Transcribing Broadcast News: The LIMSI Nov96 Hub4 System

J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker

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

ABSTRACT

In this paper we report on the LIMSI Nov96 Hub4 system for transcription of broadcast news shows. We describe the development work in moving from laboratory read speech data to real-world speech data in order to build a system for the ARPA Nov96 evaluation. Two main problems were addressed to deal with the continuous flow of inhomogenous data. These concern the varied acoustic nature of the signal (signal quality, environmental and transmission noise, music) and different linguistic styles (prepared and spontaneous speech on a wide range of topics, spoken by a large variety of speakers).

The speech recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora. The base acoustic models were trained on the WSJ0/WSJ1 corpus, and adapted using MAP estimation with 35 hours of transcribed task-specific training data. The 65k language models are trained on 160 million words of newspaper texts and 132 million words of broadcast news transcriptions. The problem of segmenting the continuous stream of data was investigated using 10 MarketPlace shows. The overall word transcription error of the Nov96 partitioned evaluation test data was 27.1%.

INTRODUCTION

The goal of the ARPA Hub4 task is to transcribe radio and television news broadcasts. The shows contain signal segments of various acoustic and linguistic nature, with abrupt or gradual transitions between segments. The signal may be of studio quality or have been transmitted over a telephone or other noisy channel (ie., corrupted by additive noise and nonlinear distorsions), as well as speech over music and pure music segments. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc. The linguistic style ranges from prepared speech to spontaneous speech. Acoustic models trained on clean, read speech, such as the WSJ corpus, are clearly inadequate to process such inhomogeneous data.

Our development work aimed at addressing the two principle types of problems encountered in transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training specific acoustic models for the different acoustic conditions. The first step in dealing with the inhomogeneous data was to develop a segment classifier, so as to divide the data into the main different segment types. The segment classifier was developed and evaluated using MarketPlace data. Even though the evaluation was carried out using partitioned data, the segment classifier was used to detect unlabeled bandlimited speech. In the partitioned evaluation, the focus conditions correspond to different speaking styles (prepared or spontaneous speech) and to different acoustic environments (high quality, degraded acoustic conditions, and speech over music). In contrast to previous evaluations using read-speech data where the longest sentences were on the order of 30s, the partitioned segments can be several minutes long. Therefore a chopping algorithm was developed so as to limit to 30s the amount of data to be processed as a single unit.

In order to address variability observed in the linguistic properties, we analyzed differences in read and spontaneous speech, with regard to lexical items, word and word sequence pronunciations, and the frequencies and distribution of hesitations, filler words, and respiration noises. As a result of the analysis, the phone set was enlarged to explicitly model filler words and breath noise, resulting in specific context-dependent acoustic models. These phenomena were also explicitly represented in the language model. Compound words were introduced as a means of modeling reduced pronunciations for common word sequences.

Our 1996 Hub4 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. Acoustic modeling uses cepstral parameters 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 were selected based on their frequencies in the training data, with backoff to right-context, left-context, and context-independent phone models. Word recognition is carried out in two passes for each speech segment. In the first pass a word graph is generated using a bigram language model and in the second pass decoding uses the word graph generated by the 1st pass and a trigram language model.

In the remainder of this paper we provide an overview of the development work carried out in preparation for the Nov96 Hub4 evaluation. The initial word error of 39.2% obtained using our Nov95 Hub3 65k word recognizer was reduced to 25.2% on the Nov96 development data.

DEVELOPMENT WITH MARKETPLACE

Our 65k word recognizer developed for the Nov95 ARPA NAB evaluation [6, 5] was used to recognize a MarketPlace radio show taken from the Nov95 Hub4 sample training data distributed by NIST[1]. The wideband acoustic models were trained on the WSJ0/1-si355 training data containing a total of 46k sentences[6], comprised of 37k sentences from the WSJ si284 corpus, 130 sentences/speaker from 57 long-term and journalist speakers in WSJ0/1, and 1218 sentences from 14 of the 17 additional WSJ0 speakers not included in si284. Only the data from the close-talking, Sennheiser HMD-410 microphone was used. For telephone speech models, we used telephone channel models developed for the Hub2 test in 1994[3]. These models were trained on a bandlimited version of the WSJ si284 corpus, and adapted using MAP estimation[7] with 7k WSJ sentences of telephone speech data taken primarily from the Macrophone corpus. No task-specific acoustic training data was used. For language modeling data, we used newspaper texts and read speech transcriptions predating July 30, 1995. This data includes the August'94 release of the CSR standard LM training texts distributed by LDC (years 88-94), the 1994 NAB development data (excluding the devtest data), the WSJ0/WSJ1 read speech transcriptions (85,343 sentences), and the 1994 and 1995 financial domain material (Hub3 LM material).

A segmentation algorithm was developed using nine half-hour MarketPlace shows as task-specific training data (1 show was kept aside to test the segmenter). A small left-to-right tied mixture HMM with 64 Gaussians was built for each of the following signal types: background noise, pure music, speech on music, wide-band speech, and telephone speech. The models were trained using the segmentations and labels provided by BBN[8]. Viterbi decoding on the 5 models (fully connected) is used to segment the data and assign each speech frame to one of the 5 classes.

A show is transcribed as follows: First the show is segmented using the tied mixture models. Segments identified as background noise and pure music are discarded. The telephone speech segments are then decoded with the telephone speech models and all the other segments are decoded using the wideband models. Unsupervised MLLR adaptation [9] is performed using all the data of a given type in the current show. Since sentence boundaries are not known, each segment is decoded as a single unit.

	Identified class			
Test data	S	T	MS	M
Wide-band speech (S)	99.9	0.0	0.0	0.0
Telephone speech (T)	1.2	98.8	0.0	0.0
Music+speech (MS)	32.0	0.0	66.4	1.6
Music (M)	7.5	0.0	1.7	90.8

Table 1: Segmentation results in terms of the percentage of frames correctly and incorrectly classified for each class of data.

The spectrograms in Figure 1 show examples of the broadcast news data along with the reference (manual) and automatically determined segmentations. The top spectrogram shows three segments, a sequence of music (M), music+speech (MS), followed by music. The boundaries delimiting the speech are somewhat difficult to locate. The most difficult of these boundaries occur where music is fading in or out. The lower spectrogram shows a portion of wideband speech surrounded by telephone speech (T). The bandlimiting is clearly visible and easily detected by the system.

The segmentation error at the 10 ms frame level on the complete MarketPlace show kept aside for development was 6%. As can be seen in Table 1 most of the segmentation errors are due to the misclassification of the music+speech frames (32.0% are classified as speech) and the music frames (7.2% are classified as speech). Music+speech frames are often classified as speech when the music is fading out because the signal is not very different from a speech signal with slight background noise. In this show there were no segments labeled as noise (N) by the transcribers, and no noise segments were detected by the segmenter.

The overall word error rate of the transcription for the same MarketPlace show is 24.6%. The error rate is seen to be much lower on wideband speech (16.2%), and much higher on telephone speech (42.6%) and music+speech (37.1%). The higher error rate observed for the telephone speech is not only due to the channel (reduced bandwidth and possible distortions), but also to the fact that most of this speech is spontaneous in nature, whereas much of the wideband speech is prepared. Also contributing to the overall error rate are insertions due to words recognized in a few music segments which are erroneously labeled as music+speech.

DEVELOPMENT WITH BROADCAST NEWS

For the Nov96 evaluation, the scope of the task was enlarged to include multiple sources of broadcast news (ra-

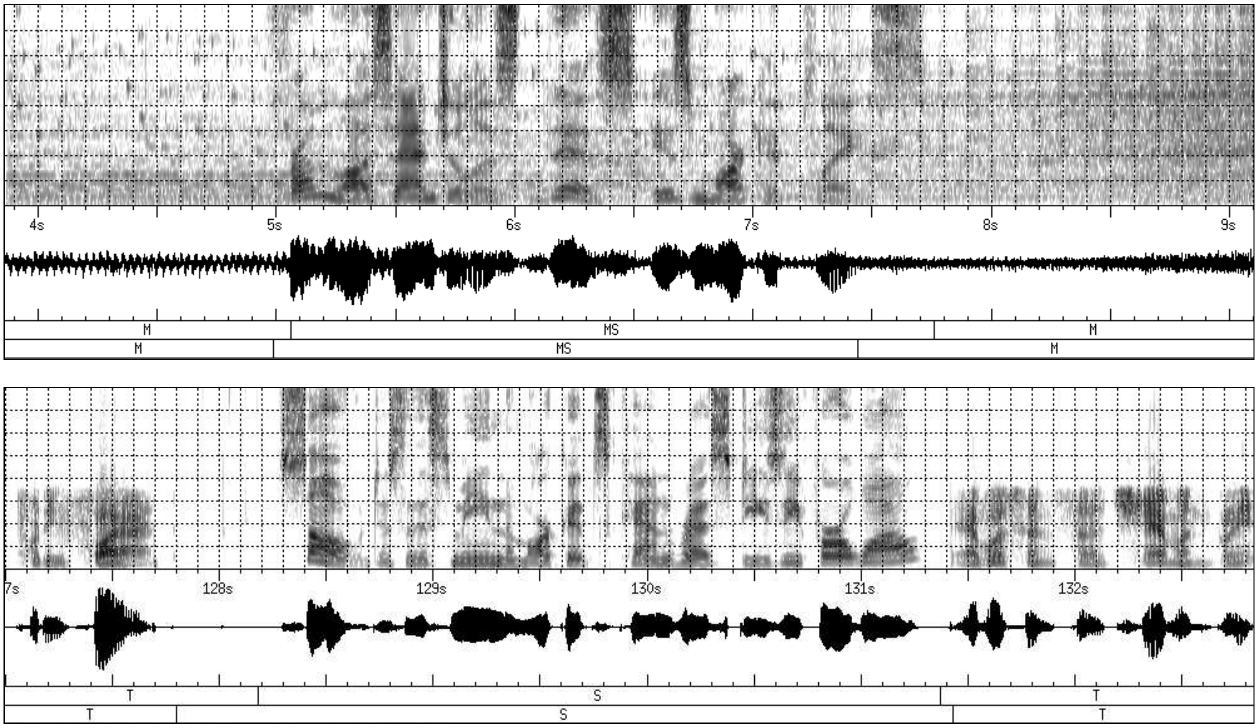


Figure 1: Spectrograms illustrating segmentations of sequences extracted from a MarketPlace radio broadcast. The upper transcript is the reference, and the lower is the result of automatic segmentation. The labels are: S (wideband speech), T (telephone speech), MS (music+speech), and M (music).

dio, TV) and different types of shows (such as CNN Headline News, NPR All Things Considered, ABC Prime Time News). The test data included episodes of shows not appearing in the training material. The 1996 evaluation consisted of two components, "partitioned evaluation" component (PE) and the "unpartitioned evaluation" component (UE). All sites were required to evaluate on the PE, which contains the same material as in the UE, but has been manually segmented into homogeneous regions, so as to control for the following *focus conditions*[11]:

F0- Baseline broadcast speech
F1- Spontaneous broadcast speech
F2- Speech over telephone channels
F3- Speech in the presence of background music
F4- Speech under degraded acoustical conditions
F5- Speech from non-native speakers
Fx- All other combinations

About 35 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.

The development data were taken from 6 shows: ABC Prime Time, CNN World View, CSPAN Washington Journal, NPR MarketPlace, NPR Morning Edition, and NPR The World.

Using our Nov95 Hub3 65k word recognizer, an initial word error 39.2% was obtained on the Nov96 development data. The available acoustic and language model training data were used to generate a new vocabulary list and language models, to extend the pronunciation lexicon, and to train type-specific acoustic models for the different acoustic data types. With the final setup used for the evaluation, a word error of 25.2% was obtained on the same development test set. In the remainder of this section, we describe our 1996 Hub4 system.

Acoustic features

The speech analysis is relatively standard, but differs in a few points from what we have used in previous evaluations[6]. A 30ms analysis window is used with a 10ms frame step. For each 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 basis using cepstral mean removal and variance normalisation. Thus each cepstral coefficient for each segment 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. This feature vector has fewer parameters than the 48-component feature vector used previously, but has better performance on the Hub4 data (3% relative gain).

Acoustic models

Different acoustic model sets were trained to address different aspects of the problem, such as segmentation, sex-identification, and word decoding. Gaussian mixture models (64-components) similar to those used for segmentation as described for the MarketPlace data, were used to separate telephone and wideband speech. For each segment, type-specific Gaussian mixture models were used to identify the sex of the speaker. For word decoding, type-specific acoustic model sets, similar to last year's Hub3 models[5] were used.

Various approaches were investigated to build acoustic models from the available WSJ-si355 and Hub4 training data. The most effective solution for our system was the following:

1. Train large sets of gender-dependent tied-state models on the secondary channel of the WSJ0/1-si355 data. The resulting acoustic model sets, M0, contained 7000 mixture distributions. These models were not used for the evaluation.
2. Use MAP estimation techniques to adapt the M0 seed models to the Hub4 1994 and 1995 training data, providing the baseline Hub4 model sets M1 (0-8kHz band) for the F0 and F1 data, and M2 (bandlimited to 0-3.5kHz) for use with the F2 data.
3. For the F3 and F4 conditions, adapt the M1 models using phone-based (one full regression matrix per phone) supervised MLLR and the F3 and F4 parts of the training data, resulting in models M3 and M4, respectively.
4. For the F5 data (non-native speakers), adapt the M0 models to British English data (WSJ0CAM)[10] prior to adaptation with the Hub4 training data to create the model set M5.
5. Unsupervised MLLR adaptation is carried out for each test segment prior to the final decoding pass.

The M1 models were used to process the F0 and F1 segments. The M2 models were used to process the F2 segments, as well as all other segments labeled as telephone speech by the Gaussian classifier. The M3, M4, and M5 models were used to process the F3, F4, and F5 data respectively. The model set to process the Fx segments was selected as follows:

```
if (telephone-data) use M2 models
else if (non-native-speaker) use M5 models
    else if (background-noise) use M4 models
        else use M1 models
```

where the telephone decision was based on the output of the Gaussian segment classifier, and all other attributes were taken from the provided segment annotation.

The different model types described above aim to deal with the varied acoustic conditions found in the Hub4 data. In order to better model the observed speaking styles, 2 new phone symbols were added to the existing phone set to explicitly model filler words and breath noises. These new phones were only trained with the Hub4 acoustic data since they are infrequent in the WSJ read-speech data.

For computational reasons, a smaller set of acoustic models was used in first bigram pass used to generate a word graph. These position-dependent, cross-word triphone models cover about 3500 contexts, with 6000 tied states and 32 Gaussians per state. For trigram decoding a larger set of 5300 position-independent, cross-word triphone models with 7000 tied states was used. The modeled triphone contexts were selected based on their frequencies in the WSJ training data. For the breath noise and filler word specific phones, the contexts were selected according to their observed frequencies in the Hub4 training data. In total there were 20 model sets: 5 conditions × 2 genders × 2 decoding passes.

Language models

The language models were trained on newspaper texts (the 1995 Hub3 and Hub4 LM material – 161M words), on the broadcast news (BN) transcriptions (years 92-96, 132 M words), and the 430 K words in the transcriptions of the 1995-1996 acoustic training data.

The 1995 Hub3 and Hub4 LM training texts were reprocessed 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[4].

The BN training texts were cleaned in an analogous manner to the previous text materials. However, since in the BN texts word fragments are represented with a "hyphen", compound words were not split in the version distributed by LDC. We retreated all the transcriptions in order to split hyphenated words, as the occurrence of word fragments was marginal compared to other situations where the hyphen needed to be treated.

The 65k recognition vocabulary included all words occurring in the transcriptions (17883 from the 1996 BN transcripts and 6332 from 1995 MarketPlace), completed with the most common words found in the texts. The LMs and vocabulary selection were optimized on the 1996 Hub4 developement test set. The resulting lexical coverage on the 1996 Hub4 devtest data is 99.34%.

We experimented with different weighting factors for the available text materials and transcripts. The perplexities as a function of type of data are given in Table 2 comparing

	Hub+BN	*Hub+BN+3×trn*	*Hub+BN+10×trn*
F0	199	193	190
F1	150	148	147
F2	140	139	138
F3	228	227	221
F4	261	263	262
F5	181	177	177
Fx	175	173	173
Overall	**180**	**177**	**175**

Table 2: Perplexity with a trigram LM as a function of the weighting factor applied to the acoustic training transcriptions.

	BN+10×trn	*Hub+BN+10×trn*
F0	351	**320**
F1	**239**	255
F2	**221**	228
F3	380	**356**
F4	473	**409**
F5	325	**304**
Fx	296	297
Overall	302	295

Table 3: Perplexities of the bigram LMs with compound words.

weighting factors of 0, 3 and 10 for the acoustic training transcripts. A weight of 10 ensured that all trigrams occurring in the transcriptions were included in the LM. As shown in Table 2, weighting the training transcripts by 10 gave a slight, yet consistent improvement in perplexity, and also led to a relative increase in word accuracy on the F0 devtest data of 2%. The addition of other newspaper texts from any date led to a degradation both in terms of perplexity on the Hub4 devtest texts and recognition accuracy.

The 1996 training transcripts were processed to map filler words (such as UH, UM, UHM) to a unique form {FW}, and the frequencies of filler words and breath noises were estimated for the different types of segments. These estimates were used in reprocessing the text materials. For breath noises, the observed proportion is different for the different segments (about 4.5% for the F0 and F1 segments, but only about 3% in the F3 and F4 segments). We hypothesized that the lower proportion in the F3 and F4 segments was an artifact due to the background music and noise which may have masked the breath noises. We also observed that while most breath noises appear at phrase boundaries, they also occur at other locations. We thus decided to process all of the training texts (1995 Hub3 and Hub4 and BN training texts) adding a fixed proportion of breath (4%), mostly near punctuation markers, respecting a minimum and maximum distance between two breath markers. A larger difference across segment types was observed for filler words, from 0.25% in prepared speech to about 3% in spontaneous speech. However, even though the global proportions were different, the filler words tend to occur in similar contexts for the different segment types.
After systematic examination of their relative proportions in the training transcriptions, we constructed a "degrading" filter which adds filler words in the text with a parametrizable global proportion, so that the relative proportion of the fillers near specific common words was similar to that observed in the training transcription.

The resulting language models were tested using perplexity and recognition word error. Construction of different LMs for prepared and spontaneous speech according to the proportion of fillers found in the transcriptions, led to a gain in terms of perplexity, but did not reduce the recognition word error. We found that adding a small proportion of filler words (0.5%) improved the recognition accuracy, but adding a large proportion (3-5%) reduced performance.

As was done last year, the training texts were processed to treat the 1000 most frequent acronyms as whole words instead of as sequences of independent letter. This year we also added 300 compound words for common word sequences.

We split the different segments into 2 homogeneous groups from the LM point of view: one group corresponding to prepared speech with F0, F3, F4, F5 segments, and the other to spontaneous speech with F1, F2 segments. For the 1st bigram decoding pass, different LMs were used for prepared speech (cut-off 8, 2M bigrams) and spontaneous speech (cutoff 3, 1.9M bigrams). In the latter case the newspaper training texts were not used. The bigram perplexities for these two language models are given in Table 3 for the different data types. For the spontaneous speech data (F1 and F2), a lower perplexity is obtained when the LM is estimated on only the Broadcast News transcriptions. Using this LM also gave a relative word error reduction of 2% on the spontaneous speech portions of the development data. For the prepared speech a lower perplexity is obtained when the newspaper texts are included in the training material.

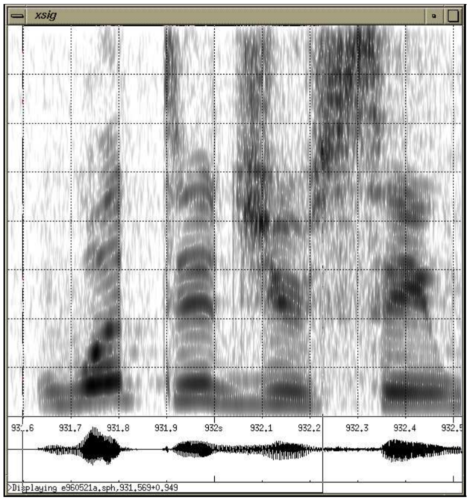
For the 2nd pass, while the use of different trigrams for prepared and spontaneous speech LMs led to a gain in terms of perplexity, the word accuracy was worse on the development data. We therefore used a single 65k trigram LM trained on all the texts mentioned above (cut-off 1-2, 7.6M bigrams and 13.4M trigrams).

Recognition Lexicon

The 65k vocabulary contains 64,968 words and 72,488 phone transcriptions. 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 emissions, and are not used in transcribing other lexical entries. The training and test lexicons were created at LIMSI and include some input and/or derivations from the TIMIT, Pocket and Moby lexicons. A pronunciation

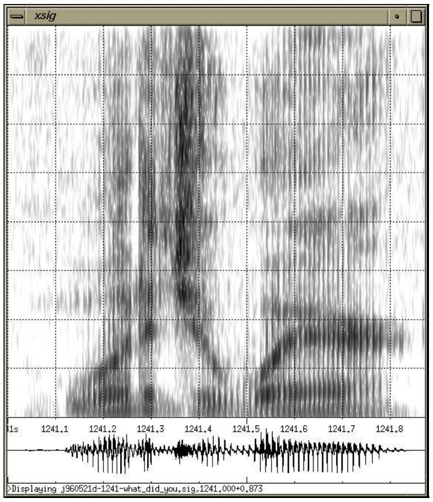
WHAT_DID_YOU	wa{t}dIdyu
	wa{t}dIdyx wa{t}dIJx w[ax]Jx
I_DON'T_KNOW	Ydon{t}no
	Yd∧no Ydno
DON'T_KNOW	don{t}no
	d∧no
LET_ME	lEtmi
	lEmi
LET_HIM	lEthIm
	lEtM lEm
I_AM	Y@m
	Yxm Ym
GOING_TO	go|Gt[ux]
	g[∧c]nx

Figure 2: Some example compound words and their pronunciations. Original concatenated pronunciation (1st line) and reduced forms (2nd line).



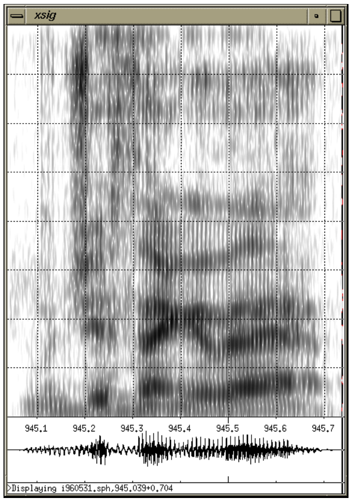
>Displaying e960521a.sph,931.569+0.949

Figure 3: Spectrogram of the word sequence "what did you see" (file e960521a).



>Displaying j960521d-1241-what_did_you.sig,1241.000+0.873

Figure 4: Spectrogram of the word sequence "what did you wear" (file j960521d).



>Displaying i960531.sph,945.039+0.704

Figure 5: Spectrogram of the word sequence "what did you think of that" (file i960531).

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.

This year 12,300 new words were added to the LIMSI master lexicon for American English, which contains 95k entries. The new words consisted of 3800 entries to cover 1996 BN training data and an additional 8500 forms included in the new 65k LM. Many of the new words were proper names, whose pronunciations could be verified only if the word appeared in the training data.

As in last year's system, the lexicon contains the most common 1000 acronyms found in the training texts[5]. This year compound words were used to represent frequent word sequences which provided an easy way to allow reduced pronunciations such as /lεmi/ for "let me" and /g∧nx/ for "going to". Some example compound words and their pronunciations are given in Table 2. The first line corresponds to the original pronunciation formed by concatenation of the component words. The second line contains reduced forms added for the compound word.

Example spectrograms of sentences including the word sequence "what did you" are shown in Figures 3 - 5. In the first spectrogram, the speaker said all three words clearly and palatalized the /dy/ into a /J/. In the second, the speaker produced a flap for the combined final /t/ in "what" and the initial /d/ in "did". In the third example, the sequence was reduced to /w∧Jx/.

Decoding

Prior to decoding, segments longer than 30ms are chopped into smaller pieces so as to limit memory required for the trigram decoding pass. The chopping algorithm is as follows. 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. A Gaussian classifier is then used to estimate the gender for each segment using different model sets for each condition, and to label the Fx data as either wideband or telephone band.

Word recognition is performed in three steps: 1) word graph generation, 2) trigram pass, 3) segment-based acoustic model adaptation. A word graph is generated using a bigram backoff language model. This step uses a gender-specific sets of position-dependent triphones with about 6000 tied states and a small bigram language model (about 2M bigrams). Differents acoustic models are used for the different segment types. The model set is chosen based on the segment label. The sentence is then decoded using the word graph generated in the first step with a large set of acoustic models (position-independent triphones with about 7000 tied states) and a trigram language model (including 8M bigrams and 13M trigrams). Finally, unsupervised acoustic model adaptation is performed for each segment using the MLLR scheme, prior to the last decoding pass with the adapted models and the trigram LM.

Experimental results

The performance of the system at various stages of the development process is shown in Figure 6. The word error on the devtest data with M0 models and last year's Hub3 LM was 39.2%. Word graphs generated with the M0 models were then used to evaluate different acoustic and language models. The use of segment-based adaptation of the acoustic models gives a small improvement of 4% relative. Using type-specific acoustic models (sets M1 through M5) reduced the word error to 34.9% (7% relative). The combined use of the type-specific acoustic models and a language model trained on the Hub4 data resulted in a word error of 30.9%, an additional relative error reduction of 9%. After generating word graphs with type-specific acoustic models and the Hub4 LM, a word error of 25.2% was obtained. The use of WSJCAM0 data reduces the word error on F5 devdata by 12% (not shown in the figure).

M0 models, Hub3 LM, without adaptation	⇒	39.2%
M0 models, Hub3 LM, with adaptation	⇒	37.5%
M1 to M5 models, Hub3 LM, (M0 graphs)	⇒	34.9%
M1 to M5 models, Hub4 LM (M0 graphs)	⇒	30.9%
Generate graphs with M1 to M5, Hub4 LM	⇒	25.2%

Figure 6: Performance progression on the 1996 development data. The model set M0 was used in development work, but not in the final system. These models were adapted to the focus conditions using the BN training data, resulting in model sets M1-M5.

	Development data		*Evaluation data*	
Label	*Duration*	*WordErr*	*Duration*	*WordErr*
F0	25 min	11.5%	31 min	20.8%
F1	28 min	25.6%	32 min	26.0%
F2	19 min	34.3%	10 min	27.1%
F3	11 min	22.0%	7 min	20.3%
F4	16 min	19.0%	9 min	33.3%
F5	9 min	19.5%	2 min	27.8%
Fx	19 min	43.7%	14 min	46.1%
Overall	127 min	25.2%	106 min	27.1%

Table 4: Word error rates for the PE on the 1996 devdata and official NIST results on the evaltest data. (F0: baseline broadcast speech, F1: spontaneous broadcast, F2: speech over telephone channels, F3: speech in background music, F4: speech under degraded acoustic conditions, F5: non-native speakers, FX: other)

The evaluation test data were taken from 4 shows. The overall word error rate is 27.1% and the per show word errors are the following: CNN Morning News (29.7%), CSPAN Washington Journal (25.6%), NPR The World (30.5%) and NPR MarketPlace (23.0%). The word error by segment type is given in Table 4, along with the results on the development data. While there are substantial differences across the focus conditions, the overall error rates are comparable for the two data sets.

The word error on the F0 devdata is about half that of other conditions. The same is not true for the eval data, partially due to a long weather report which was spoken very quickly and had a high OOV rate. Speech over background music (F3) appears to be easier to handle than speech in noisy conditions (F4). This may be because speech over music usually occurs at the beginning and end of broadcasts, and is meant to be intelligible.

SUMMARY

In this paper we have described the LIMSI Nov96 Hub4 system and the development work in preparation for the evaluation. The 1996 Hub4 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 text data for language modeling. It is a multipass system, with more accurate acoustic and language models used in successive passes. Segment-based unsupervised adaptation is carried out prior to the final trigram decoding pass.

Our development work addressed primarily two problems encountered in transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. To deal with the varied acoustic conditions, the base acoustic models were trained on the secondary channel of the WSJ0/1 corpus, instead of the Sennheiser channel. Type-specific acoustic models were estimated for the different focus conditions using the 35 hours of task specific training data. To deal with the continuous flow of data, a chopping algorithm was developed so as to limit the amount of data to be processed as a single unit. New phones were added so as to explicitly model filler words and breath noises, as these phenomena are frequent in the broadcast news data. These effects were also directly represented in the language model. The development test data was used to optimize the recognition vocabulary and language models. Over 12000 new words were added to the lexicon, as well as compound words to allow modeling of reduced forms observed in spontaneous speech.

The problem of segmenting broadcast news shows has been investigated using 10 MarketPlace shows distributed as Nov95 training data. Compared to reference labels provided by BBN, the frame classification rate was 94%.

Using our Nov95 Hub3 65k word recognizer trained on the secondary channel of the WSJ corpus, an initial word error 39.2% was obtained on the Nov96 development data. After the development period, a word error of 25.2% was obtained on the same development test data with the evaluation setup. On the partitioned evaluation data from 4 shows, an overall word error of 27.1% was obtained (official NIST score).

REFERENCES

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

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

[3] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *ICASSP-95*, Detroit, May 1995.

[4] 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, TX, Jan. 1995.

[5] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "The LIMSI 1995 Hub3 System," *Proc. DARPA Speech Recognition Workshop*, Arden House, Feb. 1996.

[6] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*, Atlanta, May 1996.

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

[8] F. Kubala, T. Anastasakos, H. Jin, J. Makhoul, L. Nguyen, R. Schwartz,N. Yuan, "Toward Automatic Recognition of Broadcast News," *Proc. DARPA Speech Recognition Workshop*, Arden House, Feb. 1996.

[9] 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), 1995.

[10] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, "WSJ-CAM): A British English Speech COrpus for Large Vocabulary Continuous Speech Recognition," *ICASSP-95*. Detroit, May 1995..

[11] R. Stern et al., "Specification for the ARPA November 1996 Hub 4 Evaluation," Nov. 1996.