voice Technology for Telecommunications Applictions*, 1992.

[3] H.Valbret, E.Moulines, and J.P.Tubach. Voice transformation using psola technique. *Speech Communication*, 11:175–187, 1992.

[4] E.Moulines. Algorithmes de codages et de modification des paramètres prosodiques pour la synthèse de la parole à partir du texte. *Thèse E.N.S.T. Paris*, 1990.

[5] H.Wakita. Normalization of vowels by vocal tract length and its application to vowel identification. *IEEE Trans. ASSP*, ASSP-25:183–192, 1977.