CONVERSATIONAL TELEPHONE SPEECH RECOGNITION

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

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,lamel,schwenk,gadda,clz,lefevre}@limsi.fr

ABSTRACT

This paper describes the development of a speech recognition system for the processing of telephone conversations, 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 state-of-the-art performance on conversational speech. Some major changes on the acoustic side include the use of speaker normalization (VTLN), the need to cope with channel variability, and the need for efficient speaker adaptation and better pronunciation modeling. 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 24.9% whereas the original BN transcription system had a word error rate of about 50% on the same data.

1. INTRODUCTION

It is well known that transcribing conversational telephone speech is a significantly more challenging task than the transcription of broadcast news (BN). This paper reports on recent work at LIMSI on moving from the transcription of BN data to conversational telephone speech data. Conversational telephone speech recognition has been one of the focal tasks in annual speech recognition benchmarks organized by NIST, using the SwitchBoard (SWB) family of resources distributed by the LDC [6]. The benchmark tests have demonstrated many of the difficulties encountered in automatic processing of conversational speech [11, 12, 13, 14].

The LIMSI SWB speech-to-text system relies on the same basic components as the LIMSI BN system [3]. Additional features specific to the SWB system are: vocal-tract length normalization (VTLN), multiple regression class MLLR adaptation, pronunciation probabilities, neural-network language model, and consensus decoding. Some of these techniques (in particular VTLN and pronunciation probabilities) which had not helped in our BN transcription system, quite significantly improve the performance of our SWB system.

The remainder of this paper is as follows. We first overview the LIMSI BN system which served as the starting point for the SWB system. Then the modifications in acoustic modeling, language modeling and decoding are described and the performance improvements demonstrated.

2. BASELINE SYSTEM

The LIMSI broadcast news transcription system has two main components, an audio partitioner and a word recognizer [3]. The word recognizer uses continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree. The acoustic feature vector has 39-components comprised of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Word recognition is performed in three steps. The first step generates initial hypotheses which are used in cluster-based acoustic model adaptation using the MLLR technique [9] prior to word graph generation in the second step. Both of these steps use a 3-gram language model (LM). The final hypotheses are generated with a 4-gram LM and acoustic models adapted with the step 2 hypotheses. The first decoding pass uses a small set of acoustic models with about 5500 contexts and 6300 tied states. The second and third pass acoustic models cover about 11000 phone contexts, represented with a total of 11700 tied-states, and 16/32 Gaussians per state, respectively. State-tying constructs one tree for each state of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states. The set of 184 questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

Given the level of development of our BN models, it was of interest to benchmark the system on conversational data without any modifications. These first experiments were done using the H5 Eval98 test set which is significantly harder than the Eval01 test used in the rest of the paper (the word error rate is about 25% higher than on Eval01). The experiments all use the same 3-pass BN decoding strategy and run in less than 10xRT. Using both the BN acoustic and language models results in a word error rate of 61.2%. Keeping the same word list (the OOV rate of the H5 Eval98 data is under 1% with the BN wordlist) and retraining the language model on the SWB transcriptions reduces the word error rate to 57.0%. Using the SWB LM with acoustic models trained on the SWB data reduces the word error rate to 46.7%. This initial experiment demonstrates that a large part of the mismatch between the BN system and the SWB data is linked to the acoustic models, and that simply training models on the SWB data with our BN recipes was not enough to achieve state-of-the-art performance on conversational data.

3. ACOUSTIC MODELING

BN audio data are for the most part wideband, with telephone speech data accounting for only a very minor portion of the data. The SWB data consist of 2-channel telephone recordings, and although each conversation side was recorded on a separate channel, there is significant echo problem for much of the data. As for BN, the front-end uses 39 PLP-like cepstral features derived from a Mel frequency spectrum every 10ms. These are estimated on the 0-3.8kHz band as opposed to the 8kHz bandwidth used for BN. Cepstral mean and variance normalization are carried out on each conversation side, whereas for BN they are computed per speaker cluster as determined by the data partitioner.

The SWB phone models use the same topology and are constructed in the same manner as the BN models. They are trained on all of the available transcribed CallHome (240 conversation sides) and SwitchBoard (4866 conversation sides) and SWB cellular (460 conversation sides, excluding the 2001 eval set) data sets, using the ISIP transcriptions. About 3% of the CallHome data and 10% of the SwitchBoard data were rejected during forced alignment. The acoustic models are trained on a total of about 229 hours of data, with roughly equal amounts of male and female data.

The acoustic models are position-dependent triphones (28k phone contexts) with about 11600 tied states, obtained using the same divisive decision tree based clustering algorithm and the same set of questions as used in the BN system. The most frequent phone contexts in the training data are modeled, with separate cross-word word-internal statistics. Two sets of gender-dependent acoustic models were built using MAP adaptation of SI seed models [5]. A second set of gender-specific acoustic models were estimated by MAP adapting the above models with SWB cellular data.

All results reported in this paper were obtained on the H5 Eval01 test set composed of 3 subsets of 20 conversations from the SwitchBoard-1, SwitchBoard-2, and SwitchBoard-cellular corpora, for a total of about 6 hours of audio data.

Vocal tract length normalization

Vocal tract length normalization, a technique which performs a simple speaker normalization at the front-end level [1], is now often used in LVCSR. The normalization consists of performing a frequency warping to account for differences in vocal tract length, where the appropriate warping factor is chosen from a set of candidate values by maximizing the test data likelihood based on a first decoding pass transcription and some acoustic models (some sites use GMMs with no need for transcriptions). In the past we tried applying VTLN to the BN task, but were unsuccessful at significantly improving our state-of-the-art results. However given the significant gains reported by other sites on conversational telephone speech data [7], we decided to reconsider our point of view. Following [7], the MEL power spectrum is computed with a VTLN warped filter bank using a piecewise linear scaling function. We found the classical maximum-likelihood estimation procedure to be unsatisfying, as iterative estimates on the training data did not converge properly, even though a significant word error reduction on conversational speech was obtained. This problem can be attributed to the fact the VTLN Jacobian is simply ignored during the ML estimation, although the normalization of the feature variances should largely compensate for this effect. Properly compensating the VTLN Jacobian would require building models for each possible warping value and would double the computation time to estimate the warping factors, we therefore investigated changing the procedure to avoid the Jacobian compensation.

The VTLN warping factors are still estimated for each conversation side by aligning the audio segments with their word level transcription for a range of warping factors (between 0.8 and 1.25), but we use single-Gaussian gender-dependent models to determine the ML warping factor. By using gender dependent models (as proposed in [2]) the warping factor histogram becomes unimodal and is significantly more compact. This effect and the use single Gaussian models (as proposed in [17]) significant reduces the need for Jacobian compensation and makes the estimation procedure very stable.

Even though the models used to estimate the warping factors (for the training and test data) are trained separately on the female and male data, the gender-dependent models used by the recognizer are trained on all the data using a standard MAP estimation procedure from SI seed models trained on all the data. Experimental results are given in Table 1 for gender-dependent models trained without VTLN, models trained with SI warping and models with F/M warping before and after MLLR adaptation (i.e. 2 pass decode). It can be seen that without acoustic model adaptation (Table 1 top) VTLN reduces the word error rate about 2%, this gain is reduced to 1.5% after MLLR adaptation (Table 1 bottom). A additional gain of 0.4% is obtained by using the gender-dependent warping.

VTLN	MLLR	SWB1	SWB2	CELL	all
n	n	28.0	36.1	42.2	35.6
SI	n	26.7	33.5	40.4	33.7
n	y	26.1	32.0	38.1	32.2
SI	y	24.4	30.2	36.9	30.7
F/M	y	24.2	30.1	36.2	30.3

Table 1: Word error rate on the 3 subsets of Eval01 using VTLN and MAP trained gender-dependent model without (top) and with (bottom) MLLR adaptation. (SI: gender-dependent VTLN warping, F/M: gender-dependent VTLN warping)

Dealing with cellular data

As mentioned in above, a set of models were trained to better match the cellular data. For each conversation side, the set of acoustic models used by the decoder is chosen by computing the likelihood ratio (by forced alignment with the first decoding pass hypotheses) for the standard SWB models and the SWB cellular models and comparing the likelihood ratio to a fixed threshold. The likelihood ratio is computed ignoring all frames with an energy lower than 20dB under the peak energy for each speaker turn. The decision threshold is set to have a negligible rate of false detections of cellular data. As shown in Table 2[1], the combination of the switch and the SWB cellular models reduces the word error on the cellular test data, without hurting performance on the non-cellular subsets.

cell-sw	SWB1	SWB2	CELL	all
n	21.5	26.3	31.3	26.5
y	21.3	26.3	30.2	26.0

Table 2: Word error rate on the 3 subsets of Eval01 with and without the cellular switch.

[1] It should noted that results from different tables cannot be directly compared as they were obtained with sligtlty different system configurations (all close to optimal).

4. LANGUAGE MODELING

One of the main challenges in language modeling for conversational speech is the sparseness of the language model training resources. For the BN task it is relatively easy to find a variety of related texts that can be processed and used as training materials. For conversational speech, the only available source is the transcripts of the audio data. In this work, two approaches are investigated to deal with this problem. The first approach is to select "conversational style" texts from other sources, such as BN data, to provide additional training data. The second approach is LM smoothing using a neural network.

Language model construction is as follows: Separate backoff n-gram LMs were estimated on the following audio training corpora transcriptions: 2.7M words of the SWB1 LDC transcriptions, 2.9M words of SWB1 ISIP transcriptions, 230k words of SWB cellular training transcriptions, and 215k words of CallHome corpus transcriptions. An additional backoff LM was built using 240M words of commercially produced BN transcripts.

A single interpolated backoff LM was built from these 5 models using an EM procedure to estimate the interpolations coefficients. The resulting LM has 12M 2-grams, 21M 3-grams and 13M 4-grams. The perplexity on the Eval01 test data is 83.2 (the decompounded perplexity is 60.4). This interpolated model is our baseline SWB 4-gram LM.

Word list selection

The recognition vocabulary contains 41670 words, and is comprised of all words found at least twice in the SWB data and words appearing at least 90 times in BN commercial transcripts. The singletons are not included in the word list as they are mostly foreign words that are unlikely to be observed again or typographical errors. A major difference from BN is that some interjections such as "uh-huh" and "mhm" (meaning yes) and "uh-uh" (meaning no) that were considered to be non-lexical items, need to recognized since they provide feedback in conversations and help maintain contact. As for BN, compound words are used for about 300 frequent word sequences subject to strong reduction. In contrast, acronyms which are frequent in BN are relatively rare in SWB, and are not treated as words. The lexical coverage is 99.7% on Eval01.

Data Selection

One way to cope with the limited amount of in-domain LM training data is to select similar data from domains where much larger corpora are available. The BN training texts include a fair amount of spontaneous speech, although the speaking style can be somewhat different from SWB. Following the work of Iyer and Ostendorf [8], we selected articles from the BN training corpus which are similar in style to the SwitchBoard data, and used the pooled data for MAP adaptation of the baseline LM. Two data selection methods are jointly applied. The first method relies on the posterior probability $\Pr(\text{SWB}|a) = \Pr(a|\text{SWB})/(\Pr(a|\text{SWB}) + \Pr(a|\text{BN}))$ for each BN article a, which is then used to weight the n-gram counts of the selected data (about 30% of the BN corpus). The second selection method relies on spontaneous speech indicators constituted of manually selected words and word pairs specific to spontaneous speech (about 550 words and 2700 word pairs). The frequencies of these words and word pairs in the SWB data are used to estimate a small 2-gram LM. With this oral feature LM, sentences and articles are selected from the BN data ((about 7%) using the perplexity as similarity measure.

Table 3 gives the word error rates on the Eval01 test set for two language models, the baseline SWB 4-gram trained as described above, and a 4-gram trained with the proposed data selection method. Unsupervised MLLR adaptation is used in both cases. The data selection method is seen to reduce the word error rate on each subset with an average absolute reduction of 0.3%,

LM	SWB1	SWB2	CELL	all
4g	21.7	26.8	30.6	26.5
4g select	21.5	26.6	30.2	26.2

Table 3: Word error rate on Eval01 with and without data selection (after MLLR adaptation).

LM Smoothing

The main idea of this connectionist approach is to project the words onto a continuous space, allowing for smooth interpolations. We believe this to be particularly important when only a small amount of LM training material is available. The neural network learns the projections of the discrete word indices onto the continuous space and and estimates the n-gram probabilities (see [15] for details).

Corpus	*ISIP*	*LDC*	*CH*	*CELL*	*interpol. w BN*
backoff	115.7	113.7	189.2	151.2	83.0
neural	106.4	104.9	181.6	150.9	78.8

Table 4: Perplexities of the backoff and the neural 4-gram LM estimated on different transcription sets (SWB ISIP, SWB LDC, CallHome, and SWB Cellular).

Table 4 summarizes the perplexities obtained on the SWB subcorpora with the neural LM estimated on the different sets of acoustic data transcriptions. For comparative purpose, a backoff 4-gram LM was also built on the same data sets using a modified version of Kneser-Ney smoothing using the SRI LM toolkit [16]. The neural LM achieves perplexity improvements of up to 8% relative on all corpora. The perplexity of these LMs interpolated with the backoff 4-gram LM described above is shown in the last column. Although this LM achieves only a small reduction in perplexity on Eval01, 78.8 (58.0 decompounded), the absolute word error is reduced from from 25.3% to 24.9% for the last decoding pass.

5. PRONUNCIATION LEXICON

The pronunciation lexicon has a total of 49648 phone transcriptions for the 41670 words. The pronunciations are based on the same 48 phone set used for BN (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The basic pronunciations are taken from the LIMSI American English lexicon, in which frequent inflected forms have been verified to provide more systematic pronunciations. The pronunciation probabilities are estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing for unobserved pronunciations. We had tried using probabilities for the different pronunciation variants in the BN task, but did not observe any gain. For SWB, the absolute improvement is 0.9% without acoustic model adaptation and 0.4% with acoustic model adaptation.

6. DECODING

The decoding procedure has been substantially modified from that of the BN system. The main changes concern the VTLN warp

	VTLN	MLLR	cell-sw	LM	CN+PP	SWB1	SWB2	CELL	all
Pass 1	n	n	n	3g	n	28.0	36.1	42.2	35.6
Pass 2	y	n	n	3g select	n	25.2	31.8	36.6	31.4
	y	n	n	4g select	n	24.6	31.4	36.2	30.9
	y	n	n	4g select	y	23.3	29.8	33.9	29.1
Pass 3	y	2	n	4g select	y	21.3	26.4	30.7	26.3
Pass 4	y	5	y	4g select + NN	y	20.3	25.3	28.9	24.9

Table 5: Word error rates on Eval01 data for each decoding pass (cell-sw: cellular data switch, CN+PP: confusion network with pronunciation probabilities).

factor estimation, the acoustic model adaptation procedure and the use of lattice rescoring with consensus decoding and pronunciation probabilities. Decoding is carried out in 4 passes. In the first pass the speaker gender is identified for each conversation side using Gaussian mixture models, and a fast 3-gram decode is performed to generate approximate transcriptions. These transcriptions are only used to compute the VTLN warp factors for each conversation side and to identify the type of communication channel (cellular or non-cellular). All of the following passes make use of the VTLN-warped data. Each pass generates a 2-gram word lattice which is then expanded with a 4-gram LM and converted to a confusion network with posterior probabilities.

The posterior probabilities of the lattice edges are estimated using the forward-backward algorithm. Each lattice is converted in a confusion network by iteratively merging lattice vertices and splitting lattices edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [10], but our algorithm appears to be significantly faster for large lattices.

The words with the highest posterior in each confusion set are hypothesized. The resulting transcriptions are used in the next decoding pass for unsupervised MLLR adaptation [9] of the acoustic models. Two regression classes (speech and non speech) are used in the third pass, whereas 5 phonemic regression classes (non speech, voiceless consonants, voiced consonants, and two vowel classes) are used for the 4th pass.

Table 5 shows the word error rates after each decoding pass for the three subsets of the Eval01 test set. The large error reduction between pass 1 and pass 2 is due to the combination of VTLN, the 4-gram LM, the pronunciation probabilities and the confusion network decoding (the contribution of each component is given in Table 5 for this pass). The gain for pass 3 comes from unsupervised acoustic model adaptation. Finally the gain in pass 4 is due to the additional MLLR adaptation with 5 regression classes, the cellular switch and the neural network language model.

7. CONCLUSION

In this paper we have described our work in developing a conversational speech recognizer starting from a state-of-the-art broadcast news transcription system. We found that processing conversational speech requires significant modifications in acoustic modeling, pronunciation and linguistic modeling, as well as in the decoding strategy. The initial word error rate of the BN system on conversational data was around 50%. In order to bring the word error down we had to modify significantly our baseline BN system in addition to using the SWB training data to retrain the models. The following features have been added to our baseline system: vocal tract length normalization, MLLR with multiple phonemic classes, specific acoustic models for cellular data, dictionary with pronunciation probabilities, interpolation of language models with data selection and neural network LM, and confusion network decoding. We also refined our non-speech models. All the improvements lead us to reduce significantly the word error rate to 25%.

REFERENCES

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

[2] P. Dogin, A. El-Jaroudi, J. Billa, "Parameter Optimization for Vocal Tract Length Normalization," ICASSP'00, Istanbul, June 2000.

[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, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP'2000*, **3**:794-798, Beijing, October 2000.

[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), pp. 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, San Francisco, March 1992.

[7] 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*, 1999.

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

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

[10] L. Mangu, E. Brill, A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeeech'99*, 495-498, Budapest, Sep. 1999.

[11] 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, Orlando, May 2002.

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

[13] A. Ljolje et al., "The AT&T LVCSR-2000 System," *Proc. NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[14] T. Hain, P.C. Woodland, G. Evermann, D. Povey, "The CU-HTK March 2000 Hub5e Transcription System," *Proc. NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[15] H. Schwenk and J.L. Gauvain, "Connectionist Language Modeling for LVCSR," *Proc. ICASSP'02*, Orlando, May 2002.

[16] A. Stolcke, "SRILM - An extensible language modeling toolkit." *Proc. ICSLP'02*, **2**:901-904. 2002.

[17] L. Welling, R. Haeb-Umbach, X. Aubert and N. Haberland, "A study on speaker normalisation using vocal tract normalization and speaker adaptive training," ICASSP'98, pp. 797-800, May 1998.