TECHNIQUES DE COMPENSATION POUR LA RECONNAISSANCE DE LA PAROLE BRUITÉE

Driss MATROUF, Jean-Luc GAUVAIN

LIMSI-CNRS, BP 133, F-91403 Orsay, France
Tél: 69 85 80 67 - Fax: 69 85 80 80 - e-mail: {driss,gauvain}@limsi.fr

ABSTRACT

The performance of speech recognizers degrade substantially when there is a mismatch between the training and testing conditions. The goal of noise compensation is to minimize the effects of the mismatch, so as to bring the recognition performance as close as possible to the obtained under matched conditions. Different approaches to achieve robustness have been studied. These approaches may be split into two groups. The first class of approaches make use of a channel model for additive and convolutional noises. Techniques in this category include spectral substraction[3], cepstral mean substraction[10], noise masking, the CDCN algorithm[9], speech and noise decomposition[2], and parallel model combination[4][7][8]. The second class of approaches makes no assumptions about the underlying noise, and uses some optimality criterion (generally Maximum Likelihood (ML)) to find a transformation that is applied either to the signal or to the acoustic models used by the recognizer. Techniques in this category include ML linear regression[11][6] and ML stochastic matching[5]. In this paper we discuss the properties of above mentioned techniques. The results of our analyses have led to the implementation of noise compensation in the LIMSI large vocabulary CSR system. Evaluation results on the Nov95 ARPA NAB CSR test data are given both with and without noise compensation.

INTRODUCTION

La présence de bruit engendre une dégradation significative des performances des systèmes de reconnaissance de la parole, en particulier lorsque les conditions d'apprentissage et de test sont différentes. Les techniques de compensation ont pour but de réduire les effets de cette différence et d'obtenir un taux d'erreur comparable à celui obtenu lorsque les conditions d'apprentissage et de test sont identiques. Les techniques utilisées pour appréhender le problème du bruit peuvent se répartir en deux classes.

La première classe repose sur un modèle du canal de transmission avec deux types de bruits: additif et convolutif, i.e. le signal observé est $y(t) = (s(t) + n(t)) * h(t)$, où $n(t)$ et $h(t)$ désignent respectivement les bruits additif et convolutif. Au sein de cette classe, on peut distinguer deux approches différentes. La première consiste à estimer $s(t)$ à partir du signal bruité $y(t)$ et de statistiques sur $h(t)$ et $n(t)$. Parmi les techniques suivant cette approche, on peut citer la soustraction spectrale[3], la soustraction du cepstre moyen[10], le masquage de bruit et l'algorithme CDCN (Codeword-Dependent Cepstral Normalization)[9]. La deuxième approche consiste à adapter les modèles retenus pour le signal propre $s(t)$ afin d'obtenir des modèles représentatifs du signal bruité $y(t)$. On peut citer la décomposition de modèles[2], et la combinaison parallèle des modèles[4][7][8].

La deuxième classe ne suppose aucun modèle *à priori* du bruit, mais utilise un critère d'optimalité généralement basé sur le maximum de vraissemblance pour estimer une transformation à appliquer sur le signal ou sur les modèles acoustiques du système de reconnaissance. Dans cette classe, on peut citer l'adaptation par régression linéaire (MLLR: Maximum Likelihood Linear Regression)[6]et la compensation stochastique[5].

L'objectif de cette étude est de faire une analyse permettant de choisir les techniques les plus pertinentes à incorporer dans le système de reconnaissance du LIMSI. Cette analyse est basée soit sur des résultats expérimentaux, soit sur une étude théorique qui met en évidence les limites des techniques non retenues. Seuls les résultats des tests validant les techniques retenues sont présentés. Des tests comparatifs avec les autres techniques ne sont pas présentés car d'une part les évaluations des différentes techniques ont été généralement réalisées sur des corpus différents et d'autre part, la place impartie ne permet pas d'exposer l'ensemble des résultats avec les conditions expérimentales correspondantes.

TECHNIQUES BASÉES SUR UN MODÈLE *À PRIORI*

Dans cette section nous analysons les techniques fondées sur un modèle *à priori*. Les modèles de bruits généralement utilisés sont: $y(t) = s(t) + n(t)$, $y(t) = s(t) * h(t)$ ou $y(t) = (s(t) + n(t)) * h(t)$, avec $n(t)$ et $h(t)$ désignent respectivement les bruits additif et convolutif et $y(t)$ le signal observé.

Soustraction spectrale

Cette technique consiste à effectuer une décomposition spectrale de chaque trame du signal bruité. Puis chaque canal du spectre est atténué selon que le niveau mesuré localement dans le canal dépasse plus ou moins l'estimation du bruit. La densité spectrale du bruit est estimée dans les périodes de silence[3]. L'utilisation d'un seuil pour éviter les valeurs négatives causées par la variance du bruit et d'un facteur de surestimation pour augmenter le rapport signal sur bruit[12] introduisent des formes spectrales complétement inconnues du système de reconnaissance de la parole qui se base principalement sur une classification des formes spectrales à court terme du signal de parole. En plus, la non prise en compte de l'effet du bruit sur les variances des modèles limite considérablement l'utilisation de cette technique dans le cadre de la reconnaissance automatique de la parole.

Soustraction du cepstre moyen

Cette technique consiste à soustraire le cepstre moyen à long terme. En général on soustrait le cepstre moyen sur la phrase. Cette idée est couramment utilisée pour éliminer les distortions liées au changement du canal d'enregistrement[10]. Après soustraction cepstrale on aboutit à des paramètres cepstraux indépendants du matériel d'enregistrement utilisé. Il est clair que la moyenne estimée dépend de la proportion du silence (bruit) dans la phrase. Cette dépendance est indésirable. Pour résoudre ce problème, un mélange de deux gaussiennes a été utilisé: une pour les trames de parole et une autre pour les trames du silence (bruit)[13]. Cette classification se fait en utilisant l'algorithme EM avec comme paramètre de classification l'énergie[1]. Ainsi la soustraction du cepstre moyen se fait comme suit:
$\hat{y}^c(t) = y^c(t) - [\gamma\bar{y}_p^c + (1-\gamma)\bar{y}_b^c]$, $y^c(t)$ est la représentation cepstrale d'une trame à l'instant t. γ désignant la probabilité que $y^c(t)$ soit de la parole. $\bar{y}_p^c$ et $\bar{y}_b^c$ désignent respectivement la moyenne des vecteurs cepstraux de parole et du silence.

Principe du masquage et la décomposition parole/bruit

L'idée du masquage du bruit[2], appelé aussi l'approximation du MAX peut se résumer par la formule:
$Y_t^l = log(X_t + N_t) \simeq max[X_t^l, N_t^l]$, Y_t^l représente le niveau du logarithme de l'énergie dans un canal d'un banc de filtres de la parole bruitée à l'instant t, X_t^l et N_t^l représentent respectivement le niveau du logarithme de l'énergie de la parole et du bruit dans un canal donné. C'est cette idée qui est à l'origine de toutes les méthodes de compensation actuellement utilisées dans les systèmes de reconnaissance. En effet, ce même principe a été étendu pour donner lieu à la technique de la décomposition de la parole et du bruit (SND: Speech and Noise Decomposition)[2]. Il s'agit d'une méthode optimale pour reconnaître la parole et le bruit simultanément. L'approximation du MAX faite par Varga[2] donne une bonne estimation de la densité de probabilité de la parole bruitée. Cependant, elle suppose que le bruit agit indépendamment d'un canal à l'autre dans le domaine log-spectral. En plus cette technique impose l'utilisation des matrices de covariance pleines pour modéliser la corrélation entre les composantes.

Combinaison parallèle des modèles

La combinaison parallèle des modèles (PMC) [7][8] a été introduite pour surmonter les limitations de la SND. Elle est cependant directement inspirée de cette dernière. Elle procède à toutes les opérations inverses pour revenir du domaine cepstral au domaine spectral (changement des densités de probabilités suivant l'opération) et inversement.

L'approximation log-normale[7][8] est souvent utilisée pour revenir du domaine spectral au domaine cepstral. Elle suppose que la somme de deux variables aléatoires log-normalement distribuées est elle-même log-normalement distribuée. Il est important de noter que l'approximation log-normale devient inacceptable dans le cas de grandes variances, c'est la raison pour laquelle il faut répartir l'espace acoustique en plusieurs sous-espaces de petites variances.

La différence principale entre la SND et la PMC est le type d'approximation: la PMC utilise l'approximation log-normale ou procède à une intégration numérique, par contre la SND utilise l'approximation du MAX définie précédemment. Une autre différence est que la SND opère dans le domaine log-spectral alors que la PMC opère dans le domaine cepstral. Il est important de noter que la SND augmente considérablement le nombre de paramètres dans le système. La difficulté des calculs et le nombre d'approximations dans le cadre de la PMC augmentent rapidement avec la complexité des formules utilisées pour les paramètres cepstraux différentiels, ce qui représente un grand inconvénient de l'approche.

Compensation du bruit convolutif dans le cadre de la PMC

L'idée retenue est la soustraction du cepstre moyen. Après avoir appliqué la PMC, on obtient des modèles correspondant à la parole bruitée. Pour obtenir des modèles pour la parole bruitée avec soustraction du cepstre moyen, il suffit de soustraire le cepstre moyen de toute la parole bruitée à toutes les moyennes des gaussiennes composant les modèles. Mais it à priori on ne dispose pas des données bruitées pour pouvoir calculer cette moyenne. Pour résoudre ce problème, on utilise une somme de gaussiennes représentant toute la parole propre; on fait la composition de celle-ci avec le modèle de bruit dont on dispose pour obtenir un modèle composé de plusieurs gaussiennes représentant toute la parole bruitée. Grâce à ce modèle on peut estimer d'une manière satisfaisante le cepstre moyen de la parole bruitée qu'on peut ensuite soustraire à toutes les moyennes des gaussiennes. Il est important de noter que l'utilisation d'un mélange de gaussiennes pour représenter toute la parole propre est nécessaire car une seule gaussienne aurait une très grande variance et l'approximation log-normale de la PMC ne sera plus valable dans ce cas. Des tests effectués au LIMSI montrent que cette technique donne des résultats tout-à-fait satisfaisants.

Utilisation de la composition directement sur les données

Pour résoudre certains problèmes liés à la PMC, on utilise une nouvelle technique développée au LIMSI[14] utilisant la composition directement sur les données. Elle consiste à stocker pour chaque gaussienne d'un état donné, d'un modèle donné, les vecteurs cepstraux qui ont permis son estimation. La composition consiste dans ce cas à faire la somme trame à trame dans le milieu spectral des trames stockées pour la gaussienne en question et des trames représentant le bruit. Les trames représentant le bruit proviennent généralement d'une analyse à court terme du signal recueilli dans les périodes de silence au moment du test. On applique les opérations nécessaires pour revenir au domaine cepstral. On obtient ainsi les vecteurs cepstraux qui permettent de réestimer la moyenne et la variance de la gaussienne en question. Comme pour la PMC, on suppose que le bruit n'affecte pas la distribution des trames entre les gaussiennes d'un état donné. Il est clair qu'il s'agit d'une supposition dont la validité dépend du type de bruit et de son niveau. Pour surmonter ce problème, il est possible d'adopter le même raisonnement au niveau de l'état et non au niveau de la gaussienne. Ceci implique une reclassification des trames (en utilisant l'algorithme EM) dans chaque état. Il est également possible d'utiliser l'ancienne classification comme initialisation de l'algorithme EM. Ainsi seul le calcul nécessaire pour ajuster les modèles est effectué. Cette technique est similaire à Data Driven PMC (DDPMC) utilisée par Gales et Young[8], la différence principale est que la technique DDPMC regénère aléatoirement les vecteurs associés à chaque état au lieu de les stocker au moment de l'apprentissage. Ceci est nécessairement coûteux en temps de calcul si on veut générer un nombre suffisant de vecteurs pour avoir une validité statistique.

La composition directement sur les données permet aussi de réaliser la soustraction du cepstre moyen. Pour faire ceci correctement, il faut estimer h ou n à partir du bruit $h * n$[14]. Le bruit peut être estimé d'une manière itérative partant des trames n_0 de silence dans les données de test. Ces trames de silence peuvent être utilisées pour calculer le cepstre moyen de la parole bruitée, qui est ensuite soustrait au cepstre moyen des données d'adaptation pour obtenir une première estimation de $\tilde{h}$. Le filtre $\tilde{h}^{-1}$ est alors appliqué aux données d'adaptation pour obtenir une meilleure estimation de n. En pratique 5 itérations suffisent pour avoir une bonne estimation de h.

TECHNIQUES BASÉES SUR UN CRITÈRE D'OPTIMALITÉ

Ces techniques ne supposent en général aucun modèle *à priori*. Elle sont basées sur un critère d'optimalité, généralement le maximum de vraisemblance. Parmi les travaux les plus importants dans ce cadre, on peut citer la régression linéaire (MLLR)[6][11].

Le système est adapté à un nouvel environnement par des transformations linéaires des paramètres. Les transformations sont estimées en alignant les données d'adaptation avec les états des modèles. Il est supposé qu'on ne dispose que d'une petite quantité de données dans le nouvel environnement, ce qui rend impossible l'adaptation individuelle de chaque paramètre du modèle (les données ne couvrent pas d'une manière suffisante la totalité des modèles). Pour cela on procède à une classification des paramètres où chaque classe subira la même transformation.

La moyenne d'une densité gaussienne est adaptée en utilisant une transformation linéaire W. W est une matrice $(n \times (n+1))$ contenant éventuellement une colonne supplémentaire pour modéliser *l'offset*. La moyenne adaptée est donnée par: $\mu_{ad} = W\hat{\mu}$. La valeur choisie pour W est celle qui maximise la vraisemblance que les données d'adaptation soient générées par les modèles adaptés. La technique utilisée pour ceci est l'algorithme EM, c'est-à-dire, la détermination de la fonction auxiliaire puis sa maximisation. D'une

Tab. 1 - *Taux d'erreur (sur les mots) avec et sans technique de compensation pour les deux types de données (C0: propres, P0: bruitées).*

PMC compens.	*MLLR adapt.*	*Taux d'erreur (%)*	
		P0	*C0*
non	non	>50.0	10.4
oui	non	20.5	10.4
oui	oui	17.5	9.1

manière similaire, on peut adapter les variances.

RÉSULTATS ET CONCLUSIONS

Les résultats de ces analyses nous ont amenés à incorporer certaines techniques de compensation du bruit dans le système de reconnaissance de parole continue du LIMSI[14] qui utilise un vocabulaire de 65.000 mots. Nous présentons ici les résultats obtenus sur les données de test ARPA NAB CSR de novembre 1995 qui ont été enregistrées dans un environement bruité (47 à 61dBA). Ces données de test comprennent 300 phrases prononcées par 20 locuteurs (15 phrases par locuteur,$\cong$ 20 mots/phrase). Chaque énoncé a été enregistré simultanément avec deux microphones: un microphone Sennheiser HMD-410 (rapport S/B moyen égal à 30dB), et un microphone inconnu du système (rapport S/B de 7 à 18dB).

Le même système de reconnaissance, entraîné uniquement sur des données propres enregistrées avec un microphone Sennheiser HMD-410, a été utilisé pour décoder les deux ensembles de données (C0: Sennheiser microphone, P0: autres microphones). Les résultats en terme de taux d'erreur sur les mots sont donnés dans le tableau 1. On peut observer que l'utilisation des techniques de compensation améliore d'une manière significative les performances sur les données bruitées. La technique PMC compense le bruit additif lié à l'environnement et le bruit convolutif lié au microphone. Le grand avantage de cette technique est qu'elle améliore de manière très significative les performances dans le cas de présence de bruit additif ou convolutif (> 50%/20.5% d'erreurs), tout en gardant les mêmes performances dans le cas d'absence de bruit (10.4%/10.4% d'erreurs). L'adaptation MLLR compense les différences résiduelles entre les données d'apprentissage et celles du test correspondant d'une part au bruit non representé dans le modèle du canal de transmission, et d'autre part, à la variabilité interlocuteur. Une seule matrice de regression (49 × 48) a été utilisée pour transformer les moyennes des modèles. D'après la dernière ligne du tableau 1, on constate que la technique MLLR apporte des améliorations dans tous les cas. Ces résultats, obtenus avec un système de reconnaissance de parole continue, indépendant du locuteur, avec un vocabulaire de 65.000 mots (vocabulaire de test illimité), montrent la pertinence des techniques de compensation retenues.

RÉFÉRENCES

[1] D. V. Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech & Language*, pp. 151-167, 1989.

[2] A.P. Varga, R.K. Moore, " Hidden Markov model decomposition of speech and noise," *ICASSP-90*, pp. 845-848.

[3] S.F Boll, " Suppression of Acoustic Noise in Speech Using Spectral Substraction," *IEEE*, pp. 113-120, 1979.

[4] F. Martin, K. Shikano, Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," *EuroSpeech'93.*

[5] A. Sankar, C.-H. Lee, "Robust Speech Recognition based on Stochastic Matching," *ICASSP-95*, pp. 121-124.

[6] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, pp. 171-185, 1995.

[7] M.J.F. Gales, S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech & Language*, pp. 289-307, 1995.

[8] M.J.F. Gales, S.J. Young, "A fast and flexible implementation of parallel model combination," *ICASSP-95*, pp. 133-136.

[9] A. Acero, R.M. Stern, "Enviromental Robustness in Automatic Speech Recognition," *IEEE Acoustics, Speech & Signal Processing*, pp. 849-852. April 1990.

[10] B. Atal, "Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification," *Journal of the Acoustic Society of America*, **55**, pp. 1304-1312. June 1974.

[11] O. Siohan, Y. Gong, J-P. Haton, "Noise adaptation using linear regression for continuous noisy speech recognition" *EuroSpeech'95*, pp. 465-468.

[12] O. Cappé, "Techniques de réduction de bruit pour la restauration d'enregistrements musicaux" *thèse TELECOM Paris 93 E 019*, sept 93.

[13] X. Huang, A. Acero, F. Alleva, M-Y. Hwang, L. Jiang and M. Mahajan "Microsoft windows highly intelligent speech recognizer: whisper" *ICASSP-95*, pp. 93-96.

[14] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96.*