Exploitation d'une marge de tolérance de classification pour améliorer l'apprentissage de modèles acoustiques de classes en reconnaissance de la parole

Denis Jouvet, Arseniy Gorin, Nicolas Vinuesa
Speech Group, INRIA – LORIA,
615 rue du Jardin Botanique, 54602 Villers les Nancy
{denis.jouvet,arseniy.gorin}@loria.fr

RESUME

Ce papier présente la prise en compte d'une marge de tolérance lors la classification des données d'apprentissage pour la fabrication de modèles acoustiques de classes pour la transcription automatique de la parole. En effet, bien que la classification automatique des données permette d'aller au-delà de la traditionnelle partition hommes/femmes, le nombre de classes utilisables est généralement limité par la fiabilité des modèles acoustiques associés aux classes, qui malheureusement va en diminuant avec le nombre de classes. Les expériences présentées montrent que la prise en compte d'une marge de tolérance lors de la classification des données d'apprentissage permet d'accroitre la quantité des données associées à chaque classe, et donc la fiabilité des modèles acoustiques associés aux classes. Les évaluations menées sur les données de la campagne ESTER2 ont montré la possibilité de fabriquer ainsi des modèles de classes aboutissant à de meilleures performances que l'utilisation des modèles habituels spécialisés hommes/femmes.

ABSTRACT

Exploitation of a classification tolerance margin for improving the estimation of class-based acoustic models for speech recognition

This paper presents the introduction of a classification tolerance margin in the classification of the training data for building class-based acoustic models for automatic speech transcription. Indeed, although automatic classification of speech data makes it possible to go beyond the traditional male / female partition, the number of usable classes is actually limited by the reliability of the associated acoustic models which, unfortunately, decreases when the number of classes increases. The reported experiments show that using a tolerance margin in the classification process increases the amount of training data associated to each class, and consequently increases the reliability of the acoustic models of the classes. The performance evaluation carried on the ESTER2 data have shown that it is possible with the proposed approach to build class-based acoustic models that lead to better speech recognition performance than with the usual gender-based acoustic models.

MOTS-CLES :Reconnaissance de la parole, classification automatique, modèles acoustiques de classes, marge de tolérance de classification,
KEYWORDS :Speech recognition, automatic classification, class-based acoustic models, classification tolerance margin

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 763–770,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les modèles acoustiques sont l'un des constituants fondamentaux des systèmes de reconnaissance de la parole. Ils modélisent la réalisation acoustique des sons (phonèmes) de la langue, et doivent tenir compte des multiples sources de variabilité qui viennent affecter le signal de parole et qui impactent sur les performances de la reconnaissance automatique de la parole (Benzeghiba *et al.*, 2007). Comme les performances de reconnaissance sont d'autant meilleures que les modèles acoustiques utilisés sont en adéquation avec les conditions acoustiques du signal de parole à reconnaître, les systèmes de transcription automatique de la parole fonctionnent typiquement en plusieurs passes. La première est consacrée à la découpe du signal en segments homogènes, puis à l'identification des caractéristiques de chaque segment. La reconnaissance est ensuite effectuée en utilisant des modèles acoustiques correspondant à la classe du segment à reconnaître, en général des modèles acoustiques dépendant du sexe du locuteur.

L'augmentation du nombre de composantes gaussiennes des densités acoustiques améliore les performances de reconnaissance grâce à une meilleure modélisation des variantes des réalisations acoustiques qui résultent des multiples sources de variabilité affectant le signal de parole. Toutefois, la forte dispersion des réalisations dues aux variabilités du signal limite la précision des modèles acoustiques. L'emploi d'une modélisation multiple est une voie pour pallier ce problème. Ainsi, au lieu d'un seul jeu de modèles acoustiques, il est possible de fabriquer plusieurs jeux de modèles acoustiques, chaque jeu correspondant à un sous-ensemble de variabilités. Le décodage peut alors être effectué avec le modèle acoustique adéquat ou bien plusieurs décodages peuvent être faits en parallèle, et les résultats combinés par une approche de type ROVER (Fiscus, 1997).

La modélisation acoustique dépendante du locuteur est la modélisation acoustique la plus précise. Différentes techniques existent pour adapter un modèle générique aux données d'un locuteur comme les voix propres (Kuhn *et al.*, 1998), l'interpolation de modèles de classes (Gales, 1998) ou de locuteurs de référence (Teng *et al.*, 2007). Une autre orientation consiste à exploiter le principe des réseaux bayesiens dynamiques (Zweig, 1998) pour rendre la modélisation acoustique dépendante d'une variable auxiliaire représentant les variabilités considérées, comme le pitch (Stephenson *et al.*, 2004), des facteurs cachés (Korkmazsky *et al.*, 2004) ou encore la classe du locuteur (Cloarec & Jouvet, 2008).

L'estimation des modèles acoustiques correspondant à un ensemble réduit de variabilités (ex. classe restreinte de locuteurs) peut conduire à des modèles non fiables lorsqu'il n'y a pas assez de données correspondant à cette classe. Or c'est malheureusement fréquemment le cas lorsque le nombre de classes augmente. Ce papier présente une approche visant à accroître la quantité de données utilisée pour l'apprentissage des modèles acoustiques de chaque classe. L'approche repose sur la prise en compte de l'incertitude de classification pour les données qui se trouvent à la frontière entre plusieurs classes, et s'inspire du traitement de l'incertitude sur les frontières de segmentation pour l'estimation de modèles dépendant de la vitesse d'articulation (Jouvet *et al.*, 2011).

L'organisation du papier est la suivante. Après un rappel sur l'utilisation de modèles acoustiques de classes et la classification automatique des données d'apprentissage, la section 2 présente l'introduction d'une marge de tolérance dans la classification automatique. La section 3 présente les expériences menées en transcription automatique de la parole sur les données de la campagne d'évaluation ESTER2 (Galliano *et al.*, 2009) et commente les résultats obtenus. Finalement, une conclusion termine le papier.

2 Introduction d'une marge de tolérance dans la classification automatique

Avant d'introduire la prise en compte d'une marge de tolérance dans la classification automatique des données d'apprentissage, cette section rappelle l'utilisation de modèles de classes en reconnaissance de la parole, ainsi qu'une approche de classification automatique permettant la fabrication d'un nombre arbitraire de classes de données.

2.1 Utilisation de modèles de classes en reconnaissance de la parole

Au lieu de la découpe traditionnelle en 2 classes (hommes vs. femmes), la classification automatique permet de considérer un nombre arbitraire de classes C_k, $k = 1, \dots, K$. L'ensemble des GMMs$\{\Phi_k, k = 1, \dots, K\}$ correspondant aux différentes classes permet de classifier n'importe quelle donnée (segment de parole) X:

$$X \in C_k \Leftrightarrow P(X|\Phi_k) \geq P(X|\Phi_l) \quad \forall l \tag{1}$$

La donnée est affectée à la classe C_k du GMM Φ_k qui conduit à la plus grande vraisemblance. Les modèles acoustiques Λ_k (modèles acoustiques des phonèmes) associés à cette classe C_k sont alors utilisés pour décoder ce signal de parole X:

$$\widehat{W} = \text{argmax}_W \, P(X|W, \Lambda_k).P(W) \tag{2}$$

Les modèles acoustiques Λ_k des phonèmes sont typiquement obtenus par adaptation d'un modèle générique sur les données d'apprentissage (données d'adaptation) de la classe C_k.

2.2 Classification automatique des données d'apprentissage

Le problème de la classification automatique est la détermination simultanée des classes de données et des GMMs associés. L'approche choisie ici repose sur une détermination incrémentale des classes et des GMMs associés. A chaque étape du processus le nombre de classes est multiplié par deux. On commence par 1 classe correspondant à l'ensemble des données, puis on fabrique 2 classes, puis 4, 8, 16, … classes. Si le nombre désiré de classes n'est pas une puissance de 2, on peut limiter l'augmentation du nombre de classes en ne considérant que les plus grosses ou les plus dispersées.

La figure 1 représente les différentes étapes de traitement pour la classification automatique, et commence par 1 seule classe et le GMM associé correspondant à l'ensemble des données à classifier. Ensuite, à chaque étape du processus, le nombre de classes est multiplié par 2. Pour cela, chaque GMM Φ_k est dupliqué, et les valeurs des moyennes des gaussiennes sont légèrement modifiées de manière aléatoire pour obtenir deux GMMs Φ_{k1} et Φ_{k2}. Ces GMMs servent alors à classifier les données conformément à l'équation (1), i.e. chaque donnée est affectée à la classe du GMM qui maximise la

vraisemblance. Pour chacune des classes obtenues, un nouveau GMM est appris. Les étapes de classification et d'apprentissage des GMMs sont répétées autant que nécessaire, jusqu'à convergence (critère de vraisemblance) ou nombre maximal d'itérations.

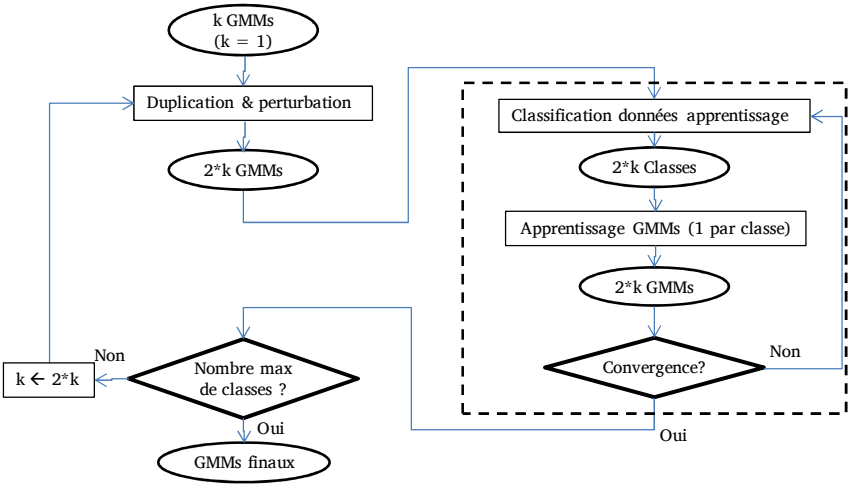

FIGURE 1 – Etapes de la classification automatique.

2.3 Exploitation d'une marge de tolérance lors de la classification

L'idée sous-jacente consiste à exploiter de manière optimale les données qui sont à la frontière des classes. En effet les données à la frontière de deux classes peuvent être affectées à l'une ou l'autre des classes, voire aux deux classes. Cela revient à considérer qu'il y a une incertitude sur la frontière. L'introduction d'une marge de tolérance δ dans l'équation (1) permet de gérer une telle incertitude, et d'affecter à plusieurs classes les données qui se trouvent à la frontière des classes :

$$X \in C_k \Leftrightarrow \frac{1}{T} Log\ P(X|\Phi_k) \geq \max_l \frac{1}{T} Log\ P(X|\Phi_l) - \delta \qquad (3)$$

Lorsque la marge de tolérance δ vaut 0, l'équation (3) conduit à la même classification que l'équation (1). Lorsque l'on augmente la marge de tolérance δ, de plus en plus de données se trouvent affectées à plusieurs classes, ce qui augmente, en moyenne, la quantité des données associées à chaque classe.

3 Etude expérimentale

Les expériences de reconnaissance automatique de la parole avec des modèles de classes ont été menées sur les données de la campagne d'évaluation ESTER2 (Galliano *et al.*, 2009).

3.1 Contexte expérimental

Les données d'apprentissage du corpus ESTER2, environ 190 heures, ont servi pour l'estimation des GMMs de classification, ainsi que pour l'estimation des modèles acoustiques des phonèmes associés à chaque classe. Les évaluations ont été menées sur les données françaises du corpus de développement, et correspondent à environ 4h30 de signal audio et 36800 mots.

Les expériences ont été menées avec le système de reconnaissance Sphinx (2011). La transcription de la parole est effectuée en 2 passes : une première passe pour la segmentation du signal et l'identification des caractéristiques des segments, puis une seconde passe pour effectuer le décodage avec le modèle correspondant aux caractéristiques estimées (qualité studio vs téléphone, et homme vs femme pour l'approche classique, ou classe du locuteur pour l'approche proposée ici). L'identification de la classe du locuteur repose sur les GMMs appris, et l'application de l'équation (1), i.e. identification de la classe correspondant au maximum de vraisemblance.

Les modèles acoustiques des phonèmes sont composés de 4 500 senones (états/densités partagés) et chaque densité a 64 composantes gaussiennes. Les modèles des phonèmes dépendant du contexte sont d'abord appris pour les conditions studio et téléphone, puis adaptés au type du locuteur (homme vs femme) ou aux classes de locuteurs, selon les expériences, en combinant successivement une adaptation MLLR (une matrice de régression par phonème) puis une adaptation MAP des paramètres. L'adaptation aux classes se fait à partir des données associées à chaque classe en fonction de la marge de tolérance choisie, d'après l'équation (3).

Le lexique de prononciations comprend environ 64 000 mots, et un modèle de langage trigramme est également utilisé.

3.2 Analyse de quelques classes

Lorsque la marge de tolérance de classification augmente, plus de données sont affectées à chaque classe, i.e. les classes se recouvrent de plus en plus.

C'est ce qui est représenté sur la figure 2 pour la classification à 32 classes. L'axe vertical représente sur une échelle logarithmique la durée totale (en secondes) du signal de l'ensemble des segments affectés à chaque classe (axe horizontal) en fonction de la marge de tolérance (0,0, 0,5, 1,0 et 1,5). Les classes ont été rangées par ordre décroissant de la quantité des données affectées à la classe pour la classification traditionnelle (équivalent à une marge de 0,0).

On voit que la quantité des données associées à chaque classe est très variable, et va d'une vingtaine de minutes à plus de 13 heures. A quelques exceptions près, l'introduction d'une marge de tolérance dans la classification augmente de manière significative la quantité de données associées à chaque classe. L'idée sous-jacente de l'approche proposée, est que, pour une marge de tolérance raisonnable, les données complémentaires associées à chaque classe sont similaires à celles du noyau de la classe, et donc ne perturberont pas l'apprentissage, mais au contraire seront bénéfiques, car l'ensemble d'adaptation plus grand devrait rendre l'estimation des paramètres plus pertinente.

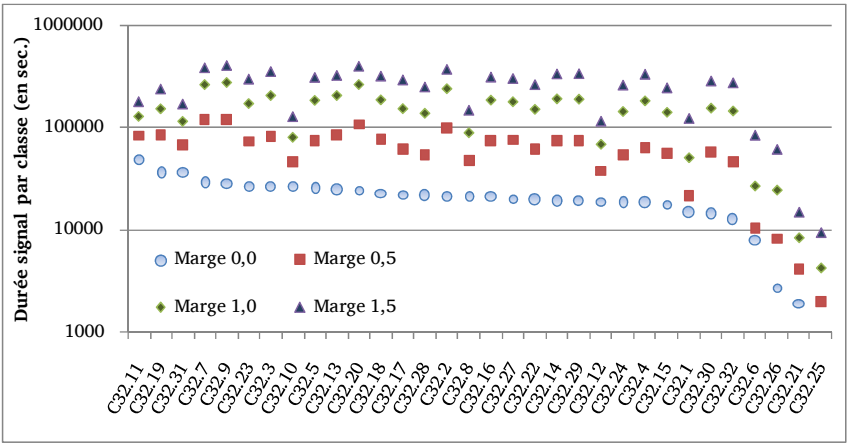

FIGURE 2 – Impact de la marge sur la taille des sous-ensembles de données associées à chaque classe (modèle 32 classes).

3.3 Evaluation des performances de reconnaissance

Le tableau suivant indique les taux d'erreur de reconnaissance sur les données françaises de l'ensemble de développement d'ESTER2 en fonction du nombre de classes du modèle, et de la marge de tolérance utilisée pour la classification des données d'apprentissage.

En l'absence de classification, i.e. en utilisant uniquement un modèle générique pour les données de qualité studio, et un autre pour les données de qualité téléphone, le taux d'erreur obtenu est de 25,97%.Lorsque l'on utilise en plus une classification homme/femme, le taux d'erreur descend à 24,91%.

Marge de tolerance	0,0 (aucune)	0,5	1,0	1,5	2,0	2,5
2 classes	**24,97**	25,16	25,55	25,37	25,56	25,66
4 classes	24,77	**24,69**	24,99	25,21	25,09	25,29
8 classes	24,88	**24,66**	**24,71**	25,01	24,95	25,29
16 classes	25,15	25,14	**24,54**	**24,52**	24,90	24,98
32 classes	25,97	24,82	**24,32**	24,59	24,51	25,04

TABLE 1 – Taux d'erreur (%) en fonction du nombre de classes et de la marge de tolérance.

Les modèles acoustiques des phonèmes pour chaque classe ont été adaptés en applicant successivement une adaptation MLLR (une matrice de régression par phonème) puis une adaptation MAP des paramètres. Cette combinaison donne en effet de meilleures performances que l'adaptation MAP seule, en particulier lorsque le nombre de classes est élevé.

La deuxième colonne du tableau (marge 0,0) montre que l'approche de classification standard est rapidement limitée par le manque de fiabilité des modèles estimés dès que le nombre de classes est important. Les performances se dégradent à partir de 8 classes.

Les résultats montrent également qu'en introduisant une légère marge de tolérance dans la classification des données d'apprentissage, on améliore la qualité des modèles acoustiques des classes, et donc globalement les performances de reconnaissance. Par contre, lorsque la marge de tolérance est trop grande (par exemple 2,0 et 2,5), on introduit dans la classe des données trop disparates qui viennent pénaliser la précision des modèles.

Globalement, les résultats montrent qu'en introduisant une marge de tolérance raisonnable lors de la classification des données du corpus d'apprentissage on peut utiliser de manière efficace un nombre important de classes de données, et obtenir des taux d'erreurs significativement meilleurs qu'avec la traditionnelle classification homme/femme.

4 Conclusion

Ce papier a analysé l'apprentissage de modèles acoustiques de classes de données dans le cadre de la transcription de parole. La classification automatique des données permet de fabriquer un nombre arbitraire de classes. Cependant lorsque le nombre de classes augmente, la quantité de données affectées à chacune des classes diminue, ce qui fait que le gain en précision du modèle est pénalisé par le manque de fiabilité des estimations (manque de données). En conséquence le nombre de classes utilisables est rapidement limité.

Ce papier présente l'utilisation d'une marge de tolérance de classification pour pallier ce problème d'estimation. Elle consiste à introduire une tolérance sur la frontière des classes, ce qui permet d'accroître la quantité des données affectées à chaque classe. Les résultats expérimentaux ont montré que la prise en compte d'une petite marge de tolérance permet d'améliorer la pertinence des modèles appris, et d'obtenir des taux de reconnaissance significativement meilleurs que ceux résultant de la traditionnelle classification homme/femme.

Dans ces expériences une classification automatique simple des données d'apprentissage a été exploitée. On peut supposer raisonnablement qu'une classification plus élaborée, telle que celles proposées dans (Krstulovic *et al.*, 2007) qui focalisent sur certaines classes phonétiques ou exploitent des modèles acoustiques des phonèmes ou dans (Beaufays *et al.*, 2010) qui cherche à maximiser la dissimilarité entre classes, pourrait conduire à des performances encore meilleures.

Références

BEAUFAYS, F., VANHOUCKE, V. et STROPE, B. (2010). Unsupervised discovery and training of maximally dissimilar cluster models. *In Proc. INTERSPEECH'2010*, Makuhari, Japon, sept. 2010.

BENZEGHIBA, M., DE MORI, R.,DEROO, O., DUPONT, S.,ERBES, T., JOUVET, D.,FISSORE, L.,LAFACE, P.,MERTINS, A., RIS, C., ROSE, R., TYAGI, V. etWELLEKENS, C. (2007). Automatic speech recognition and variability: a review. *Speech Communication*, vol. 49, pp. 763-786, 2007.

CLOAREC, G. et JOUVET, D. (2008). Modeling inter-speaker variability in speech recognition. *In Proc. ICASSP'2008*, Las-Vegas, USA, mars 2008.

FISCUS,J.G. (1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). *In Proc. ASRU'97*, Santa Barbara, CA, USA, 1997, pp. 347-354.

GALES, M.J.F. (1998). Cluster adaptive training for speech recognition. *In Proc. ICSLP'98*, Sydney, Australie, 1998, pp. 1783-1786.

GALLIANO, S., GRAVIER, G., etCHAUBARD, L. (2009). The ESTER 2 evaluation campaign for rich transcription of French broadcasts. *In Proc. INTERSPEECH'2009*, Brighton, UK, pp. 2583-2586, sept. 2009.

JOUVET, D., FOHR, D. et ILLINA, I. (2011). About handling boundary uncertainty in a speaking rate dependent modeling approach. *In Proc. INTERSPEECH'2011*, Florence, Italie, août 2011.

KORKMAZSKY, F., DEVIREN, M., FOHR, D. et ILLINA, I. (2004). Hidden factor dynamic Bayesian networks for speech recognition. *In Proc. ICSLP'2004*, Jeju Island, Corée, 2004.

KRSTULOVIC, S., BIMBOT, F., BOËFFARD, O., CHARLET, D., FOHR, D. et MELLA, O. (2007). Selecting representative speakers for a speech database on the basis of heterogeneous similarity criteria. *Speaker Classification II*, Christian Müller (réd), Lecture Notes in Computer Science, 4441, Springer Berlin, 2007, pp. 276-292.

KUHN, R., NGUYEN, P., JUNQUA, J.-C.,GOLDWASSER, L.,NIEDZIELSKI, N.,FINCKE, S., FIELD, K.etCONTOLINI, M. (1998). Eigenvoices for speaker adaptation. *In Proc. ICSLP'98*, Sydney, Australie, 1998, pp. 1771-1774.

SPHINX.[Online] Available: http://cmusphinx.sourceforge.net [consulté en 2011].

STEPHENSON, T.A., MAGIMAI-DOSS, M. et BOURLARD, H. (2004). Speech recognition with auxiliary information. *IEEE Trans. on Speech and Audio Processing*, 2004, vol. 12, pp. 189-203.

TENG, W., GRAVIER, G., BIMBOT, F. et SOUFFLET, F. (2007). Rapid speaker adaptation by reference model interpolation. *In Proc. INTERSPEECH'2007*, Anvers, Belgique, 2007, pp. 258-251.

ZWEIG, G. (1998). *Speech recognition with Dynamic Bayesian Networks*. Ph. D. Dissertation, Univ. California, Berkeley, 1998.