

Reconnaissance du Locuteur basée sur des Signatures Glottiques

Thomas Drugman, Thierry Dutoit

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

ABSTRACT

The great majority of current speaker recognition systems are based on features related to the vocal tract. However some studies have shown that the glottal flow conveys relevant information about the speaker identity. This paper proposes the use of some glottal signatures in speaker recognition. These signatures are extracted from a speaker-dependent dataset of pitch-synchronous residual frames. Experiments of speaker identification are led on both TIMIT and YOHO databases. It is shown that the proposed approach outperforms other state-of-the-art methods based on glottal features.

Keywords : Speaker Recognition, Glottal Analysis, Residual Signal, Voiceprint

1. INTRODUCTION

Développer un système de reconnaissance du locuteur efficace implique une bonne connaissance de ce qui définit l'individualité d'un locuteur. Bien que des informations de haut niveau (comme par exemple l'usage de mots) puissent être envisagées, des attributs acoustiques de bas niveau sont généralement utilisés [11]. Ces derniers sont, la plupart du temps, extraits du spectre d'amplitude du signal de parole. Ils visent à paramétrer la contribution du conduit vocal, qui est une caractéristique importante de l'identité du locuteur. D'un autre côté, très peu de travaux ont étudié la possibilité d'utiliser en reconnaissance du locuteur des attributs émanant de la source glottique. Pourtant des différences significatives dans les formes d'onde glottiques ont été observées entre différents types de locuteurs [6].

Principalement, deux signaux véhiculent de l'information quant au comportement de la glotte : le flux glottique et le signal résidu. Le flux glottique est le débit d'air expulsé dans la trachée et passant à travers les cordes vocales. Son estimation directement à partir du signal de parole est un problème typique de séparation aveugle, puisqu'aucune des contributions glottique et du conduit vocal ne sont observables. Il est donc requis d'adopter un processus d'estimation incorporant une connaissance précise du mécanisme de production. De cette façon, le flux glottique peut être estimé, par exemple, par une analyse spectrale sur la phase fermée de la glotte. Par cette technique, Plumpe et al. [10] ont extrait un ensemble d'attributs temporels paramétrisant le flux glottique ainsi estimé. Dans un canevas similaire, Gudnason et al. [5] ont caractérisé le flux glottique par des coefficients de cepstre réel. Ces deux approches ont abouti à une amélioration, en termes d'identification du locuteur, en combinant ces

paramètres glottiques à des attributs extraits du spectre d'amplitude de la parole (tels que les coefficients LP ou MFCC). D'un autre côté, le signal résidu désigne le signal obtenu par filtrage inverse, après avoir enlevé la contribution de l'enveloppe spectrale. Le signal résidu qui en résulte véhicule de l'information pertinente quant à l'excitation et, contrairement au flux glottique, a l'avantage d'être obtenu facilement. Dans [12], Thevenaz et al. ont suggéré d'utiliser, en vérification du locuteur, des coefficients LPC du signal résidu. Plus récemment, Murty et al. [8] ont mis en évidence, en reconnaissance du locuteur, la complémentarité de la phase résiduelle avec les MFCCs conventionnels. Dans cette dernière étude, l'information contenue dans la phase résiduelle a été extraite via des réseaux de neurones.

Le but de cet article est d'étudier la potentialité d'utiliser des *signatures glottiques* en reconnaissance du locuteur. La recherche d'un invariant dans le signal de parole, caractérisant univoquement une personne (comme pour les empreintes digitales), a toujours attiré la communauté scientifique [7]. Comme ceci semble utopique dû à la nature inhérente du mécanisme de phonation, nous préférons ici le terme de "*signature vocale*" pour désigner un signal contenant une information pertinente quant à l'identité du locuteur. Cet article est structuré comme suit. En Section 2, nous détaillons la façon d'extraire ces signatures vocales à partir du signal de parole et de les inclure dans un système de reconnaissance du locuteur. La Section 3 présente des résultats d'identification du locuteur menés sur les bases de données TIMIT et YOHO. Finalement, la Section 4 conclut cet article.

2. SIGNATURES GLOTTIQUES

2.1. Signatures Glottiques utilisées dans cette Etude

Les *signatures glottiques* utilisées dans cette étude proviennent du Modèle Déterministe plus Stochastique (DSM) du signal résidu que nous avons proposé dans [3] pour la synthèse paramétrique de parole. Ce modèle émane d'une analyse menée sur un ensemble de trames de résidu normalisées et pitch-synchrones. La Figure 1 présente le diagramme utilisé pour obtenir cet ensemble particulier de données à partir d'une collection d'enregistrements d'un locuteur donné. Tout d'abord, une analyse de prédiction linéaire (LP) classique, capturant l'enveloppe spectrale, est réalisée sur les signaux de parole. Les résidus sont ensuite obtenus par filtrage inverse. Les instants de fermeture glottique (GCI) sont

alors identifiés en localisant les discontinuités les plus marquées dans le signal résidu, comme expliqué dans [2]. En parallèle, le pitch est estimé via la librairie Snack Sound Toolkit [9], disponible publiquement. Les trames de résidu pitch-synchrones sont ensuite isolées par un fenêtrage de Blackman centré sur un GCI et long de 2 périodes de pitch. Les trames résultantes sont finalement normalisées en prosodie, c-à-d à la fois en pitch et énergie. Cette opération de normalisation en pitch est réalisée par décimation/interpolation sur un nombre fixé d'échantillons (de façon à ce que les trames de résidu aient toutes la même longueur).

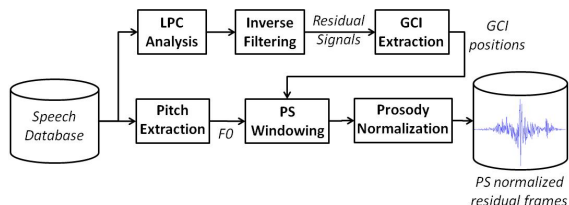


FIGURE 1: Diagramme permettant d'obtenir, pour un locuteur donné, un set de trames de résidu normalisées et pitch-synchrones.

Une fois que le set de trames de résidu est disponible, certaines caractéristiques dépendantes du locuteur et liées au modèle DSM sont extraites sur celui-ci. D'après ce modèle [3], le signal résidu voisé $r(t)$ est composé d'une structure déterministe basses-fréquences $r_d(t)$ et d'une composante stochastique hautes-fréquences $r_s(t)$, supposée modéliser principalement les turbulences présentes dans le débit d'air glottique. Le spectre est donc divisé en deux bandes délimitées par la *fréquence maximale de voisement* F_m (fixée à $4k\text{Hz}$ au sein de cette étude). Le signal résidu synthétisé est alors obtenu comme décrit en Figure 2. La partie déterministe est modélisée par une forme d'onde unique dépendante du locuteur et appelée *premier résidu propre*. Cette forme d'onde est définie comme le premier vecteur propre obtenu par application d'une Analyse en Composantes Principales (PCA) sur le set de trames de résidu. Quant à la composante stochastique, elle est modélisée par un bruit Gaussien hautes-fréquences modulé temporellement par une enveloppe d'énergie pitch-synchrone. Cette *enveloppe d'énergie* est extraite du set de données précédent en moyennant l'enveloppe de Hilbert du contenu hautes-fréquences des trames de résidu.

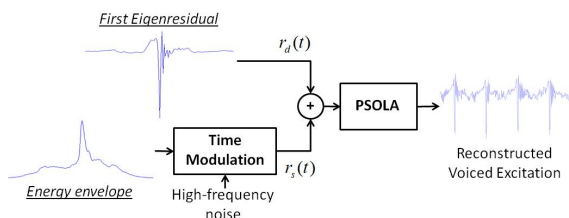


FIGURE 2: Reconstruction de l'excitation voisée selon le Modèle Déterministe plus Stochastique (DSM) du signal résidu. Les 2 signatures glottiques utilisées dans ce travail sont le premier résidu propre et l'enveloppe d'énergie.

En conclusion, le modèle DSM du signal résidu fait usage de deux formes d'onde dépendantes du locuteur, ci-après nommées *signatures glottiques* : le premier résidu propre (ou *résidu propre* tout court) et l'enveloppe d'énergie. La

Figure 3 illustre la forme de l'enveloppe d'énergie pour deux locuteurs masculins. Des différences dans les formes d'onde suggèrent que les signatures glottiques proposées ont le potentiel pour être utilisées en reconnaissance automatique du locuteur.

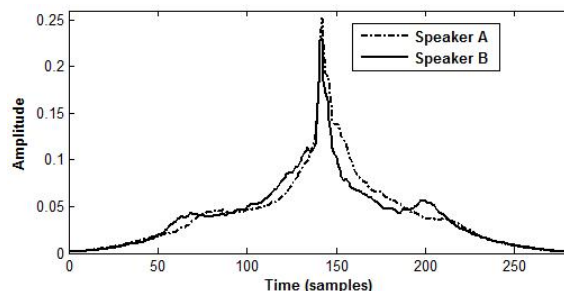


FIGURE 3: Formes d'onde de l'enveloppe d'énergie pour deux locuteurs masculins différents.

2.2. Intégration des signatures Glottiques en Identification du Locuteur

Afin d'être incorporées dans un système d'identification du locuteur, les signatures glottiques sont estimées à la fois sur le test d'entraînement et de test. Une *matrice de confusion* $C(i, j)$ entre le locuteur i et le locuteur j est ensuite calculée. Dans ce travail, le carré de l'erreur temporelle relative (RTSE) a été choisi comme mesure entre deux formes d'onde différentes. Si $v_{k,l,training}$ et $v_{k,l,test}$ désignent la $k^{ième}$ signature glottique (dans notre cas, $k = 1, 2$ respectivement pour le résidu propre et l'enveloppe d'énergie) pour le locuteur l , estimée respectivement sur les sets d'entraînement et de test, la matrice de confusion $C_k(i, j)$ en utilisant uniquement la $k^{ième}$ signature glottique est définie comme :

$$C_k(i, j) = \sqrt{\frac{\sum_{n=0}^{N-1} (v_{k,i,test}(n) - v_{k,j,training}(n))^2}{\sum_{n=0}^{N-1} v_{k,j,training}(n)^2}} \quad (1)$$

où N est le nombre d'échantillons pour la normalisation en pitch. La matrice de confusion $C(i, j)$ est finalement obtenue comme :

$$C(i, j) = C_1(i, j) \cdot C_2(i, j) \quad (2)$$

Notez que plusieurs opérations pour combiner les deux matrices sont possibles. D'après nos expériences, la multiplication a donné les meilleurs résultats, bien que les différences de performance observées étaient relativement faibles.

Finalement, l'identification d'un locuteur i est réalisée en cherchant la plus petite valeur dans la $i^{ième}$ ligne de la matrice de confusion $C(i, j)$. Le locuteur est alors identifié correctement si la position de minimum est i . En d'autres mots, quand des enregistrements sont présentés au système, le locuteur identifié est celui dont les signatures glottiques sont les plus proches (au sens Euclidien) des signatures glottiques extraites sur ces enregistrements.

3. EXPÉRIENCES

Les expériences décrites dans cette Section ont été menées sur les bases de données TIMIT et YOHO. La base de données TIMIT [4] comporte 10 enregistrements prononcés par 630 locuteurs (438 hommes et 192 femmes) échantillonnés à 16 kHz. Quant à la base de données YOHO [1], elle contient de la parole de 138 locuteurs (108 hommes et 30 femmes) échantillonnée à 8 kHz. Ces enregistrements ont été collectés dans un environnement réel de bureau lors de 4 sessions sur une période de 3 mois. Pour chaque session, 24 phrases ont été prononcées par locuteur. Dans nos expériences, les données ont été séparées pour chaque locuteur (et chaque session pour YOHO) en 2 parts égales pour l'entraînement et le test. Ceci est fait de manière à garantir que, pour chaque étape, suffisamment de trames de résidu soient disponibles pour estimer de façon fiable les signatures glottiques.

3.1. Résultats sur la base de données TIMIT

Pour donner une première idée sur le potentiel d'utiliser les signatures glottiques en reconnaissance du locuteur, la Figure 4 montre les distributions de $C_1(i, j)$ respectivement quand $i = j$ et quand $i \neq j$. En d'autres mots, ce graphique montre les histogrammes de la RTSE (voir Equation 1), en échelle logarithmique, entre les résidus propres estimés respectivement pour le même locuteur et pour des locuteurs différents. Il peut être clairement observé que la mesure d'erreur est bien plus grande (environ 15x en moyenne) quand la signature glottique n'appartient pas au locuteur considéré. Cependant, un faible recouvrement des distributions est noté, ce qui peut mener à certaines erreurs d'identification du locuteur.

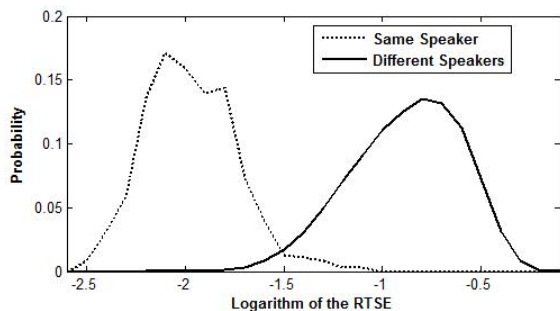


FIGURE 4: Distributions du carré de l'erreur temporelle relative (RTSE) entre les résidus propres estimés respectivement pour le même locuteur et pour des locuteurs différents.

La Figure 5 illustre l'évolution du taux d'identification avec le nombre de locuteurs considérés dans la base de données. Pour cela, l'identification a été réalisée en utilisant une seule des deux signatures glottiques, ou en utilisant leur combinaison comme suggéré par l'Equation 2. Comme attendu, la performance se dégrade quand le nombre de locuteurs augmente, puisque le risque de confusion devient plus important. Cependant cette dégradation est relativement lente dans tous les cas. Une autre observation importante est le clair avantage de combiner les informations des deux signatures glottiques. En effet, ceci mène à une amélioration de 7.78% comparé à l'utilisation unique du résidu propre.

Le Tableau 1 résume les résultats obtenus sur la base de

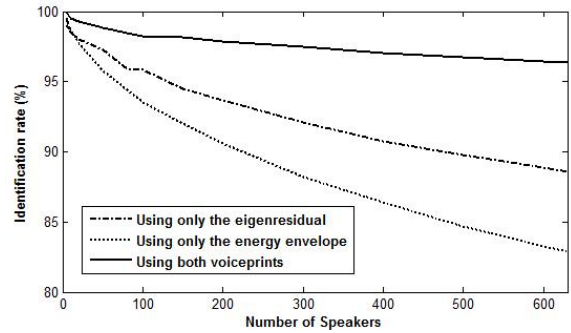


FIGURE 5: Evolution du taux d'identification avec le nombre de locuteurs pour la base de données TIMIT.

données TIMIT. Les taux d'identification pour 168 locuteurs sont aussi donnés pour des motifs de comparaison. En effet, dans [10] Plumpe et al. ont extrait un ensemble de 12 paramètres temporels caractérisant le flux glottique estimé par une analyse sur la phase fermée de la glotte. En utilisant ces attributs, ils ont rapporté un taux de mauvaise classification de 28.64% sur un sous-ensemble de 168 locuteurs. Sur le même sous-ensemble, Gudnason et al. ont rapporté dans [5] un taux de mauvaise classification de 5.06% en utilisant des coefficients du cepstre de la source vocale. Ces résultats peuvent être comparés aux 1.98% que nous avons obtenus en utilisant les deux signatures glottiques. Finalement, notons que Gudnason et al. [5], en utilisant leurs attributs glottiques, ont aussi obtenu un taux de mauvaise classification de 12.95% sur la totalité de la base de données TIMIT (630 locuteurs). Avec les signatures glottiques proposées, un taux de mauvaise classification de 3.65% est atteint.

	168 locuteurs	630 locuteurs
Résidu propre	5.88	11.43
Enveloppe d'énergie	8.76	17.14
Avec les 2 signatures	1.98	3.65
Plumpe et al.	28.64	/
Gudnason et al.	5.06	12.95

TABLE 1: Taux de mauvaise classification (%) sur la base de données TIMIT obtenus en utilisant une seule des deux signatures glottiques, ou leur combinaison.

3.2. Résultats sur la base de données YOHO

Comparé à la base de données TIMIT, le corpus YOHO diffère en deux principaux aspects : 1) les enregistrements sont maintenant échantillonnés à 8 kHz, 2) les enregistrements ont été collectés en plusieurs sessions sur une période de 3 mois. Le premier point implique pour notre système que les GCIs sont plus difficiles à localiser, et de surcroît que les signatures glottiques vont perdre leurs détails hautes-fréquences (qui peut contenir de l'information pertinente pour distinguer des locuteurs). Concernant le second aspect, on peut s'attendre à une plus grande variabilité intra-locuteur lorsque les sessions d'entraînement et de test sont espacées sur une longue période de temps. Les résultats que nous avons obtenus sur le corpus YOHO en utilisant les 2 signatures vocales sont présentés en Figure 6. Ces résultats sont détaillés selon la période séparant les enregistrements d'entraînement et de test. De plus, les pourcentages des cas pour lesquels le locuteur correct est reconnu en seconde ou troisième position (au

lieu d'être en première position) sont également donnés. De ce graphe il peut être remarqué que le système marche parfaitement quand les enregistrements proviennent de la même session. Au contraire, quand le test est fait dans une session ultérieure, l'identification chute brutalement jusqu'à 70%. Cette chute est essentiellement imputable à la discordance entre les conditions d'entraînement et de test. Il peut être observé que le taux d'identification décroît ensuite d'environ 5% pour toute session ultérieure. Comme attendu, ceci résulte de la plus grande variabilité du locuteur quand l'intervalle de temps entre sessions augmente. Notons aussi que, quand les conditions d'entraînement et de test diffèrent, entre 12% et 16% des locuteurs sont identifiés en seconde ou troisième position. On peut s'attendre à ce que la combinaison des signatures glottiques proposées avec des attributs basés sur le spectre de magnitude de la parole enlève l'essentiel de cette ambiguïté. Finalement, dans un but de comparaison, Gudnason et al. ont rapporté dans [5] un taux de mauvaise identification de 36.3% en utilisant les coefficients cepstraux de la source vocale (avec des enregistrements de test répartis sur les 4 sessions). En moyennant nos résultats sur la totalité des sessions, nous avons trouvé un taux de mauvaise classification de 29.3% en utilisant les 2 signatures glottiques.

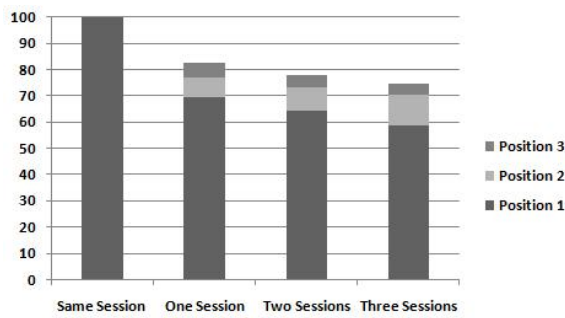


FIGURE 6: Taux d'identification (%) pour la base de données YOHO quand les sessions d'entraînement et de test peuvent être séparées sur une longue période. La proportion de locuteurs pour lesquels les signatures glottiques sont reconnues en seconde ou troisième position est également indiquée.

4. CONCLUSION

Cet article a étudié la potentialité d'utiliser des signatures glottiques en reconnaissance du locuteur. Ces signatures vocales ont été dérivées d'une analyse, pour un locuteur donné, d'un set de trames de résidu pitch-synchrones et normalisées en prosodie. Des résultats d'identification du locuteur ont été rapportés sur les bases de données TIMIT et YOHO. Dans ces expériences, les signatures glottiques proposées ont donné de meilleurs résultats que d'autres études similaires basées sur des attributs glottiques. Cependant, il a été montré que la performance est dégradée quand les sessions d'entraînement et de test sont espacées dans le temps.

Plusieurs améliorations pourraient être apportées à l'approche actuelle. En effet, les résultats ont été obtenus en utilisant *uniquement* les signatures glottiques proposées. D'après l'évidence d'une complémentarité entre les MFCCs et les caractéristiques basées sur l'excitation ([8], [10], [5]), il est raisonnable de penser qu'incorporer les signatures vocales proposées dans un

système de reconnaissance du locuteur mènerait à une amélioration appréciable. Deuxièmement, l'application d'une compensation de canal pourrait réduire la discordance entre les sessions d'entraînement et de test. En effet, différentes conditions d'enregistrement imposent différentes caractéristiques au signal de parole. Parmi celles-ci, les différences en réponse de phase peuvent affecter sensiblement l'estimation des signatures glottiques (puisque l'information du résidu est essentiellement contenue dans sa phase). Ces deux possibles améliorations sont l'objet d'un travail en cours.

5. REMERCIEMENTS

Thomas Drugman est supporté par le Fonds National de la Recherche Scientifique (FNRS).

RÉFÉRENCES

- [1] J. Campbell. Testing with the yoho cd-rom voice verification corpus. In *Proc. ICASSP*, pages 341–344, 1995.
- [2] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.
- [3] 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.
- [4] W. Fisher, G. Doddington, and K. Goudie-Marshall. The darpa speech recognition research database : Specifications and status. In *Proc. DARPA Workshop on Speech Recognition*, pages 93–99, 1986.
- [5] J. Gudnason and M. Brookes. Voice source cepstrum coefficients for speaker identification. In *Proc. ICASSP*, pages 4821–4824, 2008.
- [6] I. Karlsson. Glottal waveform parameters for different speaker types. In *STL-QPSR*, volume 29, pages 61–67, 1988.
- [7] L.G. Kersta. Voiceprint identification. In *Nature* 196, pages 1253–1257, 1962.
- [8] S. Murty and B. Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. In *IEEE Signal Processing Letters*, volume 13, pages 52–55, 2006.
- [9] [Online]. The snack sound toolkit. In <http://www.speech.kth.se/snack/>.
- [10] 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.
- [11] D.A. Reynolds. An overview of automatic speaker recognition technology. In *Proc. ICASSP*, volume 4, pages 4072–4075, 2002.
- [12] P. Thevenaz and H. Hugli. Usefulness of the lpc-residue in text-independent speaker verification. In *Speech Communication*, volume 17, pages 145–157, 1995.