Surveillance vocale de réseaux de communication professionnels par la reconnaissance du locuteur

Alexandre Preti[1,2], Bertrand Ravera[2], François Capman[2], and Jean-François Bonastre[1]

[1] Université d' Avignon, LIA
339 chemin des meinajaries, Agroparc BP 1228, 84911, Avignon Cedex 9, France
{alexandre.preti, jean-francois.bonastre}@univ-avignon.fr
[2] THALES Multimedia Processing
147 Bd Valmy, 922047, Colombes, France
{bertrand.ravera., francois.capman}@fr.thalesgroup.com

ABSTRACT

Even if the speaker recognition field is very dynamic, few studies concern the constraints linked to the use of a speaker recognition system inside a professional telecommunication network. This paper deals with this problem and proposes some adaptation of such system in the focus of a real world network monitoring application. Both real-time constraints and distributed architectures are investigated. We propose a frame-by-frame on-line processing for feature extraction, frame selection and normalization. The links between the network speech coder and the speaker recognition system are also investigated, for both the ETSI TETRA speech codec (at 4600 bit/sec) and the NATO STANAG 4591 (at 2400 bit/sec). The proposed solutions are compared with a classical unconstrained front-end (off-line processing).

Keywords: speaker verification, front-end, real-time processing.

1. INTRODUCTION

L'authentification par signature vocale suscite beaucoup d'intérêt dans de multiples domaines, notamment dans le renforcement de la sécurité des réseaux de communication dits professionnels (militaires ou privés). Les utilisateurs accèdent au réseau par un terminal d'acquisition mobile, un contrôle d'identité lors de l'utilisation de ce terminal peut permettre de détecter des utilisateurs non autorisés. Pour cela il est nécessaire d'implanter un système de reconnaissance du locuteur déporté en ligne qui, en temps réel, authentifie la voix de l'utilisateur. Une implantation sur le terminal semble difficile du fait des ressources limitées. Dans cette optique les caractéristiques d'un réseau de communication professionnel doivent être prises en compte. La voix est codée à bas débit pour éviter une consommation trop importante de bande passante et la transmission des communications repose sur une architecture distribuée. Dans cet article nous considérons différentes configurations pour le traitement des paramètres extraits du signal de parole, cette phase est communément appelée le « traitement amont». Pour une surveillance en temps réel de l'authenticité de la voix de l'utilisateur il est nécessaire d'adapter le système de vérification du locuteur qui, dans la plupart des cas, est conçu pour un traitement hors-ligne.

Ce travail présente une solution complète pour l'extraction des paramètres acoustiques, adaptée à un traitement en temps réel, en ligne. Le système de reconnaissance du locuteur qui a servi de base à ce travail est présenté dans la section 2. L'extraction des paramètres acoustique classiquement utilisée pour un traitement hors-ligne est décrite dans la section 3. La section 4 présente les contraintes des réseaux de communication professionnels. La solution d'extraction et de normalisation en ligne, temps réel, des paramètres acoustiques est présentée dans la section 5. La section 6 est dédiée à la validation expérimentale et à l'analyse de ses résultats.. Une conclusion sur le travail présenté termine cet article.

2. DESCRIPTION DU SYSTÈME DE RECONNAISSANCE DU LOCUTEUR

Le système de reconnaissance du locuteur ayant servi de base à ce travail a été développé au LIA [1]. Il utilise une modélisation statistique des paramètres du signal de parole à base de mélanges de Gaussiennes. Le modèle du monde est estimé par l'algorithme EM sur plusieurs heures d'enregistrement. Comme l'estimation du modèle du monde est faite avant la mise en route du système de vérification il n'est pas nécessaire d'implanter un traitement en ligne pour sa création. Les modèles de locuteurs sont dérivés du modèle du monde par adaptation *Maximum A Posteriori* [2]. Le modèle du monde comme les modèles de locuteur utilisent 512 Gaussiennes à matrice de covariance diagonale. Aucune normalisation de score n'est appliquée dans les résultats présentés.

3. TRAITEMENT AMONT DES PARAMÈTRES

Comme dans la plupart des systèmes de reconnaissance état de l'art, nous utilisons une analyse cepstrale pour extraire les paramètres du signal de parole. Une détection d'activité vocale permet d'éliminer les trames de silence ou de faible énergie du signal qui dégradent les performances de la reconnaissance vocale [3]. Nous utilisons une classification à base de Gaussiennes pour déterminer les trames de plus haute énergie à sélectionner [3]. Les paramètres cepstraux sont ensuite normalisés selon une loi de moyenne nulle et de variance unité pour réduire les effets dus au canal de transmission. Ces trois étapes utilisent des enregistrements complets (fichiers) pour estimer les paramètres nécessaires.

3.1. Le standard ETSI Aurora

Le standard ETSI Aurora [4] a été créé à l'origine pour la reconnaissance de la parole sur des architectures distribuées. Le terminal a pour charge d'extraire les paramètres cepstraux et de les transmettre après compression. Le flux compressé est ensuite reçu par un serveur distant pour effectuer la reconnaissance. Les dégradations dues au codage bas débit de la voix ou au codage canal sont ainsi évitées. Le standard Aurora propose 13 paramètres cepstraux statiques et un paramètre d'énergie, calculés toutes les 10 ms. Un étage de quantification est appliqué aux paramètres cepstraux. Le vecteur de paramètres est étendu en ajoutant les dérivées premières et secondes des cepstres.

3.2. La sélection de trames

Cette étape permet d'éliminer les trames inutiles comme le bruit, le silence ou les trames de faible énergie. Ces dernières sont en effet connues pour dégrader les performances de reconnaissance [4]. Le critère d'énergie est utilisé pour sélectionner les trames utiles. Un modèle GMM à trois composantes est appris indépendamment sur chaque enregistrement. Les trames sélectionnées sont celles appartenant à la Gaussienne de moyenne maximale (soit correspondant aux trames de plus grande énergie).

3.3. La normalisation des paramètres cepstraux

Les paramètres cepstraux sont normalisés par l'application d'une soustraction de la moyenne cepstrale et d'une normalisation de la variance. Ce processus est appliqué sur chaque fichier d'enregistrement. Il a pour but de réduire les bruits stationnaires de convolution et de réduire les effets de la variation entre les environnements des sessions d'apprentissage et de tests. Les meilleurs résultats de reconnaissance sont obtenus lorsque les paramètres de moyenne et de variance sont estimés sur un enregistrement entier (fichier).

4. CONTRAINTES DES RÉSEAUX DE COMMUNICATION

La chaîne de transmission est une des caractéristiques majeures des réseaux de communications professionnels. L'architecture typique de ce type de réseau est décrite dans la figure 1. L'extraction des paramètres peut être réalisée à différents endroits de l'architecture. Ils peuvent être extraits sur le terminal (point 1 figure 1). Mais, dans ce cas précis, ils doivent être encodés et transmis via un lien « data » comme le prévoit le standard Aurora. Il y a alors transmission du flux audio pour la communication et du flux data pour les paramètres. Cette configuration est considérée comme la référence en termes de performance puisque le signal originel, non codé, est utilisé. L'extraction des paramètres peut également être réalisée après le décodage bas débit de la voix (point 3 figure 1). Enfin, une configuration optimisée (point 2 figure 1) consiste à extraire les paramètres dans le domaine compressé à partir du train binaire émis. Cette configuration ne nécessite pas le signal décodé, elle se montre économe en termes de bande passante comme de ressource de calcul. Dans ce travail, nous prenons comme exemple deux vocodeurs très répandus sur les réseaux de communications professionnels, le TETRA (Terrestrial Trunked Radio) [5] de débit 4,6 kbit/s et le MELP (OTAN STANAG-4591 Mixed Excitation Linear Prediction) [6] de débit 2,4 kbit/s. Le codage bas débit de la voix est réalisé pour ces deux codeurs par une analyse LPC. Les coefficients de prédiction sont quantifiés puis émis (quantification vectorielle). Il est alors possible d'extraire les paramètres cepstraux (LPCC) à partir de ces coefficients.

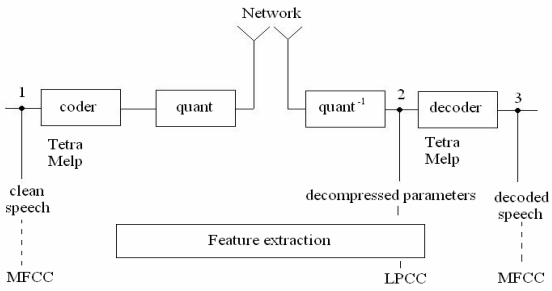

Figure 1 : Architecture d'un réseau de communication professionnel.

5. IMPLÉMENTATION TEMPS RÉEL

Cette section décrit les modifications apportées à chaque étape du traitement amont des paramètres pour permettre une implémentation temps réel du système de reconnaissance du locuteur. Les étapes concernées sont la détection d'activité vocale, l'extraction et la normalisation des paramètres. Ce travail a pour but l'intégration d'un monitoring réseau temps réel sur des réseaux de communications professionnels où la voix est codée à bas débit par les vocodeurs TETRA ou MELP.

5.1. Détection d'activité vocale

Comme présenté en section 3.2, le système de reconnaissance du locuteur de référence utilise un détecteur d'énergie à base de GMM pour la sélection des trames. Pour chaque fichier, ce processus utilise l'enregistrement complet pour estimer le niveau d'énergie minimal des trames à sélectionner. Cette solution n'est pas adaptée à l'application visée car elle ne permet pas un traitement temps réel, en ligne. Nous proposons l'utilisation d'une détection d'activité vocale (DAV) basée sur le standard Aurora. Le standard Aurora implémente un DAV à la trame avec un retard d'exécution de seulement 6 trames. Elle utilise une mesure d'accélération de l'énergie en sous bandes robuste au bruit pour chaque trame. Des premiers résultats montrent que la DAV Aurora sélectionne plus de trames (environ 80 % du signal) que le détecteur à base de Gaussiennes (environ 50% du signal). Comme les performances de reconnaissance sont altérées par l'ajout de trames de bruit, de silence ou de faible énergie, nous proposons de ne

sélectionner que les trames classées comme voisées (information intégrée dans le standard Aurora) pour diminuer le nombre de trames sélectionnées.

5.2. Soustraction cepstrale et normalisation de la variance

Pour chaque dimension de l'espace des paramètres acoustiques, les coefficients sont normalisés en utilisant une soustraction de la moyenne et une normalisation de la variance. Cette normalisation nécessite une estimation robuste des paramètres de moyenne et de variance caractérisant chaque coefficient. Dans le système de référence, ces paramètres sont estimés a posteriori, fichier par fichier, en utilisant toutes les trames de l'enregistrement courant. Des solutions utilisant une fenêtre glissante sur le signal sont proposées dans [7, 8] et présentent des performances semblables à une estimation sur la totalité du signal enregistré. En suivant cette approche nous avons implémenté une normalisation à la trame basée sur une estimation avec facteur d'oubli [8]. Cette procédure consiste en l'utilisation d'une fenêtre d'initialisation de N trames (seules les trames sélectionnées par la DAV sont prises en considération). Les trames appartenant à cette fenêtre sont normalisées lorsque la fenêtre est pleine. Ensuite la normalisation opère trame à trame sans aucun délai. Les paramètres de normalisation sont calculés sur la fenêtre d'initialisation puis mis à jour continuellement pour chaque nouvelle trame selon les équations suivantes (2,3) :

$$\sigma^2 = \beta\sigma^2 + (1-\beta)t_i^2 \quad (2)$$

$$\mu = \beta\mu + (1-\beta)t_i \quad (3)$$

$$\begin{cases} \beta = 1 \text{ ; indices de trames de [0;N]} \\ \beta = (N-1)/(N) \text{ ; sinon} \end{cases}$$

6. BASE DE DONNEES ET RESULTATS

Cette section présente la base de données et les protocoles utilisés pour la validation expérimentale de notre étage d'extraction et de normalisation « en ligne » des paramètres acoustiques. Les résultats expérimentaux sont également présentés et commentés.

6.1. Base de données BREF

La base d'enregistrements BREF [9] est composée de phrases de langue française lues dans un environnement non bruité. Le modèle du monde indépendant du genre est appris avec un sous ensemble de 40 locuteurs. 40 locuteurs sont utilisés comme clients du système de reconnaissance (20 hommes et 20 femmes). Enfin 35 autres locuteurs sont utilisés comme imposteurs. Au total environ huit mille tests clients sont effectués sur un total d'environ quatre-vingt dix mille tests. Pour respecter au mieux les contraintes des réseaux de communications professionnels nous utilisons des enregistrements de 1 minute pour l'apprentissage des modèles clients et de 8 secondes pour la phase de test. Les performances sont évaluées par des courbes DET (avec mesures EER et NIST DCF).

6.2. Détection d'activité vocale

Cette section présente les résultats de l'utilisation de la DAV Aurora modifiée pour intégrer l'information de voisement. L'extraction des paramètres est identique à celle présentée dans la section 3.1. La figure 2 présente les courbes DET de performances de la DAV modifiée et de la DAV de référence. Les meilleures performances sont obtenues avec la DAV Aurora compatible temps réel (traitement à la trame). L'utilisation du critère de trames voisées du standard Aurora apporte un gain en performance de 28 % relatif pour les mesures DCF et EER comparé à la DAV référence (traitement sur fichiers).

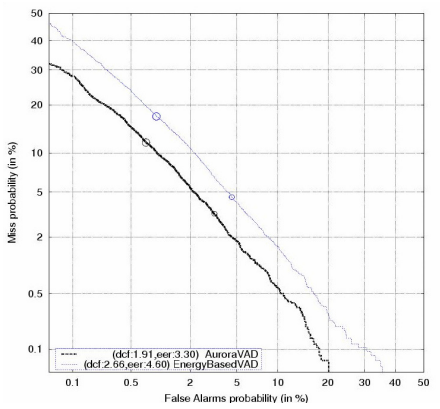

Figure 2 : courbes DET pour DAV Aurora (DCF: 1.91, EER: 3.30) et la DAV de référence (DCF: 2.66 , EER: 4.60).

6.3. Normalisation des paramètres

Cette section compare les résultats obtenus avec la normalisation en ligne proposée en section 5.2 et avec la normalisation de référence fonctionnant en mode « fichier ». Deux expériences avec deux tailles de fenêtre d'initialisation (150 et 300 trames) sont réalisées. Les résultats sont présentés sur la figure 3. Les résultats obtenus avec une fenêtre d'initialisation de 300 trames (3.49% EER) sont très proches de ceux obtenus avec une normalisation sur fichiers (3.30% EER). Cependant il est à noter que comme la durée des enregistrements est de seulement 8 secondes (environ 500 trames après DAV) 60% des trames sélectionnées se trouvent dans la fenêtre d'initialisation.

6.4. Résultats dans le domaine compressé

Cette section présente les résultats obtenus en extrayant les paramètres dans le domaine compressé, soit directement à partir des paramètres des codeurs. Pour chaque expérience, tous les enregistrements sont d'abord codés en utilisant le vocodeur TETRA ou MELP puis les LPCC sont extraits du train binaire codé. Contrairement à l'extraction des paramètres de référence (cf. section 3.1) l'échelle de Mel n'est pas appliquée aux LPCC. A des fins de comparaison nous proposons les résultats de l'extraction des paramètres sur le signal décodé (point 3 figure 1) et sur le signal non codé (point 1 figure 1). L'extraction des paramètres de référence est alors utilisée

(cf. section 3). Ces résultats sont listés dans la table 1. L'analyse de ces résultats démontre que lorsque les paramètres cepstraux (LPCC) sont extraits du train binaire des codeurs, le codeur TETRA obtient de meilleures performances que la solution utilisant le signal décodé (issu de ce codeur). Le codeur MELP montre un profil différent, il obtient en effet de meilleurs résultats en travaillant sur le signal décodé (issu du MELP) qu'en travaillant avec le train binaire. Notons que le MELP sur le signal décodé permet d'approcher les performances du système de référence. Nous pouvons émettre l'hypothèse que la reconstruction du signal en sortie du décodeur MELP ajoute de l'information utile à l'analyse cepstrale, notamment avec la transmission des dix premiers modules de la transformée de Fourier du signal résiduel de prédiction pour améliorer la synthèse du signal. Enfin, on peut noter que la perte due au codage bas débit de la voix est évaluée à 20% pour les mesures EER et DCF (différences mesurées entre les expériences sur le signal non codé, point 1 figure 1 et les expériences dans le domaine compressé, point 2 figure 3).

Table 1 : Mesures DCF et EER pour les expériences dans le domaine compressé et sur le signal décodé ou originel (les nombres entre parenthèses se réfèrent à la figure 1).

Expérience	**DCF**	**EER**
LPCC Tetra (2)	3.55	5.57
LPCC Melp (2)	3.55	5.58
TETRA décodé (3)	3.84	7.29
MELP décodé(3)	2.94	5.20
Signal non codé	2.78	4.58

7. CONCLUSION ET PERSPECTIVES

Dans ce travail nous analysons les contraintes induites par l'architecture des réseaux de communication pour la mise en place d'un système de surveillance basé sur la reconnaissance du locuteur. Nous proposons une solution complète pour le traitement des paramètres en temps réel sur ce type de réseaux de communication. La DAV Aurora à la trame apporte un gain de 28% relatif pour les mesures EER et DCF comparé à la DAV de référence. La normalisation en ligne égale les performances du traitement sur fichier. De plus nous évaluons l'impact de l'extraction de paramètres dans le domaine compressé pour les vocodeurs MELP et TETRA. Ainsi les meilleures performances pour le vocodeur TETRA sont obtenues dans le domaine compressé au contraire du vocodeur MELP qui offre une meilleure resynthèse du signal. Ces résultats encouragent la mise en place d'un système de reconnaissance du locuteur en temps réel sur les réseaux de communications professionnels sans perte de performance. De futurs travaux viseront à déterminer un seuil de décision optimal en fonction, par exemple, de la quantité de trames accumulées par le système et de qualité de celles-ci (critère de rapport signal à bruit).

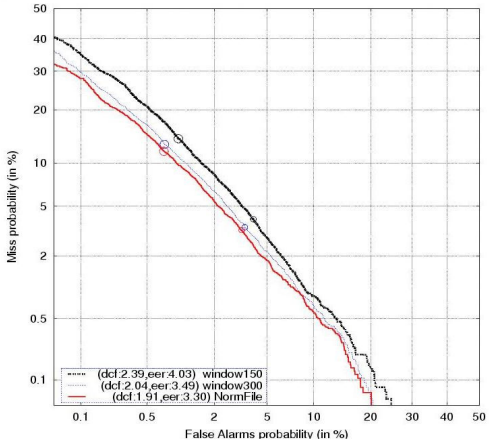

Figure 3 : Courbes DET pour des tailles de fenêtre d'initialisation de 150 and 300 trames et la référence (traitement sur fichiers).

Au vu des performances des récentes techniques de compensation des variabilités intra locuteur et inter sessions (Latent Factor Analysis) des travaux seront menés pour implémenter ces techniques en ligne.

REFERENCES

[1] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve and J. Mason. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. *In Speaker Odyssey*, South Africa, January 2008.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing,* vol. 4, pp. 430–451, 2004.

[3] L. Besacier, J-F. Bonastre, C. Fredouille. Localization and selection of speaker-specific information with statistical modeling. *Speech Communication*, vol. 31, pp.89-106, 2000.

[4] Aurora ETSI ES 202 212 V1.1.2 (2005-11).

[5] Tetra ETSI EN 300 395-1 v1.3.1 (2005-06)

[6] L.M Supplee, R.P. Cohn, J.S. Collura, and A.V. McCree. MELP: the new federal standard at 2400 bps. *In Proc IEEE ICASSP*, Munich, Germany, April 1997, vol.2, pp. 1591-1594.

[7] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. *In Proc. ISCA Workshop on Speaker Recognition : A Speaker Odyssey*, June 2001.

[8] P. Pujol, D. Macho, C. Nadeu. On Real-Time Mean-and-Variance Normalization of Speech Recognition Features. I*n ICASSP 2006 Proceedings*. Toulouse France, 2006.

[9] L. Lamel, J.-L. Gauvain, M. Eskenazi, BREF, a large vocabulary spoken corpus for French. I*n EUROSPEECH,* 1991, 505-508.