

Système d'élagage temps-fréquence pour l'identification du locuteur

L. Besacier, J.F. Bonastre

LIA/CERI

339, chemin des Meinajaries BP 1228 - 84911 Avignon Cedex 9 (France)
(laurent.besacier, jean-françois.bonastre)@lia.univ-avignon.fr
laurent.besacier@imt.unine.ch

ABSTRACT

This work is an attempt to refine decisions in speaker identification. A test utterance is divided into multiple time-frequency blocks on which a normalized likelihood score is calculated. Instead of averaging the block-likelihoods along the whole test utterance, some of them are rejected (pruning) and the final score is computed with a limited number of time-frequency blocks. The results obtained in the special case of time pruning, added to former results obtained in the special case of frequency pruning, lead the authors to experiment a joint time and frequency pruning approach. The optimal percentage of blocks pruned is learned on a tuning data set with the minimum identification error criterion. Validation of the time-frequency pruning process on 567 speakers leads to a significative error rate reduction on TIMIT and NTIMIT (up to 41% reduction on TIMIT) for short training and test durations. Finally, experiments in the case of simulated noise degradation have shown that this approach is a very efficient way to deal with partially corrupted speech.

1. INTRODUCTION

En reconnaissance automatique du locuteur, la modélisation mono-gaussienne du signal de parole a été largement remplacée par une modélisation multi-gaussienne (GMM) [Rey95], sensée représenter plus finement le signal. Les mixtures de gaussiennes peuvent être vues comme une *coopération* de modèles puisqu'elles reviennent à une somme pondérée de densités gaussiennes. Dans cet article, le travail présenté se rapporte plutôt à une *compétition* de modèles : différents modèles mono-gaussiens (associés à une sous-bande particulière) sont appliqués sur le signal de test et la décision est prise à partir du ou des N-meilleurs modèles.

Plus précisément, une portion de test est divisée en blocs temps-fréquence, chacun d'entre-eux correspondant à une sous-bande fréquentielle précise et à un segment temporel particulier. Pendant la phase de reconnaissance, les scores de vraisemblance calculés sur chaque bloc sont accumulés afin d'obtenir un score sur la portion de test totale. Cet

article propose une stratégie d'élagage de certains scores de vraisemblance pour renforcer la décision finale.

Le but visé est de sélectionner automatiquement les parties du signal les plus aptes à identifier un locuteur. Cette approche augmente la robustesse de la décision dans le cas d'un bruit localisé sur une zone temps-fréquence précise, les blocs les moins fiables étant supprimés. Dans le cas de parole non corrompue, le procédé d'élagage permet de bâtir la décision finale sur les informations les plus pertinentes et d'ignorer les parties non informatives du signal.

Dans le *paragraphe 2*, un formalisme est proposé pour décrire le système d'élagage temps-fréquence. Son potentiel est alors mis en évidence pour le cas particulier de l'élagage temporel (*paragraphe 3*). Dans le *paragraphe 4*, nous proposons une expérience ayant pour but d'évaluer le pourcentage optimal (en terme de performances d'identification) de blocs temps-fréquence supprimés. La méthode décrite dans cet article est ensuite validée sur 567 locuteurs des bases de données TIMIT et NTIMIT (*paragraphe 5*). Enfin, dans le *paragraphe 6*, nous résumons les principaux résultats obtenus et montrons le potentiel de notre système sur des données partiellement bruitées.

2. FORMALISME

2.1 Modélisation mono-gaussienne par segments

La modélisation mono-gaussienne pour la reconnaissance automatique du locuteur est décrite plus précisément dans [Bim95] et [Gis94]. Soit $\{x_t\}_{1 \leq t \leq M}$, une séquence de M vecteurs résultant de l'analyse p -dimensionnelle d'un signal de parole prononcé par le locuteur X . Ces vecteurs sont modélisés par le vecteur moyen \bar{x} et la matrice de covariance X :

$$\bar{x} = \frac{1}{M} \sum_{t=1}^M x_t \quad \text{and} \quad X = \frac{1}{M} \sum_{t=1}^M (x_t - \bar{x})(x_t - \bar{x})^T \quad (1)$$

De même, pour le signal de parole prononcé par le locuteur Y , une séquence de N vecteurs $\{y_t\}_{1 \leq t \leq N}$ peut être extraite. En supposant que tous les vecteurs acoustiques prononcés par le locuteur X sont distribués suivant une fonction

gaussienne, la vraisemblance du vecteur y_i prononcé par le locuteur Y est :

$$G(y_i / X) = \frac{1}{(2\pi)^{p/2} (\det X)^{1/2}} e^{-\frac{1}{2}(y_i - \bar{x})^T X^{-1} (y_i - \bar{x})} \quad (2)$$

En admettant de plus que tous les vecteurs y_i sont des observations indépendantes, la log-vraisemblance moyenne de $\{y_i\}_{i+1 \leq i \leq i+T}$ sur un segment de T trames, peut être écrite :

$$\overline{G_X}(y_{i+1}^{i+T}) = \frac{1}{T} \log G(y_{i+1} \dots y_{i+T} / X) = \frac{1}{T} \sum_{i=1}^T \log G(y_{i+i} / X) \quad (3)$$

2.2 Modélisation multibande

Le modèle du locuteur X à 'K-sous-bandes', peut être obtenu à partir d'un modèle mono-bande initial :

$$M_X(K) = \left\{ (X^1, \bar{x}^1), \dots, (X^k, \bar{x}^k), \dots, (X^K, \bar{x}^K) \right\} \quad (4)$$

où le locuteur X est modélisé sur la k -ième sous-bande avec la matrice de covariance X^k et le vecteur moyen \bar{x}^k .

X^k est un sous-bloc de la matrice de covariance X et \bar{x}^k est un sous-vecteur du vecteur moyen \bar{x} . Ainsi, les quantités définies dans (2) et (3) peuvent être respectivement écrites sur la k -ième sous-bande :

- $G^k(y_i / X)$ vraisemblance du vecteur acoustique y_i sur la k -ième sous-bande,

- $\overline{G_X}^k(y_{i+1}^{i+T})$ log-vraisemblance moyenne du segment $\{y_i\}_{i+1 \leq i \leq i+T}$ sur la k -ième sous-bande.

2.3 Système d'élagage temps-fréquence

Le système final combine les deux aspects *segmental* et *multibande*. Une portion de test de N trames est divisée en K sous-bandes et n segments de T trames chacun, tel que $N=nT$ (Fig. 1). Bien que non représenté sur la Fig. 1, un recouvrement entre les sous-bandes ou entre les segments est possible. Pour chaque couple (t,k) correspondant à la k -ième sous-bande du t -ième segment, un score normalisé [Gis94] $h_X^k(y_{i+1}^{i+T})$ (homogène au moins logarithme d'un rapport de vraisemblance) est calculé

$$h_X^k(y_{i+1}^{i+T}) = \max_{Z \neq X} \overline{G_Z}^k(y_{i+1}^{i+T}) - \overline{G_X}^k(y_{i+1}^{i+T}) \quad (5)$$

h est également appelé 'fonction discriminante' [Fuk90] (p.52) puisque si $h < 0$, la vraisemblance du locuteur X est plus grande que celle de tous les autres locuteurs sur le bloc (t,k) , et X est reconnu sur ce bloc. Si $h > 0$, le locuteur reconnu sur le bloc n'est pas le locuteur X .

Les $(n \cdot K)$ scores normalisés sont accumulés afin d'obtenir le score de chaque modèle de référence :

$$\tilde{h}_X(y_1^N) = ACC \left[h_X^k(y_{i+1}^{i+T}) \right]_{\substack{i \in [0, n-1] \\ k \in [0, K-1]}} \quad (6)$$

ACC est la fonction d'accumulation ; notons que $h_X(y_1^N)$ est strictement équivalente à la mesure mono-gaussienne

standard (bande totale) lorsque $n=1$ et $K=1$ (i.e. portion de test considérée dans sa globalité).

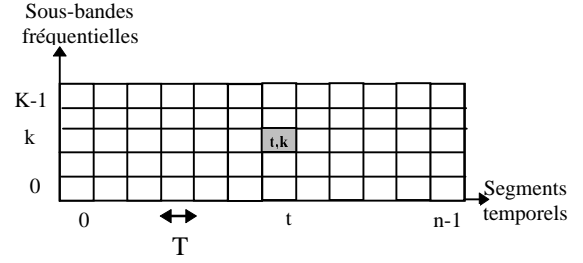


Figure 1 : Division d'une portion de test en n segments de K sous-bandes ($n \cdot K$ blocs au total)

L'utilisation de différents blocs temps-fréquence nous permet de supprimer ou désaccentuer des blocs correspondant à un événement anormal ou des blocs peu représentatifs du locuteur cible. La fonction d'accumulation profitant de cette segmentation est donnée dans (7) :

$$\tilde{h}_X(y_1^N) = \arg \min_{(p,q)} \left[\frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q h_X^k(y_{i+1}^{i+T}) \right]_{\substack{i \in [0, n-1] \\ k \in [0, K-1]}} \quad (7)$$

Dans ce cas, un score est obtenu avec la moyenne des $p \cdot q$ plus petits scores de blocs pour chaque locuteur, avec $p < n$ (n nombre de segments dans une portion de test) et $q < K$ (K nombre de sous-bandes au total). Pour une même portion de test, les blocs sélectionnés peuvent varier d'un locuteur de référence à l'autre.

Finalement, deux cas particuliers peuvent être dérivés de ce formalisme général :

- si $K=1$ et $n > 1$, il s'agit d'une méthode d'identification avec normalisation des scores au niveau segmental [Mar96] ; seul l'élagage temporel est alors considéré.
- si $n=1$ et $K > 1$, il s'agit d'une approche 'multibande' [Bes97] et seul l'élagage fréquentiel est considéré.

3. ELAGAGE TEMPOREL

3.1 Conditions expérimentales

Pour les expériences présentées ici, nous avons utilisé les bases TIMIT (parole normale) et NTIMIT (parole téléphonique). Le module d'analyse acoustique caractérise le signal toutes les 10ms par un vecteur de 24 coefficients banc de filtres. Les conditions d'analyse sont identiques à celles décrites dans [Bim95] et [Bes97]. Pour la base TIMIT, nous gardons les 24 coefficients des vecteurs spectraux. Pour NTIMIT, nous supprimons les 2 premiers coefficients et les 7 derniers qui sont en dehors de la bande téléphonique (300-3400Hz).

Le protocole d'apprentissage et de test est commun à toutes les expériences [Bes97]. Nous utilisons des durées courtes (6s d'apprentissage et 3s de test) dans le but de montrer

l'intérêt de la procédure d'élagage même lorsque peu de parole est disponible. Tous les tests sont réalisés dans le cadre d'une identification du locuteur en ensemble fermé indépendante du texte. La règle de décision est la règle du maximum de vraisemblance.

3.2 Elagage des trames

Lorsque $K=1$, le système fonctionne en mode mono-bande et seul l'élagage temporel est possible. Dans ce cadre particulier, l'influence du nombre p de segments conservés est étudiée lorsqu'un segment est composé d'une seule trame ($T=1$). Les résultats sont reportés dans la Fig. 2.

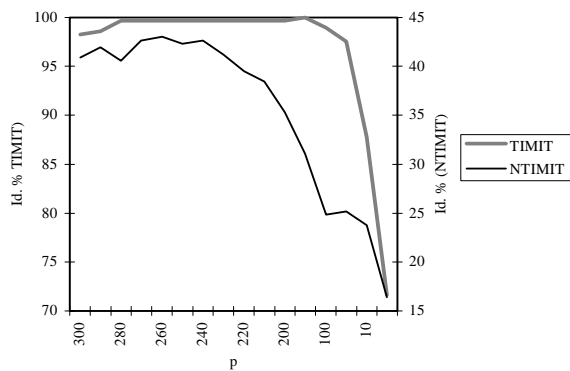


Figure 2 : Elagage temporel - 6s App./3s Test - (300-p) trames éliminées - T=1 - 63 locuteurs - 286 tests

Pour les deux bases, les résultats optimaux sont obtenus quand des trames sont supprimées : id.=100% pour $p=150$ sur TIMIT et id.=43% pour $p=260$ sur NTIMIT. Ceci montre que la sélection de l'information est importante puisque certaines trames de test peuvent contaminer le score final. De plus, il est intéressant de noter qu'une performance acceptable est obtenue sur TIMIT quand une seule trame est conservée (choix fait *a priori*) pour le calcul de la vraisemblance de chaque locuteur de référence (71.63% id.).

3.3 Les trames rejetées sont elles toujours les mêmes ?

Lors d'une mesure de ressemblance entre un signal de test et le modèle d'un locuteur donné, les trames conservées sont sélectionnées suivant le modèle concerné. Les scores de vraisemblance ne sont donc pas forcément tous calculés à partir des mêmes segments temporels. La Fig. 3 montre la distribution des trames suivant leur fréquence de sélection. Ainsi, lorsque $N_{select}=63$ (resp. $N_{select}=0$), les trames correspondantes sont conservées (resp. rejetées) par chacun des 63 modèles.

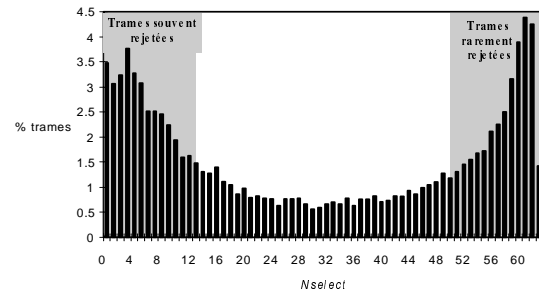


Figure 3 : Distribution des trames suivant leur fréquence de sélection - TIMIT - $p=150$ (la moitié des trames rejetées) - 63 locuteurs

Le profil des résultats suggère une certaine cohérence dans l'information convoyée par les trames puisque si une trame est rejetée, elle l'est par la majorité des modèles.

4. ELAGAGE TEMPS-FREQUENCE

4.1 Elagage fréquentiel

Une étude précédente [Bes97], a montré l'intérêt d'une procédure de sélection fréquentielle de l'information. Il est apparu que certaines bandes perturbaient le résultat final. De même, la quantité d'information spécifique du locuteur semble très inégalement répartie dans le spectre fréquentiel : des différences de performances allant de 25% d'identification à 5% d'identification ont pu être mesurées sur TIMIT pour des bandes constituées de 4 canaux consécutifs.

4.2 Elagage temps-fréquence

Cette expérience est réalisée avec une architecture en 24 sous-bandes de 20 canaux (architecture 24x20) pour TIMIT et une architecture en 15 sous-bandes de 11 canaux (15x11) pour NTIMIT. Les autres conditions expérimentales sont les mêmes que celles décrites dans la section 3.1. La taille d'un segment est $T=1$ (i.e. 1 segment=1 trame). Pour une durée de test de 3s (300 trames), le nombre total de blocs temps-fréquence est donc 7200 (300×24) pour TIMIT et 4500 (300×15) pour NTIMIT. L'influence, sur les performances, du nombre de blocs sélectionnés pq est étudiée. Les résultats sont reportés dans la Fig. 4.

Pour les deux bases, les meilleurs résultats sont obtenus lorsque des blocs sont supprimés : id.=100% pour $pq=3500$ ou 4500 sur TIMIT et id.=41.95% pour $pq=3900$ sur NTIMIT. Cependant, dans ces conditions d'expérience, il est difficile de voir le réel bénéfice d'une approche conjointe d'élagage temps-fréquence par rapport à une approche d'élagage uniquement temporel.

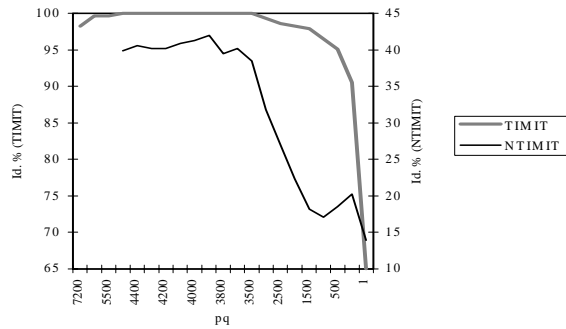


Figure 4 : Elagage temps-fréquence - 6s App./3s Test - architecture 24x20 pour TIMIT et architecture 15x11 pour NTIMIT - T=1 - 63 locuteurs - 286 tests

5. VALIDATION

Les valeurs des paramètres p (élagage temporel, *Section 3.2*) et pq (élagage temps-fréquence, *Section 4.2*) sélectionnées précédemment (à partir d'un sous ensemble composé de 63 locuteurs, de TIMIT et NTIMIT) ont été utilisées pour valider le bénéfice de la procédure d'élagage. Des tests d'identification du locuteur ont été réalisés sur un ensemble de validation distinct de l'ensemble de réglage. Ce jeu de validation est constitué par les 567 locuteurs restant de TIMIT et NTIMIT.

Les taux d'identification obtenus sont donnés dans la *Table 1*. Pour les deux bases, l'amélioration des performances est significative. La procédure d'élagage temps-fréquence permet une réduction du taux d'erreur de 41% sur TIMIT par rapport à la mesure de vraisemblance gaussienne conventionnelle. Ce bénéfice est cependant moins évident sur NTIMIT. Ceci peut s'expliquer par le plus faible nombre de bandes fréquentielles (causé par la bande passante totale réduite). De plus, l'estimation du pourcentage de blocs à rejeter reste empirique et probablement très dépendante de la base utilisée.

Table 1 : Validation de la procédure d'élagage sur TIMIT et NTIMIT (6s App./3s test - 567 locuteurs - 2639 tests)

	REFERENCE	ELAGAGE TEMPOREL	ELAGAGE TEMPS-FREQUENCE
TIMIT	n=1;K=1	K=1;p=150;T=1	K=24;pq=4500;T=1
Id. %	91.66	94.20	95.14
NTIMIT	n=1;K=1	K=1;p=260;T=1	K=15;pq=3900;T=1
Id. %	15.91	18.64	17.77

6. DISCUSSION

Cet article a présenté un système d'élagage temporel et fréquentiel pour l'identification du locuteur. Les résultats obtenus ont montré que cette technique peut augmenter significativement les performances d'un système

d'identification du locuteur dans certaines conditions expérimentales.

Pour montrer la robustesse de cette approche dans des conditions dégradées, un bruit distribué aléatoirement sur le domaine fréquentiel total a été ajouté aux signaux de parole de la base TIMIT. Pour chaque trame, N canaux sélectionnés aléatoirement sont dégradés pour différents rapports signal-sur-bruit. L'expérience est réalisée avec N=2 ou 3 et RSB=10dB ou 20dB.

Table 2 : Taux d'identification dans le cas de parole corrompue par un bruit distribué aléatoirement sur le domaine fréquentiel total (63 locuteurs, 286 tests).

Nombre de canaux corrompus	RSB (dB)	REFERENCE n=1;K=1	ELAGAGE T-F pq=4500;T=1
3	10	13.28	71.67
2	10	23.07	84.26
3	20	44.4	95.1
2	20	58.04	98.25

Le potentiel adaptatif de la procédure d'élagage dans de telles conditions est illustré par les résultats de la *Table 2*. Dans chaque cas, l'élagage temps-fréquence dépasse largement, en performances, l'approche conventionnelle (*reference*).

7. BIBLIOGRAPHIE

- [Bes97] BESACIER, L., BONASTRE, J.F., Subband approach for automatic speaker recognition: optimal division of the frequency domain. *In Audio- and Video-based Biometric Person Authentication*, Bigün, et. al. Eds., Springer LNCS 1206, 1997.
- [Bim95] BIMBOT, F., MAGRIN-CHAGNOLLEAU, I., MATHAN, L., Second-order statistical methods for text-independent speaker identification. *Speech Communication*, n°.17(1-2), August 1995.
- [Fuk90] FUKUNAGA, K., *Statistical Pattern Recognition*. Second Edition, Academic Press, Inc., San Diego. 1990.
- [Gis94] GISH, H., SCHMIDT, M., Text independent speaker identification. *IEEE Signal Processing Magazine*, pp 18-32, October 1994.
- [Mar96] MARKOV, K., NAKAGAWA, S., Frame level likelihood normalization for text-independent speaker identification using GMMs. *In Proc. ICSLP*, pp 1764-1767, 1996.
- [Rey95] REYNOLDS, D.A., Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, vol. 17, pp 91-108, August 1995.