The LIMSI 1997 Hub-4E Transcription System

Jean-Luc Gauvain, Lori Lamel, Gilles Adda

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

ABSTRACT

In this paper we report on the LIMSI system used in the Nov'97 Hub-4E benchmark test on transcription of American English broadcast news shows. There are two main differences from the LIMSI system developed for the Nov'96 evaluation. The first concerns the preprocessing stages for partitioning the data, and the second concerns a reduction in the number of acoustic model sets used to deal with the various acoustic signal characteristics.

The LIMSI system for the November 1997 Hub-4E evaluation is a continuous mixture density, tied-state cross-word context-dependent HMM system. The acoustic models were trained on the 1995 and 1996 official Hub-4E training data containing about 80 hours of transcribed speech material. The 65K word trigram language models are trained on 155 million words of newspaper texts and 132 million words of broadcast news transcriptions. The test data is segmented and labeled using Gaussian mixture models, and non-speech segments are rejected. The speech segments are classified as telephone or wide-band, and according to gender. Decoding is carried out in three passes, with a final pass incorporating cluster-based test-set MLLR adaptation. The overall word transcription error of the Nov'97 unpartitioned evaluation test data was 18.5%.

INTRODUCTION

The goal of the DARPA Hub-4 task is to transcribe radio and television news broadcasts[1, 2]. These shows contain a wide variety of segment types, of different acoustic and linguistic natures. The signal quality is quite variable with segments of clean, studio data, telephone data, as well as speech in the presence of background noise or music and pure music segments. The speech is of various linguistic styles (ranging from prepared to spontaneous speech), produced by various types of speakers (news anchors, talk show hosts, reporters, politicians, everyday people, etc.).

Our development work for the Nov'97 evaluation addressed mainly the problem of partitioning the continuous stream of acoustic data and improving as well as simplifying the acoustic models. The first step is to partition the data into relevant segment types. A speech/non-speech decision is made using Gaussian mixture models (GMMs), and the speech segments are then clustered using an agglomerative clustering algorithm and classified according to data type. Long segments were subsequently chopped into chunks of at most 30s using the chopping algorithm developed last year[6]. In our 1996 Hub-4 system, the wide variety of acoustic data was addressed by training specific acoustic models for the different acoustic conditions. This resulted in 5 different acoustic models sets to account for the different acoustic data types, whereas this year only 2 acoustic model sets are used.

In the remainder of this paper we provide an overview of the LIMSI Nov'97 Hub-4E system, summarizing some of our development work in preparation for the Nov'97 Hub-4E evaluation, and differences with our Nov'96 system.

SYSTEM OVERVIEW

Our 1997 Hub-4E system uses the same basic technology as used in previous evaluations, that is continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling[4]. For acoustic modeling, 39 cepstral parameters are derived from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10ms. Each phone model is a tied-state left-to-right CDHMM with Gaussian mixture observation densities (about 32 components). The modeled triphone contexts are selected based on their frequencies in the training data, with backoff to right-context, left-context, and context-independent phone models.

Prior to word recognition, the data is partitioned into a set of speech segments which are labeled as wide or telephone bandwidth, and according to gender. Non-speech segments are discarded. Word recognition is carried out in three passes for each speech segment. In the first pass a word graph is generated using a bigram language model. The second decoding pass uses the word graph generated by the 1st pass and a trigram language model. The final decoding pass is carried out using adapted acoustic models. As done in our 1996 Hub-4 system, some of the observed characteristics of the broadcast data were modeled by using specific phone and word models for filler words and breath noise. Compound words were used as a means of modeling reduced pronunciations for common word sequences and acronyms.

TRAINING DATA

This year about 80 hours of transcribed task-specific training data were available. These data were obtained from the following shows: ABC Nightline, ABC World News Now, ABC World News Tonight, CNN Early Edition, CNN Early Prime, CNN Headline News, CNN Prime News, CNN The World Today, CSPAN Washington Journal, NPR All Things Considered, and NPR MarketPlace. For the portion of the training data already used for our Nov'96 system, we essentially kept the LDC August'96 release of the transcriptions, as a variety of manual modifications had been carried out on these. For the remaining training data we used primarily the August'97 transcriptions, with some use of the Feb'97 transcription release from LDC.

This data was used to train the Gaussian mixture models needed for segmentation and the acoustic models for use in word recognition. In contrast to our 1996 system where some read-speech corpora were used, no additional acoustic training data were used.

The language model training data consisted of the 1995 Hub-3 and Hub-4 LM material (155M words), and on the broadcast news transcriptions from 1992 to 1996 (125M words). The transcriptions of the acoustic training data (866K words) were included 10 times.

DATA PARTITIONING

The segmentation and labeling procedure is as follows. First, the non-speech segments are detected (and rejected) using Gaussian mixture models (GMMs). Three GMMs each with 64 Gaussians serve to detect speech, pure-music and other. The acoustic feature vector used for segmentation contains 38 parameters. It is the same as the recognition feature vector except that it does not include the energy, although the delta energy parameters are included. The three GMMs were each trained on about 1h of acoustic data, extracted from the training data after segmentation with the transcriptions. The speech model was trained on data of all types, excluding pure music segments and the silence portions of speech over music segments. These models are expected to match all speech segments. The music model was trained on portions of the data that were labeled as pure music, so as to avoid mistakenly detecting speech over music segments. The silence model was trained on the segments labeled as silence during forced alignment, after excluding the segments labeled as containing speech in the presence of background music.

The segmentation algorithm generally works well, but we observed that a long, noisy speech segment in the Nov'96 evaluation data was rejected, being mistakenly labeled as music. In an attempt to avoid this kind of error, we also experimented with using a 4th GMM for speech in the presence of background music, but there were only small differences in the rejected segments. All segments labeled as music or silence were removed prior to further processing.

A segmentation/clustering process is then carried out on the speech segments using GMMs and an agglomerative clustering algorithm. 10 iterations of the GMM reestimation/clustering algorithm are run. The process is initialized using a simple segmentation algorithm based on the detection of spectral change (similar to the first step used in the Nov'96 CMU system[13]). The threshold is set so as to over-generate segments. Initially, the cluster set consists of a cluster for each segment. This is followed by Viterbi training of the set of GMMs (one 8-component GMM per cluster). The clustering technique is a bottom-up agglomerative one, where each cluster is characterized by its GMM distribution. When applied, this clustering procedure is embedded in the Viterbi training process by replacing the segmentation step in alternate iterations. The procedure is controlled by 3 parameters: the minimum cluster size, the GMM distance measure threshold, and the segment boundary penalty.

The data partition is obtained by Viterbi decoding in the last iteration. The segment boundaries are then refined using the last set of GMMs with an additional relative energy-based boundary penalty, within a 1s interval. This is done to locate the segment boundaries at silence portions, so as to avoid cutting words in two.

Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to label telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one each for wideband and telephone band).

The result of the partitioning process is a set of speech segments with cluster labels (including gender and telephone labels).

ACOUSTIC MODELING

The speech analysis in the LIMSI Nov'97 system[6, 7] results in 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficents and the log energy, along with the first and second order derivatives.

Various approaches were investigated to build acoustic models from the available WSJ-si355 and Hub-4E training data. Early experiments showed no clear gain from using the WSJ data to initialize the acoustic models, so most of the development work was carried out only with the Hub-4 data.

Last year 5 different model types were built to deal with the varied acoustic conditions found in the Hub-4 data. This year only wideband and telephone band models are used. As was done last year, specific phone symbols are used to explicitly model filler words and breath noises.

Two sets of gender-dependent acoustic models were built using MAP adaptation of SI seed models for each of wideband and telephone band speech[9]. For computational reasons, a smaller set of acoustic models is used in the bigram pass to generate a word graph. These position-dependent, cross-word triphone models cover 3521 contexts, with 8471 tied states and 32 Gaussians per state. For trigram decoding a larger set of 8900 position-dependent, cross-word triphone models with 11500 tied states was used.

The use of position-dependent models in the trigram decoding is new this year. In the past, position-dependent acoustic models were used in the first decoding pass in order to reduce the search space and the decoding time, even though slightly better performance was obtained with position-independent models. This year a slight gain was observed on the development data with position-dependent models.

LANGUAGE MODELING

The language models were trained on newspaper texts (the 1995 Hub-3 and Hub-4 LM material – 155M words), on the broadcast news (BN) transcriptions (years 92-96, 125M words), and the 866K words in the transcriptions of the 1995-1996 acoustic training data LDC releases of Aug'96 and Aug'97. The 1995 Hub-3 and Hub-4 LM training texts were processed as was done previously to clean errors inherent in the texts or arising from the preprocessing tools. They were also transformed to be closer to the observed American reading style[5]. Any attempts to use newspaper texts distributed by LDC from a more recent period (up to the limit of June 1996, about 153M words) led to a negligable decrease in perplexity and an increase in the recognition word error rate.

The 65K word LMs were built using the CMU-Cambridge Statistical Language Modeling Toolkit[3]. The trigrams from the BN training transcriptions (820K words) as well as those in 1995 Marketplace transcriptions (46K words) were added in the LM with a weight of 10. Using this toolkit, we have chosen the Witten-Bell discounting strategy because this method led to systematic (but quite small, less than 1% relative) improvements as compared to the classical Good-Turing strategy.

All distinct words in the transcriptions (25167 from the BN training and 6451 from 1995 MarketPlace) were included in the recognition vocabulary. The vocabulary selection and language models were optimized on the 1996 Hub-4 F0 and F1 evaluation test set. The OOV rate is 0.66% on the 1996 Hub-4 dev test data and 0.97% on F0-F1 part of the Nov'96 eval test set.

As was done last year, the broadcast news training texts were processed in order to be homogeneous with the previous texts, and filler words such as UH and UHM, were mapped to a unique form. All of the training texts (1995 Hub-3 and Hub-4, and BN) were reprocessed in order to add a proportion of breath markers ({breath}) (4%), and of filler words ({fw}) (0.5%)[6]. We also tried two techniques to add n-grams with {breath} and {fw}: in the first one, we constructed a language model after reprocessing the texts to include {breath} and {fw} (i.e. {breath} and {fw} are considered to be normal words), or we just added all the trigrams containing {breath} and {fw} to a trigram constructed on a non-reprocesssed text (similar to what we do for cross sentence trigrams). Although the second method resulted in a perplexity decrease of about 5%, better recognition results were obtained with the first method.

Two strategies were explored to add cross sentence trigram counts in the trigram model[12]: add the whole texts with and without sentences boundaries, and renormalize the counts; or add only the cross sentence trigrams. Both strategies led to similar results in terms of perplexity and recognition error. For the evaluation, the language models were constructed using the second approach.

LEXICAL MODELING

The recognition vocabulary contains 65,252 words and 72,788 phone transcriptions. This year the lexicon was extended to include about 2900 new words found in the 1997 February and August transcription releases. Many of the new words were proper names, whose pronunciations were verified by listening to the training data. Pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). The filler and breath phones were added to model these effects, which are relatively frequent in the broadcast data and are not used in transcribing other lexical entries. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. Frequently occuring inflected forms were verified to provide more systematic pronunciations.

As in last year's system, the lexicon contains the most common 1000 acronyms found in the training texts[8], and compound words to represent frequent word sequences[6]. This provides an easy way to allow for reduced pronunciations such as /lεmi/ for "let me" and /g$\wedge$nx/ for "going to" or a syllabic-n for the word "and" in "AT&T"

WORD DECODING

Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to limit the memory required for the trigram decoding pass. The chopping algorithm is the same as was used last year[6]. A bimodal distribution is estimated by fitting a mixture of 2 Gaussians to the log-RMS power for all frames of the segment. This distribution is used to determine locations which are likely to correspond to pauses, thus being reasonable places to cut the segment. Cuts are made at the most probable pause 15s to 30s from the previous cut.

Word recognition is performed in three steps: 1) word graph generation, 2) trigram pass, 3) cluster-based acoustic model adaptation. The word graph is generated using a 65K word bigram backoff language model. This step uses a gender-specific sets of position-dependent triphones with about 8500 tied states and a small bigram language model (about 2M bigrams). Differents acoustic models are used for telephone and wideband segments. The sentence is then decoded using the word graph generated in the first step with a large set of gender-dependent acoustic models (position-dependent triphones with about 11500 tied states) and a 65K word trigram language model (including 8M bigrams and 16M trigrams). Finally, unsupervised acoustic model adaptation (both means and variances) is performed for each cluster using the MLLR technique[11], prior to the last decoding pass with the adapted models and the trigram LM. The mean vectors are adaptated using a single block-diagonal regression matrix, and a diagonal matrix is used to adapt the variances.

EXPERIMENTAL RESULTS

For development data we used the Nov'96 development and evaluation data sets. Development was carried out using the partitioned evaluation segments, focusing mostly on the F0 (prepared) and the F1 (spontaneous) data types.

In the Nov'96 evaluation, LIMSI reported results only for the partitioned evaluation (PE) component, with an overall word error rate is 27.1%. The PE condition assumes that both the segment boundaries and the data type (F0-Fx) are known, but automatically making some of the distinctions is not so evident. We therefore explored two alternatives to see the importance of such prior information: the first uses only the segment boundaries but not the data type classification; and the second uses no prior information (unpartitioned condition).

In [7] we showed that using only two model sets (for wideband and telephone band, selected automatically) instead of 5 type-specific model sets resulted in a only slight increase in word error (about 1% relative in the second decoding pass with a trigram language model, making use of the word graphs generated with the first pass type-specific acoustic model sets). This indicated that given the segment boundaries, a priori data type classification does not critically affect overall performance. Similar results were reported by BBN last year[10].

We explored the no prior knowledge (unpartitioned) condition using the Marketplace show from the 1996 development data. A two-way classification was used dividing the data into wideband and bandlimited segments. The telephone segments in the show were correctly detected, with boundary locations close to those marked manually. All segments longer than 30s were subsequently chopped into chunks, and each chunk was processed independently. Decoding the telephone speech segments with the telephone speech models and all the other segments with the wideband models, resulted in a word error of 18.7%. This error rate can be compared with the 16.7% word error rate obtained when the segment boundaries, but not the data type, are known. Thus, a relatively small degradation is observed with a substantially simpler system.

The data partitioning procedure used for the Nov'97 evaluation was not optimized by minimizing the word error rate. It was subjectively evaluated by comparing the resulting partitions manual segmentations of various data sets. Only one complete run was carried out on the Nov'96 UE data (after the Nov'97 evaluation).

Concerning the acoustic models, this year there was double the amount of transcribed training data available compared to last year. We evaluated different acoustic model sets, training only on Hub-4 data and Hub-4 combined with other read-speech corpora (mainly WSJ). The observed differences on the F0/F1 portion of the development data were quite small (a few percent relative). The best results were obtained by training on only the 80 hours of Hub-4E acoustic data, with a relative error reduction of 7% on the Nov'96 eval data compared to last year's acoustic models which were trained on the WSJ corpus and the 40 hours of Hub-4 data.

The combined effect of the new language models and modifications to the lexicon gave an additional small (2% relative) error reduction.

Table 1 compares the results on the eval96 and eval97 data sets. The high deletion rate on the eval96 data is mainly due to 2 very noisy speech segments which were classified as non-speech. (This type of error was less frequent on the eval97 data which was of higher quality on average.) However since the word error is very high on these segments, rejecting them has only a marginal effect on the overall word error rate. The result is a higher deletion rate and a lower substitution one.

SUMMARY

In this paper we have presented the LIMSI 1997 Hub-4E system and some of our development work in preparation for the evaluation. The system architecture is very similar

Test set	*Corr*	*Sub*	*Del*	*Ins*	*Err*
eval96	77.8	15.4	6.9	3.1	25.3
*eval97**	84.1	12.4	3.5	2.5	18.5

Table 1: Word error rates for of unpartitioned evaluation on 1996 and 1997 eval test data. (* Official NIST score).

to that of our Nov'96 Hub-4 PE system, the main difference being the addition of an algorithm to partition the data. Keeping the same architecture, our development effort focused on acoustic modeling, making use of the additional transcribed training data, and improving the lexicon and language models.

The data partitioning algorithm makes use of Gaussian mixture models and an iterative segmentation and clustering procedure. The resulting segments are labeled using 64-component GMMs as pure music, wideband (male/female), and narrowband (male/female). Cepstral mean normalization is carried out on a cluster basis, instead of on individual segments. Similarly, cluster-based unsupervised adaptation of the means and variance is used.

Concerning acoustic modeling, last year we found that with 40 hours of transcribed broadcast news data, better models could be built by also using WSJ training data. This year, with 80 hours of Hub-4 data available, we reconsidered the need for the read-speech data and observed that acoustic models trained only on the Hub-4 data outperformed last year's models.

Experiments comparing focus type specific models with more general task-specific models resulted in only a slight loss in performance. This performance difference does not justify the additional burden in training and decoding with specialized model sets, as reported in [10, 7]. As a consequence, this year's system had 2 gender-dependent sets of acoustic models, compared to 5 last year. Concerning the acoustic models we observed that word position dependent contexts not only reduced computational requirements, but also gave a small improvement in performance. The word error on the Nov'97 Hub-4E data was 18.5%.

REFERENCES

[1] *Proc. DARPA Speech Recognition Workshop*, Arden House, February 1996.

[2] *Proc. DARPA Speech Recognition Workshop*, Chantilly, February 1997.

[3] P. Clarkson, R. Rosenfeld, "Statistical Language modelling using CMU-Cambridge Toolkit," *Proc. EuroSpeech'97*, Rhodes, Greece, pp. 2707-2710 September 1997.

[4] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, October 1994.

[5] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Large Vocabulary Dictation: The LIMSI Nov94 NAB System," *Proc. ARPA Spoken Language Technology Workshop*, Austin, Texas, pp. 131-138, January 1995.

[6] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 56-63, February 1997.

[7] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcription of Broadcast News," *Proc. EuroSpeech'97*, Rhodes, Greece, pp. 907-910, September 1997.

[8] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "The LIMSI 1995 Hub3 System," *Proc. DARPA Speech Recognition Workshop*, Arden House, pp. 105-111, February 1996.

[9] 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.

[10] R. Schwartz, H. Jin, F. Kubala, S. Matsoukas, "Modeling Those F-Conditions – Or Not," *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 115-118, February 1997.

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

[12] K. Seymore, S. Chen, M. Eskenazi, R. Rosenfeld, "Language and Pronunciation Modeling in the CMU 1996 Hub4 Evaluation. *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 141-146, February 1997.

[13] M. Siegler, U. Jain, B. Raj, R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 97-99, February 1997.