Developments in Continuous Speech Dictation using the ARPA WSJ Task[†]

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 recent development work in large vocabulary, American English continuous speech dictation. We have experimented with (1) alternative analyses for the acoustic front end, (2) the use of an enlarged vocabulary 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 for development and evaluation test data are given, as well as an analysis of the errors on the development data.

INTRODUCTION

Research in large vocabulary speaker-independent dictation at LIMSI[3, 4] makes use of large newspaper-based corpora such as the ARPA Wall Street Journal-based CSR corpus (WSJ)[8]. The recognizer uses phone-based CDHMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. The LIMSI recognizer has been evaluated in the last 3 ARPA CSR Benchmark tests and most recently in the November 1994 North American Business (NAB) News CSR test, Hubs 1 and 2.

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 asess 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 aim of Hub 2 *Telephone NAB News* is to demonstrate SI recognition performance on unlimited-vocabulary read-speech over long-distance telephone lines.

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

NOV94 NAB RECOGNIZER

The recognizer makes use of continuous density HMM with Gaussian mixture for acoustic modeling. Cepstral mean removal is performed for each sentence. The acoustic models are sets of context-dependent (CD), position-independent phone models, which include both intra-word and cross-word contexts. The contexts are automatically selected based on their frequencies in the training data and include triphone models, right- and left-context phone models, and context-independent phone models. Each phone model is a left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). Maximum a posteriori estimators[5] are used to derive separate male and female models from speaker-independent HMM parameters, so as to more accurately model the speech data.

For the baseline test H1-C1, the standard set of 37,518 WSJ0/WSJ1 sentences (SI-284, primary microphone) has been used for training two sets of 3309 gender-dependent acoustic models. 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. For the telephone hub, H2-P0, a reduced bandwidth analysis was carried out, and SI models were built from the primary microphone (Sennheiser) data. These models were adapted using MAP estimation with 7130 sentences: 403 sentences from the 1993 Spoke 6 development test data, 313 sentences from 1994 H2-dev data and 6,414 WSJ sentences from the macrophone corpus[1].

N-gram statistics estimated on newspaper texts are used for language modeling. Bigram and trigram backoff LMs were trained on the 230 million word CSR LM-1 training text material (LDC, Aug94). The backoff mechanism[6] 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. For the H1-C1 test, the standard 20K trigram LM provided by CMU was used. For the H1-P0 condition, a 65k trigram LM was trained on the standard CSR LM-1 training texts (years 87-94), the 1994 NAB development data

(excluding articles including the dev test prompts), and the WSJ0/WSJ1 read speech transcriptions (85,343 sentences). For H2-P0, a 40k word trigram LM was used, where the 40k vocabulary contains the most common 39,639 words in the H1-P0 65k word list.

Decoding is carried out in two forward acoustic passes[2]. The first pass is a time-synchronous graph-search which includes intra- and inter-word CD phone models and gender-dependent models, and a bigram LM. The second decoding pass makes use of a word graph generated with the bigram and incorporates a trigram LM. Prior to recognition, the gender of each sentence is identified using phone-based ergodic HMMs[7], then word recognizer is run using the model set of the identified gender. Both passes use a time-synchronous Viterbi decoder.

To improve the performance of the system we have been exploring several directions to reduce the loss of linguistic information by the front end, to increase the system robustness and to achieve higher precision modeling. In the next sections we focus on several aspects of our development work mostly carried out using a 5k system.

ACOUSTIC FRONT END OPTIMIZATION

The front end configuration used in our Nov92 and Nov93 systems was optimized using 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 30ms, 24ms, 20ms, 15ms 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%.

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 Nov94 evaluation.

		% 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.

TEXT PROCESSING/LEXICAL COVERAGE

The lexical coverage of the 5k and 20k most frequent words in the WSJ texts are 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. An obvious approach to reducing the errors due to OOV words is to increase the size of the lexicon.

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 <number> → HUNDRED AND <number> (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 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

	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
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 64k lexicons.

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.

data. As stated in the Nov94 recognizer description, 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. The OOV rate on the Dev94 test data is 0.39% which is a pretty good indicator of the 0.42% observed on the 1994 H1 test data.

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%.

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[2]. On this year's H1-C1 dev data (trigram LM) we observed only a 15% error reduction with SI-284 training. 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. The reason for this 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.

EXPERIMENTS WITH TELEPHONE DATA

We have experimented with the Nov93 Spoke 6 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 3.6kHz bandwidth every 10ms.

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 (see recognizer description) 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.

ARPA NOV94 EXPERIMENTAL RESULTS

A description of the 1994 Hubs 1 and 2 was given in the introduction. 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 2 data sets. The use of a larger vocabulary is seen

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

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

to substantially reduce the word error, mainly by reducing the OOV rate.

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 → STRANGER
MARCH'S → MARCHES
DIVORCES → DIVORCE IS
BUSIER → BUSY YOUR
NORIYUKI→ 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.

Large differences in word error are observed 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%. The same was 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). 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".

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 (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.

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. Experimental results were presented for different test sets and conditions. The system is a multipass system, where more accurate models are used in successive passes. Thus the first pass which is used to generate the initial word lattice must use accurate enough acoustic models so as to not introduce lattice errrors which are evidently unrecoverable with further processing. In practice the graph error is small ($\sim$ 2%), but poor speakers tend to have higher graph errors, and the average graph error on the telephone data is 8%. 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 about 10% was obtained. With the same 40k vocabulary the word error on telephone speech is 25.1%

The previously mentioned large difference in performance across speakers is certainly an outstanding challenge for speech recognition. There are 2 main reasons for high error rates: (1) OOV words, and (2) non-standard pronounciations, especially for poorly articulated words and high speaking rates. The first problem can be handled through increased lexicon sizes which have been demonstrated to improve performance, despite the probable introduction of homophones. The observed gain in word error rates is roughly 1.2 times the reduction in OOV rate. Concerning the second problem, 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] J. Bernstein, K. Taussig, J. Godfrey, "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project," *ICASSP-94*.

[2] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA WSJ Task," *ICASSP-94*.

[3] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System," *ARPA HLT-94*.

[4] J.L. Gauvain et al., "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**, (1-2), 1994.

[5] J.L. Gauvain, C.H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, **11**(2-3), 1992.

[6] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.

[7] L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Eurospeech-93*.

[8] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP-92*.