Vérification du locuteur : variations de performance

Juliette Kahn[1, 2, 4] Nicolas Scheffer[3] Solange Rossato[1] Jean-François Bonastre[2]
(1) LIG, (2) LIA, (3) SRI, (4) LNE
juliette.kahn@lne.fr, nicolas.scheffer@speech.sri.org,
solange.rossato@imag.fr, jean-francois.bonastre@univ-avignon.fr

RÉSUMÉ

Les progrès de performance en vérification du locuteur ces quinze dernières années sont incontestables. Les systèmes sont de plus en plus sûrs dans le sens où les taux EER ou DCF diminuent d'année en année. Pourtant, il est nécessaire de déterminer dans quelles circonstances les systèmes d'identification du locuteur sont fiables. Des études ont été menées pour analyser les performances en fonction du locuteur. Dans cet article, nous nous intérrogeons sur la variation des performances observée en fonction du signal de parole utilisé pour représenter le locuteur. Des variations très importantes des valeurs d'EER sont obtenues pour deux systèmes état de l'art. Nous proposons également une méthode pour mesurer la variation de performance propre au système. Les valeurs d'EER varient alors d'un point.

ABSTRACT

Speaker verification : results variation

Speaker verification systems have shown significant progress and have reached a level of performance that make their use in practical applications possible. Nevertheless, large differences in terms of performance are observed, depending on the speaker or the speech sample used. This context emphasizes the importance of a deeper analysis of the system's performance over average error rate. In this paper, the effect of the training excerpt on performance is investigated. The results show that the performance are highly dependent on the voice samples used to train the speaker model for two state-of-art systems. A methods to observe the variation explained by the system him-self is investigated too.

MOTS-CLÉS : Verification du locuteur, Variation de la performance.

KEYWORDS: Speaker Verification, performance.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 425–432,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les progrès des systèmes de vérification du locuteur obtenus ces quinze dernières années sont incontestables (Greenberg *et al.*, 2011).Les systèmes sont de plus en plus sûrs dans le sens où les taux EER ou DCF diminuent d'année en année pour atteindre sur des segments longs moins de 1% d'EER. La performance des systèmes est estimée à partir de deux types d'erreurs potentielles. Dans le cas d'un Faux Rejet (FR), le fichier test a bien été produit par le locuteur modélisé (test cible) mais le système considère l'hypothèse inverse. Dans le cas d'une Fausse Acceptation (FA), alors que l'auteur du fichier test est différent du locuteur cible (test imposteur), le système les considère comme identiques. Ces deux types d'erreurs sont liés par le seuil choisi pour prendre la décision. Pour comparer les performances de différents systèmes, la courbe DET, le taux d'Egal Erreur (EER) et la fonction de coût de décision (DCF) sont le plus couramment utilisés (Martin *et al.*, 1997). La courbe DET représente l'évolution des deux types d'erreur en fonction du seuil. L'EER correspond au point où le taux de FA est égal au taux de FR, le DCF introduisant une fonction de coût. Toutes ces mesures sont calculées globalement sur une grand nombre de tests.

Pourtant, pour envisager d'utiliser ses systèmes, il est nécessaire d'aller "d'un taux d'erreurs faible sur un grand nombre de test" vers la notion de fiabilité en déterminant dans quelles circonstances il est possible de faire confiance aux systèmes. Dans cet article, nous nous interrogeons sur la variation des performances observées pour deux systèmes états de l'art en fonction du signal de parole utilisé en apprentissage pour représenter chaque locuteur. Nous proposons également une méthode pour mesurer la variation de performance propre au système.

2 Variation de performance en fonction du fichier d'apprentissage

2.1 Bases de données

Pour déterminer la variation de performance due au choix du fichier d'apprentissage, nous avons utilisé le corpus NIST-08. Il est constitué d'enregistrements de parole téléphonique (conditions short 2-short 3) d'une durée de 2,5 min de la campagne NIST 2008. A l'origine, 648 fichiers d'apprentissage étaient utilisés et provenaient de 221 locuteurs hommes. Pour augmenter le nombre de modèles par locuteur, nous avons construit la base M-08 à partir des données NIST-08 par une procédure de leave-one-out. Chaque fichier d'apprentissage de NIST-08 ainsi que les fichiers test ayant servi en tests cible dans NIST-08 ont été utilisés pour créer un modèle de locuteur différent. Afin d'analyser les variations inter-modèles pour un locuteur donné, nous avons exclu les 50 locuteurs représentés par moins de 3 modèles. M-08 comprend alors 171 locuteurs représentés au total par 816 modèles. Chaque modèle est testé avec l'ensemble des fichiers sélectionnés précédemment, excepté celui ayant servi à la construction du modèle considéré, ce qui conduit à 661 416 tests imposteur et 3 624 tests cible.

2.2 Systèmes utilisés

Nous avons testé 2 systèmes différents, état-de-l'art entre 2008 et 2011, ALIZE/SpkDet et Idento.

ALIZE/SpkDet (Bonastre *et al.*, 2008) est un système de RAL basé sur le paradigme UBM/GMM (Reynolds *et al.*, 2000) développé notamment au LIA. Le modèle du monde est constitué de 1024 gaussiennes. Il inclut les techniques de Factor Analysis (Kenny *et al.*, 2005). La configuration que nous avons utilisée correspond à la soumission effectuée par le LIA lors de l'évaluation NIST 2008 (Matrouf *et al.*, 2008). Nous n'avons cependant pas effectué de normalisation des scores.

Idento (Scheffer *et al.*, 2011) est un système de RAL développé au SRI basé sur la technique des i-vector (Dehak *et al.*, 2009). Le vecteur de paramètres,de dimension 60, est composé de 20 Mel Filter Cepstral Coefficients (MFCC) ainsi que des 20 Delta et des 20 Delta-Delta.
Nous comparons les résultats obtenus en No-Norm avec ceux obtenus en ZT-Norm afin de mesurer l'influence de la normalisation sur la variation de performance.

2.3 Comparaisons de performance

Mesurer la sensibilité des systèmes automatiques aux fichiers d'apprentissage revient à quantifier les différences en terme de performance qu'amène le changement de l'enregistrement utilisé en apprentissage. La méthode que nous avons adoptée s'appuie sur les Fausses Acceptations (FA) et Faux Rejets (FR).

2.3.1 Définir FA_{ij} et FR_{ij} sur la totalité des données pour la sélection du meilleur et du pire modèle

Nous pouvons obtenir le FA_{ij} et le FR_{ij}, avec un seuil fixé à l'avance pour chaque locuteur i et chaque modèle j. Il est possible alors de déterminer pour chaque locuteur le meilleur et le pire modèle en fonction de ces taux. Le meilleur modèle est celui qui minimise la somme $FA+FR$ tandis que le pire maximise cette somme. La sélection du meilleur et du pire modèle est réalisée sur tous les fichiers de M-08.

2.3.2 Établir différentes séries de tests où seul change le fichier d'apprentissage

Une fois le meilleur et le pire modèle sélectionné pour chaque locuteur, il s'agit de mesurer l'écart de performance entre les deux modèles du même locuteur. Pour effectuer la comparaison, une cohorte de fichiers test est définie. Au lieu de comparer chaque fichier test à un fichier d'apprentissage comme cela est fait habituellement, nous comparons chaque fichier test à un ***locuteur*** dont le modèle généré par le système peut différer en fonction du fichier d'apprentissage considéré mais dont l'identité biométrique ne change pas. Une comparaison est donc ici composée d'un ***locuteur*** et d'un fichier de test. En partant des comparaisons proposées par NIST 08, nous les adaptons afin d'être certains que les 171 locuteurs interviennent dans la cohorte et qu'il sont tous testés en comparaisons cible et imposteur. Cette cohorte est le canevas qui répertorie l'ensemble des comparaisons locuteur/fichier test. Pour chaque locuteur nous pouvons choisir un fichier d'apprentissage dont nous connaissons *a priori* la performance local $FA_{ij}+FR_{ij}$ (Meilleur, Pire ou aléatoire). Une série de tests correspond au canevas tel que nous l'avons défini où le locuteur est modélisé à l'aide du fichier d'apprentissage de notre choix. Ainsi, chaque série de tests se compose exactement des mêmes locuteurs et des mêmes fichiers de test, seul change

le fichier d'apprentissage utilisé. Une fois le fichier d'apprentissage sélectionné, un locuteur est représenté par le modèle élaboré à partir d'un seul fichier d'apprentissage dans toute la série.

Afin de mesurer l'influence du choix de ce fichier sur les performances du système, nous avons réalisé plusieurs séries de tests. Pour la première série, nous avons utilisé en apprentissage pour chaque locuteur son meilleur modèle (série *Min*), puis nous avons réalisé la série de tests en utilisant en apprentissage le pire modèle du locuteur (série *Max*). Cette démarche nous permet de mesurer l'écart maximum de performance que nous pouvons observer pour les mêmes tests et les mêmes locuteurs lorsque seuls les fichiers d'apprentissage changent. Pour établir la performance du système lorsque le fichier d'apprentissage n'est ni le meilleur ni le pire, nous avons conservé les fichiers qui étaient la référence lors de l'évaluation NIST 08.

2.3.3 Comparer les performances globales

La performance globale est mesurée à l'aide d'une courbe DET et d'un taux d'EER pour chacune des séries. Nous pouvons ainsi comparer les performances obtenues et rendre compte de la variation de performance due au fichier d'apprentissage puisque c'est l'unique élément qui change entre nos séries.
La cohorte de tests choisie est celle proposée par NIST où les 171 locuteurs de M-08 sont testés en apprentissage. Comme certains fichiers que nous avons sélectionnés comme meilleurs ou comme pires étaient à l'origine utilisés comme fichiers test, nous avons du supprimer certaines comparaisons. Cette cohorte se compose de 511 comparaisons cible et 2 856 comparaisons imposteur.

La variation relative, Vr, entre les séries pour chaque système et chaque base de donnée testée est définie par l'équation 1. Cette mesure nous permet de rendre compte de la variation due aux données d'apprentissage autour de la valeur moyenne habituellement mesurée.

$$Vr = \frac{EER_{Max} - EER_{Min}}{EER_{NIST}} \tag{1}$$

2.4 Résultats

La Figure 1 présente les résultats obtenus par ALIZE/SpkDet. Si la série *Min* obtient un EER à 4.1%, la série *Max* a un EER de 21.9%. La sélection correspondant à celle de NIST conduit à un EER de 12.1%. Dans ce cas, $Vr = 1.47$. Les Figures 2 et 3 présentent les résultats obtenus en utilisant Idento respectivement sans normalisation des scores et avec une ZTNorm. Sans normalisation, l'EER varie de 3.8% pour la série *Min* à 16.8% pour la série *Max*. La série où les modèles correspondent à ceux choisi par NIST a un EER de 9.2%. Dans ce cas, $Vr = 1.41$. Des écarts de performance s'observent donc également pour un système basé sur les i-vectors, dans des proportions semblables à celles obtenues pour ALIZE/SpkDet. Pour les deux systèmes utilisés, l'EER peut varier de 1.4 fois l'EER autour de la valeur moyenne mesurée, et ce pour les mêmes locuteurs et les mêmes fichiers de test. En étudiant les fichiers de chaque série *Min* et *Max* sélectionnés pour ALIZE/SpkDet et Idento, il est apparu que seul 30% des fichiers qui sont considérés comme les pires pour le système ALIZE/SpkDet le sont aussi pour Idento. De même, 30% des fichiers qui sont considérés comme les meilleurs pour le système ALIZE/SpkDet le sont pour Idento. Il semble donc qu'il existe une certaine variabilité entre les systèmes pour

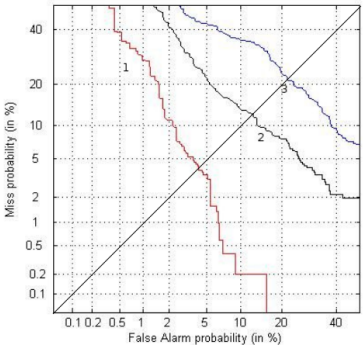

FIGURE 1 – Résultats pour ALIZE/SpkDet (*Min* (4.1%) : Rouge ; *Max* (21.9%) : Bleu ; *NIST* (12.1%) : Noir)

déterminer le meilleur et le pire enregistrement. Ceci peut être dû à la mesure de FR_{ij} qui est calculée sur peu de comparaisons cible et où une erreur a donc un impact important sur la mesure de la performance locale $FA_{ij} + FR_{ij}$. Une autre hypothèse serait qu'une partie de cette variation provienne d'une variation inhérente au système de vérification du locuteur et non au contenu des fichiers d'apprentissage. Il s'agit également de mesurer la variation propre au système afin d'en déterminer la fiabilité.

3 Variation propre au système

3.1 Base de données

Pour vérifier le comportement du système, nous avons construit, à partir des fichiers de BREF 120 (Lamel *et al.*, 1991), des fichiers de 2 minutes et 30 secondes de trames sélectionnées. Pour chaque locuteur, nous avons déterminé comme précédemment le meilleur et le pire modèle (pour plus de précision sur le protocole, lire (Kahn *et al.*, 2010)) que nous pouvions obtenir pour chacun des 111 locuteurs francophones qui composent la base de données. A partir de chacun des fichiers, nous avons construit deux modèles différents. Le premier modèle comporte toutes les trames impaires du fichier tandis que le second modèle comporte toutes les trames paires. Nous pouvons considérer que les informations utilisées pour construire les deux modèles sont équivalentes.

3.2 Mesure des écarts de performance

Comme précédemment, nous cherchons à déterminer quels sont les écarts maximum de performances que nous pouvons observer en fonction du modèle utilisé. Pour chaque locuteur, nous avons déterminé quel est le meilleur modèle entre celui construit avec les trames paires et

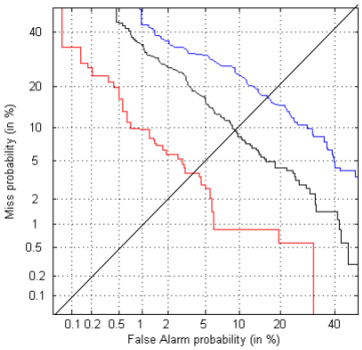

FIGURE 2 – Résultats pour Idento en NoNorm (*Min* (3.8%) : Rouge ; *Max* (16.8%) : Bleu ; *NIST* (9.2%) : Noir)

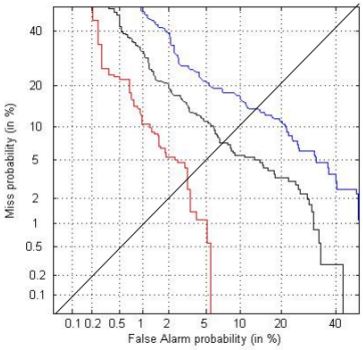

FIGURE 3 – Résultats pour Idento en ZTNorm (*Min* (3.1%) : Rouge ; *Max* (13.8%) : Bleu ; *NIST* (7.3%) : Noir)

celui construit avec les trames impaires. Nous avons effectué la même série de comparaisons en prenant les meilleurs fichiers puis les pires fichiers. Les fichiers tests des comparaisons sont ceux utilisés dans (Kahn *et al.*, 2010). ILs sont exactemlent les mêmes pour tous les locuteurs. Cette expérience a été menée uniquement avec le système ALIZE/SPkDet précédemment décrit.

3.3 Résultats

Le Tableau 1 présente les EER dans chacune des conditions.Pour les hommes, nous obtenons un $EER = 1.0\%$ pour les fichiers d'apprentissage de 2min30 de la série *Min*. En séparant en trames paires et impaires de ces fichiers d'apprentissage, les meilleurs modèles obtiennent un EER de 2.1% tandis que les pires modèles obtiennent un EER de 3.2%. Lorsque les modèles sont construits en prenant une trame sur deux des fichiers de la série *Max* ($EER = 5.8$ lorsque l'intégralité des fichiers est utilisée), les meilleurs modèles obtiennent un EER de 2.7% tandis que les pires obtiennent un EER de 3.2%.

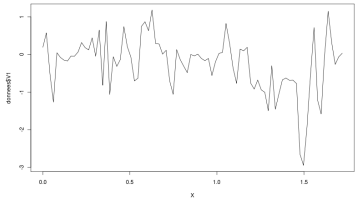

FIGURE 4 – Valeurs du LLR pour une comparaison cible

Etant donné qu'en prenant une trame sur deux, la quantité d'information est divisée par deux, il n'est pas surprenant que l'EER de 1% sur la série *Min* soit compris entre 2.1% et 3.2% lorsqu'une trame sur deux seulement est sélectionnée : Nous avons moins de trames donc moins de précision dans l'évaluation du score. Les résultats concernant la série *Max* sont plus difficiles à interpréter. En prenant une trame sur deux, nous avons un EER qui se situe entre 1.7% et 3.2% et lorsque nous prenons comme fichier d'apprentissage l'intégralité des trames, l'EER vaut 5.8%. Il apparaît que la quantité d'information utile au système n'est pas corrélée, dans ce cas, aux nombre de trames disponibles. Le même type de résultats s'observe pour les femmes.

Nous observons un écart de performance de près d'un point d'EER alors que les modèles ont été construits à partir de jeux de données statistiquement équivalents. La part de variation attribuée au système reste limitée. Afin de comprendre comment, dans le cas de la série *Max*, une quantité d'information divisée par deux peut donner lieu à de meilleurs résultats en terme d'EER, nous avons observé les scores trame à trame obtenus pour un fichier test cible (figure 4). Les scores LLR montrent des variations très brutales d'une trame à l'autre et cette instabilité peut être une piste à explorer. Elle met en évidence l'importance d'étudier le lien entre fichier d'apprentissage et fichier de test.

Genre	Catégorie d'origine des fichiers	Catégorie pour une trame sur deux	EER
Hommes	*Min* ($EER = 1.0\%$)	*Min*	2.1%
		Max	3.2%
	Max ($EER = 5.8\%$)	*Min*	2.7%
		Max	3.2%
Femmes	*Min* ($EER = 0.9\%$)	*Min*	1.2%
		Max	2.7%
	Max ($EER = 6.0\%$)	*Min*	1.2%
		Max	2.3%

TABLE 1 – EER obtenus en prenant une trame sur deux des fichiers *Min* et *Max* de BREF 2min30svs30s

4 Conclusions et perspectives

Nous avons montré que le choix du modèle d'apprentissage a des conséquences très importantes sur les performances d'un système de vérification du locuteur. Ceci est indépendant du type de locuteur puisque ce sont exactement les mêmes locuteurs qui sont comparés dans les deux séries. La série correspondant à NIST montre que si les fichiers d'apprentissage sont tirés aléatoirement, la performance du système se situe entre les deux série *Min* et *Max* et rend compte d'une performance moyenne en lissant les écarts importants dus au choix du fichiers d'apprentissage.

Par ailleurs, nous avons également étudié les écarts de performance due au système. Ces écarts sont largement plus faibles que les écart de performance observés lorsque l'on modifie le signal

d'apprentissage de chaque locuteur mais montrent que l'approche UBM-GMM présentent une certaine instabilité.

Il est par ailleurs étonnant qu'en sélectionnant une trame sur deux, les performances de la série *Max* soient si proches de celles de la série *Min*. Une analyse de la composition trame à trame des jeux de données utilisés pour l'apprentissage et le test reste nécessaire pour mieux comprendre le comportement du système, qui pourrait être du à la présence de quelques données très spécifiques comme l'illustre la Figure 4 en présentant les valeurs de LLR par trames pour une comparaison cible.

Ces séries d'expériences montrent bien la nécessité de prendre en compte la variation de performance due aux données et au système lui-même dans l'évaluation des performance de systèmes de RAL afin de les rendre fiables.

Références

Bonastre, J.-F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G. et Evans, N. (2008). ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. *In ISCA-IEEE Speaker Odyssey*, Stellenbosch.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P. et Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *In International Conference on Speech Communication and Technology (Interspeech)*, pages 1559–1562, Brighton.

Greenberg, C., Martin, A., Barr, B. et Doddington, G. (2011). Report on performance results in the nist 2010 speaker recognition evaluation. *In International Conference of the International Speech Communication Association (Interspeech)*, pages 261–264, Florence.

Kahn, J., Audibert, N., Rossato, S. et Bonastre, J.-F. (2010). Intra-speaker variability effects on speaker verification system performance. *In ISCA-IEEE Speaker Odyssey*, Brno.

Kenny, P., Boulianne, G., Ouellet, P. et Dumouchel, P. (2005). Factor analysis simplified. *In International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP)*, pages 637–640, Philadelphie.

Lamel, L., Gauvain, J. et M., E. (1991). BREF, a large vocabulary spoken corpus for French. *In European Conference on Speech Communication and Technology (Eurospeech)*, pages 505–508, Gènes.

Martin, A., Doddington, G., Kamm, T., Ordowski, M. et Przybocki, M. A. (1997). The det curve in assessment of detection task performance. *In European Conference on Speech Communication and Technology (Eurospeech)*, pages 1895–1898, Rhodes.

Matrouf, D., Bonastre, J., Fredouille, C., Larcher, A., Mezaache, S., McLarren, M. et Huenupan, F. (2008). GMM-SVM system description : NIST SRE. Montréal.

Reynolds, D., Quatieri, T. F. et B, D. R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:p19–41.

Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E. et Stolcke, A. (2011). The SRI NIST 2010 speaker recognition evaluation system. *In International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5292–5295, Brno.