Transcription de la parole conversationnelle

J.L. Gauvain, G. Adda, L. Lamel, F. Lefèvre, H. Schwenk

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133
91403 Orsay Cedex, FRANCE
{gauvain,gadda,lamel,lefevre,schwenk}@limsi.fr

ABSTRACT

This paper describes the development of a speech recognition system for the processing of conversational speech, starting with a state-of-the-art broadcast news transcription system. We identify major changes and improvements in acoustic and language modeling, as well as decoding, which are required to achieve good performance on conversational speech. Some major changes on the acoustic side include the use of speaker normalizations (VTLN and SAT), the need for better pronunciation modeling and the use of discriminative training (MMIE). On the linguistic side the primary challenge is to cope with the limited amount of language model training data. To address this issue we make use of a data selection technique, and a smoothing technique based on a neural network language model. At the decoding level, lattice rescoring and minimum word error decoding are applied. On the development data, the improvements yield an overall word error rate of about 21% whereas the original BN transcription system had a word error rate of about 50% on the same data.

1. INTRODUCTION

La transcription de conversations téléphoniques est une tâche bien plus complexe que la transcription d'émissions radio ou télédiffusées. Dans cet article nous décrivons les travaux récemment conduits au LIMSI pour faire évoluer notre système de transcription d'émissions d'information (*Broadcast News*, BN) vers un système de transcription de conversations. Cette tâche est depuis plusieurs années au coeur des campagnes annuelles d'évaluation de systèmes de reconnaissance de parole organisées par le NIST, utilisant la famille des corpus SwitchBoard (SWB) collectés par le LDC [6]. Ces évaluations ont permis de mettre en évidence les principales difficultés rencontrées pour le traitement automatique de la parole conversationnelle [13, 14, 11, 8].

Le système de transcription SWB du LIMSI repose sur les mêmes composants que le système de transcription BN. Les ajouts principaux sont : la normalisation de la longueur du conduit vocal (*Vocal Tract Length Normalisation*, VTLN), une adaptation MLLR contrainte et une adaptation MLLR avec plusieurs classes de régression, un apprentissage adaptatif (*Speaker Adaptative Training*, SAT) et un apprentissage discriminant (*Maximum Mutual Information Estimation*, MMIE), l'utilisation de probabilités de prononciation, un modèle de langage (ML) neuronal, et un décodage par consensus. Certaines de ces techniques qui s'étaient révélées peu efficaces sur les données BN, en particulier la normalisation VTLN et les probabilités de prononciation, s'avèrent très utiles pour la parole conversationnelle,

Après une description succincte du système de transcription BN qui constitue le point de départ pour nos développements, nous décrivons les changements effectués pour traiter la parole conversationnelle. Ces changements concernent les modèles acoustiques, le modèle linguistique et la procédure de décodage. Nous précisons l'impact de ces améliorations sur le taux d'erreur.

2. SYSTÈME DE RÉFÉRENCE

Le système de transcription BN du LIMSI repose sur deux composants principaux : un segmenteur audio et un décodeur lexical [3]. Le décodeur utilise des modèles de Markov cachés avec densités de probabilité continues (sommes pondérées de gaussiennes) pour les modèles acoustiques, et des statistiques n-grammes obtenus sur de grands corpus de textes pour modèle linguistique. Les modèles de Markov cachés représentent des allophones contextuels avec une structure gauche-droite à états liés. Ils modélisent des séquences de trames centisecondes avec 39 composantes, 12 coefficients cepstraux (PLP) et le logarithme de l'énergie à court-terme, avec leurs dérivées d'ordre 1 et 2.

Le décodage en mots est effectué en trois passes. La première passe produit une hypothèse qui est utilisée pour réaliser l'adaptation MLLR [10] non supervisée des modèles acoustiques. Les modèles adaptés sont utilisés dans la seconde passe pour générer un graphe de mots. Ces deux passes utilisent un modèle trigramme. L'hypothèse finale est générée avec un modèle quadrigramme et les modèles acoustiques adaptés lors de la seconde passe. La première passe utilise un jeu d'allophones représentant environ 5500 contextes avec 6300 états liés. Les passes 2 et 3 utilisent des modèles plus gros, représentant 11000 contextes phonétiques et 11700 états liés, avec respectivement 16 et 32 gaussiennes par état. L'ensemble du décodage est effectué en moins de 10 fois le temps réel. Le regroupement des états est réalisé en créant un arbre de décisions pour chaque état de chaque phonème de façon à maximiser la vraisemblance des données d'apprentissage pénalisée par le nombre d'états liés. Nous utilisons un ensemble de 184 questions relatives à la position du phonème dans le mot, et relatives aux caractéristiques acoustiques du phonème et de ses voisins immédiats.

Pour l'anglais-américain, sans contrainte particulière, le système de transcription BN a un taux d'erreur sur les mots de l'ordre de 20%[1]. Ce système a été adapté à 5 autres langages (Arabe, Français, Allemand, Mandarin et Espagnol) avec des performances comparables [4].

Considérant le niveau de développement de nos modèles BN, il paraissait intéressant d'évaluer ce système sur de la parole conversationnelle sans faire aucune modification. Cette évaluation a été menée sur les données de test de l'évaluation NIST

[1] Dans le cadre des évaluations NIST le taux d'erreur de ce système est proche de 10%.

Hub5 1998 (Eval98)[2]. Le taux d'erreur initial est 61,2%. En utilisant les transcriptions de SWB pour construire le modèle de langage, le taux d'erreur est réduit à 57% (le vocabulaire n'a pas été modifié dans la mesure où le taux de mots hors-vocabulaire reste inférieur à 1%). En combinant ce modèle de langage avec les modèles acoustiques estimées sur les données SWB, on obtient un taux d'erreur de 46,7%. Ces résultats montrent qu'une part importante de l'écart entre les modèles BN et les données SWB réside dans la modélisation acoustique, et que réestimer les modèles BN sur les données SWB en appliquant les techniques développées pour les données BN n'est pas suffisant pour obtenir des performances acceptables sur des données conversationnelles.

3. MODÉLISATION ACOUSTIQUE

Les données audio BN sont principalement de type large bande, la part des données de qualité téléphonique étant très réduite. Les conversations du corpus SWB sont de qualité téléphonique et ont été enregistrées sur 2 canaux correspondant chacun à un coté de la conversation. Comme pour les données BN, les vecteurs centisecondes comprennent 39 coefficients cepstraux obtenus à partir de spectres en échelle Mel. Ils sont estimés sur une bande réduite à 0-3.8kHz (à comparer à 0-8kHz comme pour les données BN). La moyenne et la variance de chaque coefficient cepstral sont normalisées pour chaque côté de la conversation, alors que pour les données BN la normalisation est réalisée par groupe de segments supposés correspondre à un seul locuteur.

Les modèles phonétiques SWB ont la même topologie et sont construits de la même manière que les modèles BN. Ils sont estimés sur toutes les données transcrites disponibles des corpus SwitchBoard (3606 conversations) et CallHome (120 conversations). Environ 3% des données CallHome et 10% des données SwitchBoard ont été rejetées durant l'alignement forcé entre le signal et la transcription manuelle. Au total nous utilisons 430 heures de données, comprenant une quantité à peu près égale de femmes et d'hommes. Ces données permettent de modéliser 32k contextes phonétiques avec environ 12k états liés. Deux ensembles de modèles acoustiques dépendant du genre (*gender-dependent*, GD) sont obtenus après une adaptation MAP [5] des modèles indépendant du genre (*gender-independent*, GI).

Tous les résultats mentionnés ci-après ont été obtenus sur les données de test de l'évaluation Hub5 2001 du NIST (Eval01), composé de 3 sous-ensembles de 20 conversations chacun provenant des corpus SwitchBoard-I, SwitchBoard-II et SwitchBoard-II cellulaire pour un total d'environ 6h de signal.

Normalisation spectrale

La normalisation de la longueur de conduit vocal [2] est une technique de normalisation du locuteur intervenant au niveau des paramètres acoustiques qui est largement répandue dans les systèmes de reconnaissance à grands vocabulaires. Cette normalisation repose sur une modification linéaire de l'échelle des fréquences afin de compenser les différences de longueur de conduit vocal entre les locuteurs. Le spectre de puissance en échelle Mel est estimé à partir d'un banc de filtres modifié selon une fonction linéaire par morceaux (pour ne pas sortir de la bande 0-3.8kHz). Le coefficient de normalisation est sélectionné parmi un ensemble de valeurs (0,8 à 1,25) pour maximiser la vraisemblance des données de test, en utilisant une transcription obtenue avec un décodage rapide. Bien que cette procédure d'estimation (recherche du maximum de vraisemblance) conduise à une normalisation qui permet de réduire significativement le taux d'erreur, les itérations de cette procédure ne convergent pas correctement pour les données d'apprentissage, ceci pouvant être attribué au fait que le Jacobien de la transformation est ignoré. Une prise en compte correcte du Jacobien suppose de construire un modèle acoustique par valeur possible du coefficient VTLN, ce qui doublerait les temps de calculs nécessaires à l'estimation de ce coefficient. Pour résoudre le problème sans compensation explicite du Jacobien, nous estimons le coefficient de la transformation au moyen de modèles spéciciques à chaque genre, rendant ainsi la distribution des coefficients uni-modale (pour chaque genre). Nous utilisons également des modèles mono-gaussians au lieu de modèles multi-gaussians. Après ces deux modifications, la procédure d'estimation devient très stable.

[2]Signalons que ce test est plus difficile que celui de 2001 (Eval01), qui est utilisé dans la suite de l'article (les taux d'erreur sont environ 25% plus élevés sur Eval98).

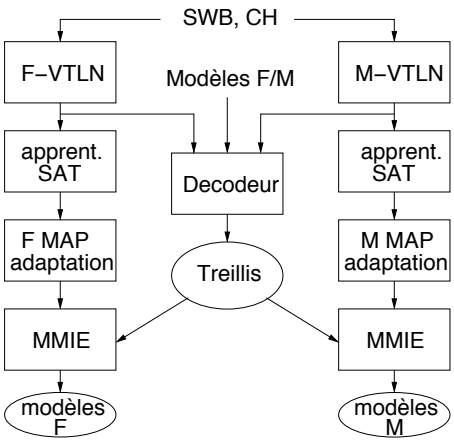

Figure 1: Procédure d'apprentissage des modèles acoustiques

Bien que les modèles servant à l'estimation des coefficients VTLN (pour les données d'apprentissage et de test) soient appris séparément sur les données des femmes et des hommes, les modèles utilisés lors de la reconnaissance sont estimés sur l'ensemble des données (cf. figure 1), puis ils sont adaptés avec les données d'un seul genre. Les résultats expérimentaux sont regroupés dans le tableau 1 pour des modèles sans normalisation VTLN, des modèles avec une estimation des coefficients indépendante du genre (GI) et dépendante du genre (GD), et avec et sans adaptation MLLR. On peut observer que sans adaptation des modèles acoustiques, la normalisation VTLN réduit le taux d'erreur d'environ 2%. Ce gain est réduit à 1,5% après adaptation des modèles. Un gain supplémentaire de 0,4% est obtenu par la procédure d'estimation dépendante du genre.

VTLN	MLLR	SWB1	SWB2	CELL	total
non	non	28,0	36,1	42,2	35,6
GI	non	26,7	33,5	40,4	33,7
non	oui	26,1	32,0	38,1	32,2
GI	oui	24,4	30,2	36,9	30,7
GD	oui	24,2	30,1	36,2	30,3

Table 1: Taux d'erreur pour les 3 sous-ensembles du corpus Eval01 avec des modèles GD, avec et sans normalisation VTLN.

Apprentissages adaptatif et discriminant

L'apprentissage adaptatif SAT [1] vise à réduire l'impact des variations inter-locuteurs lors de l'estimation des modèles acoustiques. Pour ce faire une transformation linéaire contrainte [7] est estimée pour chaque locuteur du corpus

d'apprentissage afin de rapprocher les données du locuteur du modèle multilocuteur. Un nouveau modèle est alors construit en utilisant les données d'apprentissage transformées. Ce modèle canonique est utilisé lors du décodage pour faciliter l'adaptation non-supervisée. La réduction absolue du taux d'erreur par rapport à une adaptation MLLR sans modèle SAT est de l'ordre de 0,5%. L'utilisation d'une adaptation MLLR contrainte apporte un gain additionnel de 0,5%.

L'importance de l'apprentissage discriminant a été clairement montrée pour la parole conversationnelle [17]. Nous avons développé une procédure d'apprentissage de type MMIE [17] compatible avec l'apprentissage de Viterbi utilisé dans notre système. Cette procédure commence par la génération avec un modèle bigramme d'un treillis de mots pour chaque tour de parole du corpus d'apprentissage. Ces treillis sont ensuite redécodés avec un modèle unigramme de façon à réduire l'influence du modèle de langage et à augmenter le nombre de confusions. Hormis la génération initiale des treillis, l'algorithme de réestimation n'est pas plus lent que l'algorithme non discriminant. Il permet de réduire de 1,9% absolu le taux d'erreur sans adaptation et de 1,2% avec adaptation.

4. MODÉLISATION LINGUISTIQUE

La principale difficulté de la modélisation linguistique pour la parole conversationnelle réside dans la faible quantité de données d'apprentissage disponible. Pour la tâche BN, il est relativement aisé de trouver une grande variété de textes pertinents pouvant servir de données d'apprentissage. Pour la parole conversationnelle, la seule source disponible est la transcription des données audio d'apprentissage (environ 4,5M mots correspondant à 430h). Nous proposons deux approches permettant de compenser en partie ce problème. La première approche consiste à inclure dans les données d'apprentissage des textes d'autres sources sélectionnés pour leur nature conversationnelle. La seconde approche consiste à utiliser un réseau de neurones pour lisser les probabilités n-grammes.

Le vocabulaire de reconnaissance contient 51k mots comprenant les mots apparaissant au moins 2 fois dans les données SWB (4,5M de mots), complémentés des mots les plus fréquents dans un corpus de transcriptions BN (320M de mots). L'utilisation des transcriptions BN permet d'obtenir une meilleure couverture de la langue anglaise. Une différence importante avec le vocabulaire BN est l'intégration des interjections "uh-huh" et "mhm" (signifiant oui) et "uh-uh" (signifiant non) qui n'étaient pas considérées comme des éléments lexicaux. Comme pour le système BN, les 300 séquences lexicales les plus fréquentes et sujettes à des phénomènes de réduction, sont modélisées sous forme de mots composés. Par contre, les acronymes, nombreux dans BN, sont rares dans SWB et ne sont pas traités comme des mots. La couverture lexicale atteint 99.7% sur Eval01.

Un modèle n-grammes a été construit pour chaque source de données au moyen du SRI LM toolkit [15]. Nous avons obtenu les meilleurs résultats avec le lissage de Kneser-Ney modifié. Ces modèles ont ensuite été interpolés en utilisant l'algorithme EM pour estimer les coefficients d'interpolation qui minimisent la perplexité du modèle résultant.

Afin d'augmenter la quantité de données conversationnelles utilisée pour la construction du modèle, nous avons sélectionné dans le corpus de transcriptions BN les données les plus proches des données conversationnelles (65M de mots) au moyen d'un critère de perplexité [9]. Un corpus "conversationnel" de 180M de mots sélectionné sur le WEB par l'Université de Washington a également été inclus dans les données d'apprentissage.

Dans un modèle n-gramme, les mots sont habituellement représentés dans un espace discret (indices dans le lexique) rendant toute interpolation difficile. L'idée de base du ML neuronal consiste à projeter les mots dans un espace continu et à estimer ensuite les probabilités n-grammes dans cet espace. Ainsi les distributions de probabilités sont des fonctions continues des représentations des mots, permettant une meilleure généralisation à des n-grammes non-observés. Le réseau de neurones peut estimer conjointement la projection des mots dans l'espace continu et les probabilités n-grammes en minimisant la perplexité des données d'apprentissage [16]. Le ML neuronal n'a été estimé que sur la transcription des données audio (4,5 M de mots), et il est interpolé avec le modèle n-gramme classique lors du décodage.

Le tableau 2 résume les gains en perplexité et en taux d'erreur sur les mots pour les différents ML. Tous ces résultats correspondent à une réévaluation des treillis de la dernière passe de décodage (cf. section 6).

Modèle	SWB+BN	+ BNselect & WEB	+ RN
perplexité	58,5	56,9	54,5
taux d'erreur	21,7	21,5	21,1

Table 2: Comparaison des différents modèles de langage sur le test Eval01. (SWB+BN : corpus SWB et BN, BNselect : sélection de données conversationnelles dans le corpus BN, WEB : corpus conversationnel extrait du WEB.)

5. PRONONCIATIONS

Les prononciations sont basées sur les mêmes 48 phones utilisés dans le système BN (dont 3 pour les pauses, les hésitations et les respirations). Un graphe de prononciation est associé à chaque mot afin d'autoriser des variantes telles que les réductions. Les prononciations de base sont extraites du lexique anglais-américain du LIMSI. Les formes les plus fréquemment et fortement modifiées par la nature conversationnelle des données ont été vérifiées et adaptées aux données SWB. Le dictionnaire de prononciations a un total de 60586 transcriptions phonétiques pour 51075 mots. Les probabilités des prononciations sont évaluées à partir des fréquences d'occurrences relevées dans les alignements forcés des transcriptions sur les données d'apprentissage. La prise en compte des probabilités des prononciations n'a jamais donné de gain significatif sur les performances du système BN, tandis que pour les données SWB, le gain absolu sur le taux d'erreur est de 1,9% avant adaptation et de 0,4% après adaptation MLLR.

6. DÉCODAGE

La procédure de décodage a été largement modifiée. Les principales modifications portent sur l'estimation du coefficient VTLN, la procédure d'adaptation au locuteur et le décodage par consensus utilisant des probabilités de prononciation. Le décodage est réalisé en 3 étapes. Dans la première passe, le genre du locuteur est déterminé pour chaque côté de la conversation (au moyen de 2 GMM) et un décodage trigramme rapide permet de générer une transcription initiale (cf. tableau 3, ligne 1). Cette transcription sert d'une part à estimer les coefficients VTLN pour chaque côté des conversations et d'autre part à adapter les modèles SAT qui sont utilisés dans la seconde passe. Les passes 2 et 3 utilisent les données normalisées. Chacune génère un treillis de mots avec un modèle trigramme. Ce treillis est ensuite réévalué avec un modèle quadrigramme et est transformé en réseau de consensus en intégrant les probabilités de prononciation. Les probabilités *a posteriori* des arcs du treillis sont estimées par l'algorithme avant-arrière. Les réseaux

Passe	AM	VTLN	MLLR	ML	CN	RTF	Err.
1	MMIE	non	non	3g	non	1,9	36,1
2	SAT, MMIE	oui	2	3g	non	13,9	23,6
		oui	2	4g	non	0,0	22,9
		oui	2	4g	oui	0,0	22,1
3	SAT, MMIE	oui	5	4g + RN	oui	3,1	21,1

Table 3: Taux d'erreur sur les données Eval01 pour chaque passe de décodage. (MLLR : le nombre de classe de régression est spécifié pour chaque passe. CN : décodage par réseau de consensus avec probabilités de prononciation. RTF : facteur de temps réel.)

de consensus sont obtenus en fusionnant de façon itérative les noeuds des treillis et en dupliquant les arcs jusqu'à ce qu'un graphe linéaire soit obtenu pour chaque treillis. Cette procédure permet d'obtenir des résultats comparables à ceux obtenus avec l'algorithme de regroupement d'arcs proposé dans [12] mais elle se révèle significativement plus rapide. La transcription est obtenue en prenant les mots les plus probables pour chaque ensemble de confusions.

Les transcriptions des passes 1 et 2 sont utilisées lors de la passe suivante pour les adaptations MLLR (contrainte et non contrainte). Une seule classe de régression est utilisé pour l'adaptation contrainte, alors que pour l'adaptation non contrainte nous utilisons 2 classes de régression (parole/non parole) en deuxième passe et 5 classes (non parole, consonnes non voisées, consonnes voisées, et deux classes de voyelles) pour la dernière passe. Afin d'accélérer le décodage, l'espace de recherche pour cette dernière passe est restreint au treillis de la seconde passe après transformation en graphe de mots.

Le taux d'erreur obtenu après chaque passe est donné dans le tableau 3, pour le jeu de test Eval01. La réduction importante du taux d'erreur entre la première et la seconde passe est due à la combinaison de la transformation VTLN, des adaptations acoustiques, du modèle quadrigramme, des probabilités de prononciation et du décodage par réseau de consensus (la contribution de chaque composant est détaillée pour la seconde passe). Le gain de la troisième passe vient de l'adaptation acoustique avec 5 classes de régression et de l'utilisation du ML à réseau de neurones.

Le temps de calcul pour chaque étape est indiqué dans la colonne RTF du tableau 3. L'étape la plus longue est la génération des treillis de mots qui nécessite en moyenne un temps égal à environ 14 fois la durée des données à décoder. La réévaluation des treillis avec un modèle quadrigramme et le décodage par consensus représentent par contre un temps négligeable. Au total le temps de décodage est égal à 18,9 fois la durée du signal.

7. CONCLUSIONS

Nous avons décrit le travail mené au LIMSI pour développer un système de transcription de conversations téléphoniques à partir d'un système de transcription d'émissions d'information. Il apparaît que le traitement de la parole conversationnelle requiert des modifications importantes tant en ce qui concerne les modèles acoustiques et linguistiques que le processus de décodage. Le taux d'erreur initial du système BN sur les données conversationnelles était de l'ordre de 50%. La simple réestimation des paramètres des modèles sur les données SWB ne conduit pas à un taux d'erreur satisfaisant. Il a fallu revoir l'ensemble des composants du système. Les éléments suivants on été ajoutés : une normalisation spectrale (VTLN), un apprentissage adaptatif SAT et un apprentissage discriminant des modèles acoustiques, une adaptation MLLR contrainte, l'utilisation de classes phonémiques pour l'adaptation MLLR, un modèle de langage neuronal, et une décodage par réseau de consensus utilisant des probabilités de prononciation. Tous ces raffinements ont permis d'obtenir un taux d'erreur de 21% avec un temps de décodage inférieur à 20 fois la durée des données.

RÉFÉRENCES

[1] T. Anastasakos and J. McDonough and R. Schwartz and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," *ICSLP'96*, 2:1137-1140.

[2] A. Andreoum T. Kamm and J. Cohen, "Experiments in Vocal Tract Normalisation," *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[3] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, 37(1-2):89-108, May 2002.

[4] J.L. Gauvain and L. Lamel, "Structuring Broadcast Audio for Information Access," *EURASIP journal on Applied Signal Processing*, 2003(2):140-150, February 2003.

[5] J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, 2(2):291-298, April 1994.

[6] J.J. Godfrey, E.C. Holliman, J. McDaniel, "Switchboard: Telephone speech corpus for research and development," *ICASSP'92*, 1:517-520.

[7] M. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech recognition," *Computer, Speech and Language*, 12(2):75-98, 1998.

[8] T. Hain, P.C. Woodland, T.R. Niesler and E.W.D. Whittaker, "The 1998 HTK System for Transcription of Conversational Telephone Speech," *ICASSP'99*.

[9] R. Iyer and M. Ostendorf, "Relevance weighting for combining multi-domain data for n-gram language modeling," *Computer Speech & Language*, 13:267-282, 1999.

[10] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, 9(2):171-185, 1995.

[11] A. Ljolje et al., "The AT&T LVCSR-2000 System," *Proc. NIST Speech Transcription Workshop*, 2000.

[12] L. Mangu, E. Brill, A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeeech'99*, 495-498.

[13] S. Matsoukas, T. Colthurst, O. Kimball, A. Solomonoff, F. Richardson, C. Quillen, H. Gish, P. Dongin, "The 2001 Byblos English Larve Vocabulary Conversational Speech Recognition System," *ICASSP'02*, I:721-724.

[14] A. Stolcke et al., "The SRI March 2000 Hub-5 Conversational Speech Transcription System," *Proc. NIST Speech Transcription Workshop*, 2000.

[15] A. Stolcke, "SRILM - An extensible language modeling toolkit." *ICSLP'02*, 2:901-904.

[16] H. Schwenk and J.L. Gauvain, "Connectionist Language Modeling for LVCSR," *ICASSP'02*.

[17] P. Woodland and D. Povey, "Large scale discriminative training for speech recognition," *ISCA ASR'00*, 7-16.