Developments in Large Vocabulary Dictation: The LIMSI Nov94 NAB System [†]

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

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,lamel,madda}@limsi.fr

ABSTRACT

In this paper we report on our development work in large vocabulary, American English continuous speech dictation on the ARPA NAB task in preparation for the November 1994 evaluation. We have experimented with (1) alternative analyses for the acoustic front end, (2) the use of an enlarged vocabulary of 65k words so as to reduce the number of errors due to out-of-vocabulary words, (3) extensions to the lexical representation, (4) the use of additional acoustic training data, and (5) modification of the acoustic models for telephone speech. The recognizer was evaluated on Hubs 1 and 2 of the fall 1994 ARPA NAB CSR Hub and Spoke Benchmark test. Experimental results on development and evaluation test data are given, as well as an analysis of the errors on the development data.

1. Introduction

Research in large vocabulary speaker-independent dictation at LIMSI[5, 6] makes use of large newspaper-based corpora such as the ARPA Wall Street Journal-based Continuous Speech Recognition corpus (WSJ)[14]. The LIMSI recognizer has been evaluated in the last 4 ARPA CSR Benchmark tests and most recently in the November 1994 North American Business (NAB) News CSR test, Hubs 1 and 2[4].

The goal of the Hub 1 *Unlimited Vocabulary NAB News Baseline* is to improve basic performance on unlimited-vocabulary, speaker-independent (SI) speech recognition of read-speech. The test prompts were selected from several sources of North American Business news (Dow Jones Information Services, New York Times, Reuters North American Business Report, Los Angeles Times, Washington Post). Results are reported for two systems: H1-C1, where the acoustic training data and the 20k trigram-backoff language model are fixed so as to assess and compare acoustic models; and H1-P0, where any techniques may be used to improve performance, and any acoustic and language model training data are permitted predating June 16, 1994. The LIMSI H1-P0 system used a 65k trigram language model. The aim of Hub 2 *Telephone NAB News* is to demonstrate SI recognition performance on unlimited-vocabulary read-speech over long-distance telephone lines. The LIMSI H2-P0 system was essentially our H1-P0 system adapted to the telephone channel, and with a 40k trigram language model.

2. General Recognizer Overview

In this section we give a general overview of the recognizer, which is quite similar to our Nov93 system. The primary issues addressed are acoustic modeling, language modeling, lexical representation, and the decoding strategy.

2.1. Acoustic Modeling

The recognizer makes use of continuous density HMM (CDHMM) with Gaussian mixture for acoustic modeling. The main advantage continuous density modeling offers over discrete or semi-continuous (or tied-mixture) observation density modeling is that the number of parameters used to model an HMM observation distribution can easily be adapted to the amount of available training data associated to this state. As a consequence, high precision modeling can be achieved for highly frequented states without the explicit need of smoothing techniques for the densities of less frequented states. Discrete and semi-continuous modeling use a fixed number of parameters to represent a given observation density and therefore cannot achieve high precision without the use of smoothing techniques. This problem can be alleviated by tying some states of the Markov models. However, since this requires careful design and some a priori assumptions, these techniques are primarily of interest when the training data is limited and cannot easily be increased.

The acoustic models are sets of context-dependent (CD), position independent phone models, which include both intra-word and cross-word contexts. The contexts to be modeled are automatically selected based on their frequencies in the training data. Using this approach, the most frequent triphone contexts are explicitly modeled and less frequent contexts are modeled by less specific models (right- and left-context phone models and context-independent phone models). Each phone model

[†] This work is partially funded by the LRE project 62-058 SQALE.

is a left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). The covariance matrices of all the Gaussians are diagonal. Separate male and female models are used to more accurately model the speech data. These models are obtained from speaker-independent seed models using Maximum a posteriori estimators[7].

2.2. Language Modeling

Language modeling entails incorporating constraints on the allowable sequences of words which form a sentence. Statistical *n*-gram models attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. Bigram and trigram backoff LMs language models were estimated on the 230 million word CSR LM-1 training text material (LDC, Aug94). A backoff mechanism [9] is used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved n-grams. Another advantage of the backoff mechanism is that LM size can be arbitrarily reduced by relying more on the backoff, by increasing the minimum number of required n-gram observations needed to include the n-gram. This property is used in the early decoding passes of the recognizer to reduce computational requirements.

2.3. Lexical Representation

The lexicons are represented phonemically using a set of 46 phonemes, including silence. Alternate pronunciations are provided for about 11% of the words (counted on the H1-C1 20k vocabulary and without taking into account alternate pronunciations due to optional phones). A pronunciation graph is generated for each word from the baseform transcription to which word internal phonological rules are optionally applied during training and recognition to account for some of the phonological variations observed in fluent speech. The training and test lexicons were created at LIMSI and include some input from modified versions of the TIMIT, Pocket and Moby lexicons. All pronunciations have been manually verified. Some example lexical entries are given in Figure 1. The first word "INTEREST", may be produced with 2 or 3 syllables, depending upon the speaker, where in the latter case the /t/ may be deleted. In contrast, the alternate pronunciations for "EXCUSE" reflect different parts of speech (verb or noun). In the third case, the abbreviation "CORP." may be pronounced in its full or its abbreviated form.

2.4. Decoding strategy

One of the most important problems in implementing the decoder of a large vocabulary speech recognizer is the design of an efficient search algorithm to deal with the huge search space, especially when using language models with a longer span than two successive words, such as trigrams. The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous beam search [12] which uses a dynamic programming procedure. This basic strategy has been extended by adding other features such as "fast match"[8, 1], N-best rescoring[16], progressive search[11] and one-pass dynamic network decoding[13]. The two-step approach used in our system is based on the idea of progressive search where the information between levels is transmitted via word graphs. Due to memory constraints, each step may consist of one or more passes, with each using successively more refined models. All decoding passes use cross-word CD triphone models. Prior to word recognition, sex identification is performed for each sentence using phone-based ergodic HMMs[10]. The word recognizer is then run with a bigram LM using the acoustic model set corresponding to the identified sex.

INTEREST	`IntrIst In{t}XIst`
EXCUSE	`Ekskyu[sz]`
CORP.	`kcrp kcrpXeSxn`
BAFFLING	`b@f[Ll]|G`

Figure 1: Example lexical entries, with phones in {} being optional, phones in [ ] being alternates.

The first step of the decoder uses a bigram-backoff LM with a tree organization of the lexicon for the backoff component. This one-pass frame-synchronous beam search, which includes intra- and inter-word CD phone models, and gender-dependent models, generates a list of word hypotheses resulting in a word lattice.

The tree representation of the backoff component (first introduced in our Nov92 CSR system) provides an efficient way of arbitrarily reducing the search space and of limiting the computational requirements of the first pass which represent on the order of 75% of the computation need for the entire decoding process. Additionally, this strategy allows us to use a static graph instead of building it dynamically and therefore avoids the computational bookkeeping costs associated with dynamic network decoding.

The key elements of the procedure used to generate the word graph from the word lattice are the following. In our implementation, a word lattice differs from a word graph only because it includes word endpoint information. First, a word graph is generated from the lattice by merging three consecutive frames (i.e. the minimum duration for a word in our system). Then, "similar" graph nodes are merged with the goal of reducing the overall graph size and generalizing the word lattice. This step is reiterated until no further reductions are possible. Finally, based on the trigram backoff language model a trigram word graph is then generated by duplicating the nodes having multiple language model contexts. Bigram backoff nodes are created when possible to limit the graph expansion. The trigram step may be carried out in more that one pass, using successively larger language models.

It should be noted that this decoding strategy based on multiple forward passes can in fact be implemented in a single forward pass using one or two processors. We are using a two-step solution because it is conceptually simpler, and also due to memory constraints.

3. Recent System Developments

In this section we describe the main aspects of our developmental work in anticipation of the Nov94 evaluations.

3.1. Acoustic Front End Optimization

The front end configuration used in our Nov92 and Nov93 WSJ systems was optimized using a portion of the Resource Management development data. For each frame (30 ms window), a 15 channel Bark power spectrum over the 8kHz bandwidth was obtained by applying triangular windows to the DFT output. From this 16 Bark-frequency scale cepstrum coefficients and their first and second order derivatives were computed.

We have since varied this analysis looking at different methods to obtain the cepstrum-based feature vector (LPCC vs MFCC), as well as the size of the feature vector. Analysis windows of 15ms, 20ms, 24ms, and 30ms were tried, with different spectral weightings such as the commonly used Mel and Bark frequency scales, and other intermediary interpolations. The number of filters was varied from 15 to 64, and the number of cepstral coefficients from 13 to 17.

Four sets of test data were used to assess the different analyses: the Nov92-5k, Nov93-S6, Nov93-H2 evaluation test data and the 1993 development test data SIdt-5k. In total, these contain 1275 sentences with 21,705 words from 28 speakers. All the experiments used a single set of 903 SI models trained on the standard SI-84 training set with the LIMSI Nov93 lexicons (training and 5k) which are publicly available, and the official 5k-nvp closed vocabulary LM model provided by Lincoln Labs. Even though this is nominally a closed vocabulary test, there is an out-of-vocabulary rate of 0.2%.

		% Word Error	
Test Data	# sentences	*Nov92/93*	*Nov94*
Nov93-S6	217	10.8	10.0
SIdt-5k	513	11.3	10.6
Nov92-5k	330	7.0	6.3
Nov93-h2	215	10.0	8.9
All	1275	9.9	9.1

Table 1: Experimental results on development data before and after optimization of the acoustic front end using the standard 5k-nvp closed vocabulary bigram LM.

The best configuration was found to be with a 30 ms frame and 26 cosine filters on a Mel scale over the 8kHz bandwidth, from which 15 cepstrum coefficients and a normalized energy are derived. The error rates for the new analysis (Nov94) and the old analysis (Nov92/93) are given for the individual test sets in Table 1. The overall error reduction is small (8%), but significant, and a consistent gain is obtained across the test sets, so this setup was used for the H1 systems in the Nov94 evaluation.

3.2. Use of Additional Acoustic Data

Last year we reported a word error reduction of about 30% in using the combined WSJ0/WSJ1 SI-284 training (37k sentences) as compared to SI-84 training (7k sentences) with a bigram LM[3]. On this year's H1-C1 dev data (trigram LM) we observed only a 15% error reduction when going from SI-84 training to SI-284 training. The improvement was obtained by increasing the number of CD models using a fixed threshold of 250 occurrences to model a given context.

This year we used all 85k sentences of WSJ0/WSJ1 read-speech training data, but observed only a small improvement of about 2% compared to SI-284 training with the same number of CD models. By increasing the number of CD modelsto 5000 (using the same fixed minimum count threshold of 250) increased the word error by about 4%. The reason for this disappointing result is surely due to the lack of homogeneity of the new data with the old, as all the additional data is essentially from a small number of long-term speakers. This is consistent with our previous observations that for our system better performance is obtained with the short-term speaker data (SI-84) than with comparable amounts of long term data (SI-12). In our 5k system, training comparable model sets with the long-term speakers data gives a word error 15-20% higher than that obtained with short-term speaker training.

3.3. Text processing/Lexical Coverage

The lexical coverage of the 5k and 20k most frequent words in the WSJ texts are only 90.6% and 97.5% respectively. With a 20k word vocabulary and unrestricted test data, we observe about 1.6 errors for each out-of-vocabulary (OOV) word. Thus, an obvious approach to reducing the errors due to OOV words is to increase the size of the lexicon. Our system is limited to a maximum vocabulary size of 65k words.

Prior to selecting a larger recognition vocabulary, the CSR LM-1 training texts were cleaned to remove the most frequent errors inherent in the texts or arising from processing with the distributed text processing tools. The cleaning consisted primarily of correcting obvious mispellings (such as MILLLION, OFFICALS, LITTLEKNOWN), systematic bugs introduced by the text processing tools, and expanding abbreviations and acronyms in a consistent manner. The texts were also transformed to be closer to the observed American reading style using a set of rules and the corresponding probabilities derived from the alignment of the WSJ0/WSJ1 prompt texts with the transcriptions of the acoustic data. Some example rules and their probabilities are:

HUNDRED <nb>	⟹	HUNDRED AND <nb> (0.5)
ONE EIGHTH	⟹	AN EIGHTH (0.50)
CORPORATION	⟹	CORP. (0.29)
INCORPORATED	⟹	INC. (0.22)
ONE HUNDRED	⟹	A HUNDRED (0.19)
MILLION DOLLARS	⟹	MILLION (0.15)
BILLION DOLLARS	⟹	BILLION (0.15)

The cleaning of the training texts reduced perplexity on development data by 5 points and resulted in a better coverage of the 65k lexicon. This lexicon was selected by measuring the perplexity and OOV rates on the development data (Dev93-H1, Nov93-H1 and Dev94-H1) for the most frequent 65k words in different subsets of the training texts. Our aim was to minimize the overall OOV rate, while assuring a good balance across data sets for OOV and perplexity. The 65k lexicon thus obtained consists of the 65,451 most common words of a subset of this training data (years 92-94) as this was found to provide significantly better lexical coverage than was obtained with all the data (years 87-94). In Table 2 the lexical coverage of several lexicons are given for the 1994 H1 and H2 data showing the combined effect of text cleaning and vocabulary selection. As stated earlier, the texts of the development data were removed from the LM training data so as to give better estimates of the lexical coverage on unseen data. For all test sets, the OOV rate with our 20k wordlist is significantly smaller than that of the baseline 20k wordlist. The OOV rate with the 65k wordlist on the Dev94 test data is 0.39% which is a pretty accurate indicator of the 0.42% observed on the 1994 H1 test data. The OOV rate with the 40k lexicon used in Hub 2 was 0.8% on the H1 development and evaluation test data, and higher 0.9% and 1.3% on the H2 development and evaluation test data, respectively.

	Lexicon			
Test set	*Baseline 20k*	*20k*	*40k*	*65k*
Dev94-H1	2.7	2.2	0.8	0.4
Eval94-H1	2.5	2.0	0.8	0.4
Dev94-H2	2.7	2.1	0.9	0.4
Eval94-H2	3.1	2.6	1.3	0.7

Table 2: OOV rate (%) on the H1 and H2 test sentences for 20k, 40k, and 65k lexicons.

After processing the training texts, removing all articles containing the prompts for the devtest acoustic data, and selecting the recognition lexicon, the H1-P0 65k and the H2-P0 40k language models were trained on the CSR training texts and read speech transcriptions predating June 16, 1994.

3.4. Recognition Lexicon

We also extended the training and recognition lexicons to include additional frequent pronunciations found in the training data as well as alternate pronunciations which have been seen to occur systematically. An example is the suffix "IZATION" which can be pronounced with a diphthong (/Y/) or a schwa (/x/). As always, we attempt to insure and improve the consistency of the pronunciations for similar words and different word forms. For example, in the new lexicon all words ending in "MANN" are transcribed with the phone sequence /m@n/. In previous versions this was transcribed as either /m@n/ or /mxn/ or both. We have observed that fast speakers tend to poorly articulate (and sometimes skip completely) unstressed syllable, particularly in long words with sequences of unstressed syllables. Although such long words are typically well recognized, often a nearby function word is deleted. In an attempt to reduce these kinds of errors, alternate pronunciations for long words such as AUTHORIZATION, POSITIONING, and REALISTICALLY were added to the lexicon allowing schwa-deletion or syllabic consonants in unstressed syllables. While these changes were not systematically evaluated, results with the new lexicon reduced the overall word error reported in Table 1 to 9.0%, with a small improvement on each individual test set. On the Dev94-H1 test data the improved lexicon reduced the word error from 13.0% to 12.8%.

The recognition lexicon was extended to the new 65k vocabulary. Pronunciations for the new words were generated by semi-automatically applying affix rules to existing lexical entires, or were added by hand. A substantial portion of the new lexical items were proper names, many of which are of foreign origin. In the 65k lexicon, 9% on the words have more than one pronunciation, and on average there are 1.1 pronunciations per word (not counting alternate pronunciation corresponding to optional phones). 4% of the words contain optional phones, typically stops or reduced vowels that are allowed to be deleted. The largest number of pronunciations for a single word is 8, for the word "apartheid" represented as /xpar[Tt][Ye][td]/. 5% of the entries have alternate pronunciations which are typically differences in fricative voicing or in vowel color such as USE /yu[zs]/ and DEVISE /dIvY[sz]/ (corresponding to different parts of speech), and DISNEY /dI[sz]ni/ and ADELSON /[@e]dLsxn/ (corresponding to different pronunciation variants.

3.5. Experiments with Telephone Data

In order to develop a Hub 2 system, we carried out experiments with the Nov93 Spoke 6 evaluation test data which provides parallel speech data for wideband and telephone quality speech. The multichannel data allows more accurate comparisons to be made by controlling some of the factors that affect recognition accuracy. The system was evaluated using the 5k vocabulary and standard trigram LM. For the telephone speech the acoustic feature vector contains 13 MFCCs and their first and 2nd order derivatives computed on the 3.5kHz bandwidth every 10ms.

The basic idea is to start with clean speech models and to adapt them to the telephone channel conditions. This adaptation is performed by reducing the bandwidth of the clean speech and adapting the reduced bandwidth acoustic models with telephone speech. For each of the training sets (SI-84 and SI-284) we built 3 sets of acoustic models so as to measure the recognition performance in different acoustic channel conditions and to evaluate the progressive reduction in channel mismatch. These 3 sets correspond to training with 8kHz bandwidth clean speech, training with reduced bandwidth clean speech, and to adaption of the latter model set with telephone speech.

Training	*Test data*		
Conditions	*Senn., 8k*	*Senn., Tel*	*Tel.*
SI-84	7.5	8.0	14.8
SI-84 + ad	-	8.5	12.1
SI-284	6.3	6.3	13.1
SI-284 + ad	-	7.2	10.4

Table 3: Experimental results on 1993 Spoke 6 evaluation test data using the standard 5k lexicon and trigram LM.

Experimental results are given in Table 3 for SI-84 and SI-284 training with and without telephone adaptation data, for 3 channel conditions: Sennheiser 8kHz, Sennheiser reduced bandwidth, and telephone. On the Sennheiser 8kHz data, word errors of 7.5% and 6.3% were obtained with SI-84 and SI-284 models, respectively. Using a reduced bandwidth analysis increased the word error to 8.0% for SI-84 training, but no error increase was observed for SI-284 training. For the telephone speech data, the channel mismatch has been partially compensated for by adapting the clean speech models with a relatively small amount of telephone data (only 403 sentences from Dev93-S6 for SI-84, and 7,130 sentences for SI-284). With the adapted SI-84 models, the word error on telephone data was reduced by 18%, and the word error on Sennheiser data increased by 6%. For the adapted SI-284, the word error on the telephone data was reduced by about 21%, with an increase of 14% on the Sennheiser data. Thus, the additional training data used to adapt the SI-284 models leads to a better match to the telephone channel. The word error on telephone data is about 60% higher than the error rate obtained for the Sennheiser data.

4. Nov94 NAB Systems

The system configurations used in the Nov94 NAB CSR evaluation are described in this section, along with experimental results on the H1 and H2 tests.

4.1. Nov94 NAB H1 System

The acoustic models used in the baseline test H1-C1 were trained on the standard set of 37,518 WSJ0/WSJ1 sentences (SI-284, primary microphone). The resulting two sets of 3309 gender-dependent models each have 308k Gaussians. For the primary system, H1-P0, all the available WSJ0/WSJ1 training data (85,343 sentences from 359 speakers) were used to train two sets of 3600 gender-dependent acoustic models. Each model set has 343k Gaussians.

For the H1-C1 system, the official 20k trigram language model provided by CMU was used[15]. For the H1-P0 condition, a 65k trigram LM was trained on the cleaned-up versions of the standard CSR LM-1 training texts (years 87-94), the 1994 NAB development data (excluding articles containing the dev test prompts), and the WSJ0/WSJ1 read speech transcriptions (85,343 sentences). The CMU language modeling toolkit[15] was used to build the 65k LM.

	Test data	
System	*Dev94*	*Eval94*
H1-C1, 20k	12.8	12.7
H1-P0, 40k	9.8	10.3
H1-P0, 65k	-	9.8
H2-P0, 40k	-	25.1

Table 4: Results on 1994 test data (unadjudicated[1]).

For the H1-C1 system, the first pass of the decoder used a bigram-backoff LM with a cutoff of 10. This resulted in a word graph with about 2.2M interword connections, including those corresponding to the lexicon tree of the backoff component. The resulting phone graph has 169k phone nodes and 2.6M arcs. The same bigram cutoff was used for the H1-P0 and H2-P0 systems.

4.2. Hub-1 Experimental Results

The Nov94-H1 devtest data contains 316 sentences from 20 speakers, each with prompt texts selected from North American Business news. Recognition results for the Nov94 tests are given in Table 4. For comparison, results are also given for the Dev94-H1 data containing 310 sentences from 20 speakers. The H1-C1 results are seen to be comparable for the two data sets. The use of a larger vocabulary is seen to substantially reduce the word error, mainly by reducing the OOV rate. Compared to the H1-C1 system, the H1-P0 system reduces the word error by 23%.

To better understand the errors due to OOV words, a detailed analysis of the 198 OOV words in the Dev94-H1-C1 test was carried out. On average, 1.6 word errors are generated for each OOV word. 45% of the OOV errors are single word substitutions and 45% have 2 errors. The remaining 10% generate 3 or more errors. The use of a 40k vocabulary reduces the OOV rate from 2.7% to 0.8%, so potentially 70% of the 20k OOV words can be recognized. In the 40k run, 45% the 20k OOV words were correctly recognized. Some examples of typical errors on OOV words are:

STRINGER	$\Longrightarrow$	STRANGER
MARCH'S	$\Longrightarrow$	MARCHES
DIVORCES	$\Longrightarrow$	DIVORCE IS
BUSIER	$\Longrightarrow$	BUSY YOUR
NORIYUKI	$\Longrightarrow$	NOR YOU KEEP

In the first two examples an unknown word is replaced by a homophone or a phonemically close word. The next two words DIVORCES and BUSIER generate two errors the root word and a function word to replace the suffix. In addition there are errors due to compound words such as OVERBLOWN being recognized as the sequence OVER BLOWN, which should perhaps not really be considered as errors. Reducing the OOV rate recovers on average 1.2 errors for every OOV word removed.

4.3. Nov94 NAB H2 System

Our aim for the Hub 2 test was to minimally change our H1-P0 system and to run it on the telephone data. For the telephone hub, H2-P0, a reduced bandwidth analysis was carried out as described earlier, and SI models were built from the SI-284 primary microphone (Sennheiser) data. These models were then adapted using MAP estimation with 7130 sentences: 403 sentences from the 1993 WSJ1 Spoke 6 development test data, 313 sentences from 1994 H2-dev data and 6,414 WSJ sentences from the macrophone corpus[2]. Due to time constraints we were not able to directly port our H1-P0 system to this task, and needed to limit the vocabulary size to 40k words. The 40k vocabulary list was obtained by selecting the 39,637 most common words of the 65k word list. The OOV rate of this vocabulary was 0.9% on the dev94-h2 data as given in Table 2. The 40k LM was trained on the same text material as the H1-P0 system, i.e., on the cleaned-up versions of the standard CSR LM-1 training texts (years 87-94), the 1994 NAB development data (excluding articles containing the dev test prompts), and the WSJ0/WSJ1 read speech transcriptions (85,343 sentences). We also used a single set of 1928 gender-independent CD models, compared to two sets of 3600 models as used in H1-P0. This model set had 184k Gaussians.

We observed that using comparable pruning thresholds for H2 as had been used in H1 considerably increased the decoding time, as well as the word lattice size. So in order to keep the decoding time and the memory requirements essentially the same as the H1 system, a much more aggressive pruning level was used at the risk of introducing search errors.

4.4. Hub-2 Experimental Results

The Hub 2 test data consists of 20 speakers reading about 15 sentences each for a total of 312 sentences. The prompt texts were taken from the same source as the H1 test, but the exact texts and speakers are not the same. The word error for the H2-P0 test with a 40k vocabulary is 25.1%. The error rate is over twice that of the H1-P0 40k system. This difference is larger than that observed in our development work with the matched Spoke 6 data

[1] We have chosen to provide the unadjudicated results in order to facilitate a comparison with results on the development test data. The adjudicated word error rates on the Nov94 evaluation test data are: H1-P0: 9.2%, H1-C1: 12.1%, H2-P0: 24.6%.

(see Table 3) and may be attributed to differences in the channel, as well as to the speaking style which seems to be less formal. The Hub 2 data was recorded over long distance telephone lines in unknown environments, and whereas the Spoke 6 data were recorded at SRI over external lines.

4.5. Additional Observations

Since a word graph is used to transmit information between successive passes, it is obviously important that the correct solution be in the graph. In general, we have found the word graph error to be small, on the order of 2% for the graph used in the last pass (i.e. the worst case). However, we have noticed that poor speakers tend to have higher graph error rates, which can be as high as 10%. The average graph error on the telephone data is 8%, which is significantly higher than that of the Sennheiser channel.

More generally, the system appears to not be very robust with regard to channel and speaker differences. The 40k H2 system had a word error of 25%, compared to 10% for the 40k H1 system. We also have observed large differences in word error across speakers. Concerning the Dev94-H1 test set the best speaker (4q9) had an error rate of 3.4%, whereas the worst speaker (4qg) had a word error of 42.7%. (This speaker is difficult for even humans to understand.) A large difference in error rate was also observed for the Nov94-H1 test data where the word error ranged from 1.3% for the best speaker (4t3) to 24.5% for the worst (4td). Some of the errors may be attributed to higher than average OOV rates or high perplexity sentences, where the text is not well predicted by the language model. However, the high error rates observed for poor speakers are primarily due to non-standard pronunciations and to poorly articulated words (which frequently occur for fast speaking rates). In analyzing the errors for the worst speakers, we observed many errors involving groups of frequent short words such as "WHERE DO YOU GET" which was pronounced as *"where'dya get"* and recognized as "WEREN'T GET" or "WERE TICKET".

5. Summary

In the paper we have presented our 1994 ARPA NAB CSR system and highlighted some of the more important aspects of our development work. We developed a 65k-word speech recognizer which makes use of phone-based CDHMMs with Gaussian mixture for acoustic modeling and 3-gram statistics estimated on NAB newspaper texts for language modeling. The system uses a multipass decoder, where more accurate models are used in successive passes and information is transmitted between passes via word graphs.

During our development work, we mainly worked on improving the acoustic front end, the lexical coverage, the lexicon representation and the acoustic models through the use of more acoustic data. Regarding this last point, we were disappointed to observe that by using as many as 85k sentences of acoustic training data instead of 37k sentences (SI-284 data set) does not significantly improve the model accuracy. We attribute this partly to the fact that the mixture of "long-term" and "short-term" speakers in the 85k sentences constitutes an inhomogeneous data set that our current training strategy is not able the use adequately.

In order to port our system to the telephone channel, we adapted acoustic models trained on reduced bandwidth clean speech with a relatively small amount of telephone training data coming primarily from the Macrophone corpus.

For a speaker-independent, open-vocabulary read-speech test, a word error of 9.8% was obtained with a 65k vocabulary system. Using a vocabulary of 40k words, a word error of 10.3% was obtained. With the same 40k vocabulary the word error on telephone speech from different speakers is 25.1%

Increasing the vocabulary size, at least up to 65k words, was found to reduce the average word error. This simple approach to reducing the errors due to OOV words appears to be effective despite the potential increased confusability of the lexical entries. We observed that by reducing the OOV rate, we recover on average 1.2 times as many errors as OOV words removed.

The observed large difference in performance across speakers is certainly an outstanding challenge for speech recognition. The high error rates observed for poor speakers arise mainly from non-standard pronunciations and high speaking rates which result in poorly articulated words. We have observed that better acoustic and language models do not significantly improve these errors. Modeling at the phonological level, perhaps with particular pronunciations that are invoked for frequent word sequences or for fast speakers, and speaker adaptation techniques may be needed to improve performance.

References

1. L.R. Bahl et al, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *Proceedings ICASSP-92.*
2. J. Bernstein, K. Taussig, J. Godfrey, "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project," *Proceedings ICASSP-94.*
3. J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task,"

Proceedings ICASSP-94.

4. J.L. Gauvain, L.F. Lamel, M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *Proceedings ICASSP-95*.
5. J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System," *Proceedings ARPA Human Language Technology Workshop*, 1994.
6. J.L. Gauvain,, L.F. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**, (1-2), 1994.
7. J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.2, April 1994.
8. L. Gillick, R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *Proceedings DARPA Speech & Natural Language Workshop*, 1990.
9. S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.
10. L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Proceedings Eurospeech-93*.
11. H. Murveit, J. Butzberger, V. Digalakis, M. Weintraub, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Proceedings ICASSP-93*.
12. H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. ASSP*, **32**(2), pp. 263-271, April 1984.
13. J.J. Odell, V. Valtchev, P. Woodland, S. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proceedings ARPA Workshop on Human Language Technology*, 1994.
14. D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *Proceedings ICSLP-92*.
15. R. Rosenfeld, "The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation,", *Proceedings ARPA Spoken Language Systems Technology Workshop*, 1995.
16. R. Schwartz, S. Austin, F. Kubala, J. Makhoul.,"New uses for N-Best Sentence Hypothesis within the BYBLOS Speech Recognition System," *Proceedings ICASSP-92*.