The LIMSI Nov93 WSJ System

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

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

ABSTRACT

In this paper we report on the LIMSI Wall Street Journal system which was evaluated in the November 1993 test. The recognizer makes use of continuous density HMM with Gaussian mixture for acoustic modeling and n-gram statistics estimated on the newspaper texts for language modeling. The decoding is carried out in two forward acoustic passes. The first pass is a time-synchronous graph-search, which is shown to still be viable with vocabularies of up to 20k words when used with bigram back-off language models. The second pass, which makes use of a word graph generated with the bigram, incorporates a trigram language model. Acoustic modeling uses cepstrum-based features, context-dependent phone models (intra and interword), phone duration models, and sex-dependent models. The official Nov93 evaluation results are given for vocabularies of up to 64,000 words, as well as results on the Nov92 5k and 20k test material.

1. Introduction

Our speech recognition research focuses on developing recognizers that are task-, speaker- and vocabulary-independent so as to be easily adapted to a variety of applications. In this paper we report on our efforts in large vocabulary, speaker-independent continuous speech recognition for American English using the ARPA Wall Street Journal-based CSR corpus [18]. The WSJ corpus is the designated common task for continuous speech recognition work in the ARPA community, and has become a task used for comparative development and evaluation worldwide. LIMSI participates in the ARPA-run evaluations in an effort to promote international scientific exchange.

The WSJ corpus contains large amounts of read speech material from a large number of speakers and has associated text material which can be used as a source for statistical language modeling. In these experiments two sets of standard speech training material have been used: WSJ0 and WSJ1, as well as the 37 M-word standardized text training material.

The recognizer makes use of continuous density HMM with Gaussian mixture for acoustic modeling and n-gram statistics estimated on text material for language modeling. Acoustic modeling uses cepstrum-based features, context-dependent phone models, duration models, and sex-dependent models. Statistical n-gram language models are estimated on the training corpus of newspaper text from the WSJ. In the following sections the recognizer is described and recognition results of the current system on the last two sets of evaluation test material, Nov92 [16] and Nov93 [17], are given.

2. Recognizer Overview

The recognizer uses a time-synchronous graph-search strategy which is shown to still be viable with vocabularies of up to 20k words, when used with bigram back-off language models (LMs). This one level implementation includes intra- and inter-word context-dependent (CD) phone models, intra- and inter-word phonological rules, phone duration models, and gender-dependent models [10]. The HMM-based word recognizer graph is built by putting together word models according to the grammar in one large HMM. Each word model is obtained by concatenation of phone models according to the word's phone transcription in the lexicon.

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 is that the number of parameters used to modelize 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. In the experimental section we demonstrate the improvement in performance obtained on the same test data by simply using additional training material. 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 in order to have more training data to estimate each state distribution. However, since this kind of tying 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 main characteristics of the recognizer are:

Front end: A 48-component feature vector is computed every 10 ms. This feature vector consists of 16 Bark-frequency scale cepstrum coefficients computed on the 8kHz bandwidth with their first and second order derivatives. For each frame (30 ms window), a 15 channel Bark power spectrum is obtained by applying triangular windows to the DFT output. The cepstrum coefficients are then computed using a cosine transform [3].

Acoustic models: The acoustic models are sets of CD phone models, which include both intra-word and cross-word contexts, but are position independent. Each phone model is a left-to-right CDHMM with Gaussian mixture observation densities. The covariance matrices of all the Gaussians are diagonal. Duration is modeled with a gamma distribution per phone model. The HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. Maximum a poste-

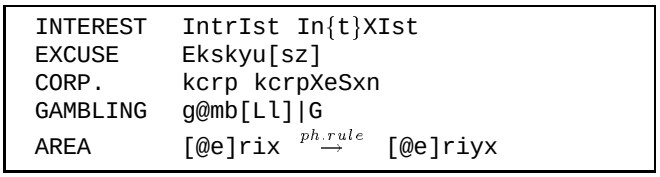

Figure 1: Example lexical entries, with phones in {} being optional, phones in [ ] being alternates.

riori (MAP) estimators are used for the HMM parameters [7] and moment estimators for the gamma distributions. Separate male and female models have been used to more accurately model the speech data. The contexts to be modeled are selected based on their frequency of occurrence in the training data. Experiments were carried out with model sets ranging in size from 493 models to 3306 models. In the experimental section we demonstrate the improvement in performance obtained by increasing the number of phone models to take advantage of the additional training material in the WSJ1 corpus.

Lexicon: The lexicon is represented phonemically using a set of 46 phonemes. The lexicon has alternate pronunciations for some of the words, and allows some of the phones to be optional. 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 predictable pronunciation variants. 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. Training and test lexicons were created at LIMSI and include some input from modified versions of the TIMIT, Pocket and Moby lexicons. Missing forms were generated by rule when possible, or added by hand. Some pronounciations for proper names were kindly provided by Murray Spiegel at Bellcore from the Orator system. Recognition lexicons containing 5k, 20k, and 64k words obtained from the standard word lists have been used in the experiments described below.

Language Model: 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. Unless otherwise specified, in this work the bigram and trigram language models provided by Lincoln Labs [18], are used. These were estimated on the 37M words training material of the WSJ. 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.

Decoding: The recognizer uses a time-synchronous graph-search strategy which includes intra- and inter-word CD phone models, intra- and inter-word phonological rules, phone duration models, and a bigram language model. Sex identification is performed for each sentence using phone-based ergodic HMMs [12]. The recognizer is then run using the set of models corresponding to the identified sex. When using a trigram LM, sentence recognition is performed in two forward passes. First, a word graph is generated using a bigram language model. Second, the sentence is decoded using the acoustic models and the trigram language model on the reduced search space provided by the word graph. Both passes use a time-synchronous Viterbi decoder.

Phonological Rules: Phonological rules are used to allow for some of the phonological variations observed in fluent speech. The principle behind the phonological rules is to modify the phone network to take into account such variations. These rules are optionally applied during training and recognition. Their use during training results in better acoustic models, as they are less "polluted" by wrong transcriptions. Using optional phonological rules during recognition can reduce the number of mismatches. The mechanism for the phonological rules allows the potential for generalization and extension by addition of new rules. The phonological rules may apply word-internally or may apply at word boundaries. The word-internal phonological rules are applied to the baseform transcriptions of the lexical entry when generating its pronunciation graph. An example of applying a phonological rule for glide insertion (in this case /y/) is shown in 1. In forming the word network, word boundary phonological rules are applied at the phone level to take into account interword phonological variations. For the present, only well known phonological rules have been incorporated in the system. These rules include word-internal rules for glide insertion, stop deletion, and homorganic stop insertion. The interword rules include palatalization, stop reduction, and voicing assimilation.

During system development phone recognition has been used to evaluate different acoustic model sets. It has been shown that improvements in phone accuracy are directly indicative of improvements in word accuracy when the same phone models are used for recognition [11]. Phone recognition provides the added benefit that the recognized phone string can be used to understand word recognition errors and problems in the lexical representation.

3. Search Strategy

One of the most important problems in implementing a large vocabulary speech recognizer is the design of an efficient search algorithm to deal with the huge search space, especially when using "long" span language models such as trigrams. The most commonly used approach for small and medium vocabulary sizes is the one-pass frame synchronous beam search [15] which uses a dynamic programming procedure. This basic strategy has been recently extended by adding other features such as "fast match" [8, 2], N-best rescoring [19], and progressive search [14]. The two-pass approach used in our system is based on the idea of progressive search [14] where the information between levels is transmitted via word graphs.

The first pass uses a bigram-backoff language model with a tree organization of the lexicon for the backoff component.[1] This one-pass frame synchronous beam search generates a list of word hypotheses resulting in a word lattice. Since the size of the second pass' search space is directly proportional to the size of this word lattice, it is desirable that this size remain as small as possible.

Two problems need then to be considered. The first is whether or not the dynamic programming procedure used in the first pass, which guarantees the optimality of the search for the bigram, generates an

[1] An advantage offered by 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 can be used in the first bigram decoding pass to reduce computational requirements.

"optimal" lattice to be used with a trigram language model. For any given word in the lattice, there will be many hypotheses with different ending points but only a few hypotheses with different starting points. This problem, which motivated forward-backward approaches [1], was in fact less severe than expected since the time information appears to not be critical for generating an "optimal" word graph from the lattice. The multiple word endings were found to provide enough flexibility to compensate for single word beginnings.

The second consideration is that the lattice generated in this way cannot be too large or there is no interest in a two-pass approach. To solve this second problem, two pruning thresholds are used during the first pass, a beam search pruning threshold which is kept to a level insuring almost no search errors (from the bigram point of view) and a word lattice pruning threshold which is used to control the lattice size.

While a complete description of the procedure used to generate the word graph from the word lattice is beyond the scope of this paper, the following steps provide the key elements behind the procedure.[2] First, a word graph is generated from the lattice by merging three consecutive frames, which is 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.

To fix these ideas, let us consider some numbers for the WSJ 5k closed vocabulary. The first pass generates a word lattice containing on average 10,000 word hypothesis per sentence, with the pruning threshold set to have a negligable number of search errors. The generated word graph before trigram expansion contains on average 1400 arcs. After trigram expansion, based on a trigram backoff LM there are on average 3900 word instanciations including silences which are treated the same way as words.

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

4. Experimental Results

Two sets of standard training material have been used for these experiments: The standard WSJ0 SI-84 training data which include 7240 sentences from 84 speakers, and the standard set of 37518 WSJ0/WSJ1 SI-284 sentences from 284 speakers. Only the primary microphone data were used for training. Using the SI-84 training data, model sets containing respectively 493 (si84a), 884 (si84b), and 1084 (si84c) models were trained, by varying the number of occurrences of a triphone required in the training material. The minimal number of occurrences were 500, 250 and 200 respectively. A set of 3306 models were trained from the SI-284 training material where each phone context had at least 250 occurrences in the training data.

While we have built n-gram-backoff LMs directly from the 37M-word standardized WSJ training text material, in these experiments all results are reported using the 5k or 20k, bigram and trigram backoff LMs provided by Lincoln Labs [18] as required by ARPA for participation in the tests.

[2] In our implementation, a word lattice differs from a word graph only because it includes word endpoint information.

5k - Conditions	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
Nov92, si84a, bg*	91.8	6.9	1.3	1.5	9.7
Nov92, si84c, bg	94.4	5.0	0.6	0.9	6.6
Nov92, si284, bg	96.0	3.6	0.3	0.9	4.8
Nov92, si284, tg	97.7	2.1	0.2	0.8	3.1
Nov93, si84c, bg	91.9	6.2	1.9	1.3	9.4
Nov93, si284, bg	94.1	4.8	1.2	0.9	6.8
Nov93, si284, tg	95.5	3.5	1.1	0.8	5.3

Table 1: 5k results - Word recognition results on the WSJ corpus with bigram/trigram (bg/tg) grammars estimated on WSJ text data. *official ARPA NOV92 evaluation results.

The WSJ corpus provides a wealth of material that can be used for system development. In our experiments, we have worked primarily with the WSJ0-Dev (410 sentences, 10 speakers), and the WSJ1-Dev from spokes S5 and S6 (394 sentences, 10 speakers). Development was done with the 5k closed vocabulary system in order to reduce the computational requirements. The Nov92 5k and 20k nvp test sets were used to assess progress during this development phase.

The LIMSI WSJ system was evaluated in the Nov92 DARPA evaluation test for the 5k-closed vocabulary using the standard bigram language models [18] with the WSJ0 SI-84 training data. The official reported results are given in the first line of Table 1 using 493 CD models (si84a), without the second derivative of the cepstral coefficients. Increasing the number of CD models and the number of features (si84c), reduced the error rate by about 30% over the system used for the Nov92 evaluation. With the same model set, a word error of 9.4% was obtained on the Nov93 test data. Using the combined WSJ0/WSJ1 SI-284 training data reduces the error by about 27% for both tests. When a trigram LM is used in the second pass, the word error is reduced by 35% on the Nov92 test and by 22% on the Nov93 test. The gap between the Nov92 and Nov93 results is mainly due to speaker differences, as the perplexity for both test sets are almost the same (111 for Nov92 versus 106 for Nov93).

Results are given in the Table 2 for the Nov92 nvp 64k test data using both open and closed 20k vocabularies. With SI-84 training (si84b) the word error rate is doubled when the vocabulary increases from 5k to 20k words and the test perplexity goes from 111 to 244. The

20k - Conditions	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
Nov92, si84b, bg	88.3	10.1	1.5	2.0	13.6
Nov92+, si84b, bg	86.8	11.7	1.5	2.7	15.9
Nov92+, si284, bg	91.6	7.6	0.8	2.6	11.0
Nov92+, si284, tg	93.2	6.2	0.6	2.3	9.1
Nov93+, si284, bg	87.1	11.0	1.9	2.3	15.2
Nov93+, si284, tg	90.1	8.5	1.4	1.9	11.8

Table 2: 20k/64k results - Word recognition results with 20,000 word lexicon on the WSJ corpus. Bigram/trigram (bg/tg) grammars estimated on WSJ text data. +: 20,000 word lexicon with open test.

higher error rate with the 20k open lexicon can be attributed to the out-of-vocabulary (OOV) words, which account for almost 2% of the words in the test sentences. Processing the same 20k open test data with a system trained on the SI-284 training data, reduces the word error by 30%. The word error on the Nov93 20k test is 15.2% with the same system. The use of a trigram reduces the error rate by 18% on the Nov92 test and 22% on the Nov93 test. As for the 5k tests, the higher error rate for the Nov93 test data can be primarily attributed to speaker differences.

The 20k trigram sentence error rates for Nov92 and Nov93 are 60% and 62% respectively. Since this is an open vocabulary test, the lower bound for the sentence error is given by the percent of sentences with OOV words, which is 26% for Nov92 and 21% for Nov93. In addition, there are inevitably errors introduced by the use of word graphs generated by the first pass. The graph error rate (ie. the correct solution was not in the graph) was 5% and 10% respectively for Nov92 and Nov93. In fact, in most of these cases the errors should not be considered search errors as the recognized string has a higher likelihood than the correct string.

A final test was run using a 64k lexicon in order to eliminate the errors due to unknown words. (In principle, all of the read WSJ prompts are found in the 64k most frequent words, however, since the WSJ1 data were recorded with non-normalized prompts, additional OOV words can occur.) Running a full 64k system was not possible with the computing facilities available, so we added a third decoding pass to extend the vocabulary size. Starting with the phone string corresponding to the best hypothesis of the 20k trigram system, an A* algorithm was used to generate a word graph using phone confusion statistics and the 64k lexicon. This word graph was then used by the recognizer with a 64k trigram grammar constructed at LIMSI using the standard 37M-word WSJ training texts. Using this approach we recovered only about 30% of the errors due to OOV words on the Nov93 64k test, reducing the word error to 11.2% from 11.8%.

5. Discussion

In this paper, we have described the LIMSI Nov93 continuous speech dictation system. The system uses CDHMM with Gaussian mixture for acoustic modeling and n-gram statistics estimated on the newspaper texts for language modeling. The recognizer uses a time-synchronous graph-search strategy which is shown to still be viable with vocabularies of up to 20k words when used with bigram back-off language models. This one level implementation includes intra- and inter-word CD phone models, intra- and inter-word phonological rules, phone duration models, and gender-dependent models. For trigram language models, decoding is performed in two forward passes. The first pass generates a word graph using a bigram language model, this graph is then used in a second acoustic pass with the trigram language model. The recognizer has been evaluated in the Nov92 and Nov93 ARPA tests with vocabularies of up to 20,000 words.

Looking at the recognition results for individual speakers, it appears that interspeaker differences are much more important than differences in perplexity. Just considering the relationship between speaking rate and word accurracy, in general, speakers that are faster or slower than the average have a higher word error. It has been observed that the better/worse speakers are the same on both the 5k and 20k tests.

Improving the acoustic modeling, by taking advantage of the available training data, has led to better system performance. By increasing the amount of training utterances from 7k to 37k, reduced the word error by about 30%. In the LIMSI Nov93 system, a trigram LM has been incorporated in a second acoustic pass. The trigram pass gives an error rate reduction of 20% to 30% relative to the bigram system. The combined error reduction is on the order of 50%. Comparable amounts of acoustic data from BREF[6, 13] and text material from the French newspaper *Le Monde* have been used to develop 5k and 20k recognizers for French. Results of evaluating this system were reported at the 1994 ARPA Human Language Technology workshop [5].

References

1. F. Alleva, X. Huang, M.-Y. Hwang, "An Improved Search Algorithm Using Incremental Knowledge for Continuous Speech Recognition," *ICASSP-93*.
2. L.R. Bahl et al, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *ICASSP-92*.
3. S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, **28**(4), 1980.
4. J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, " Speaker-Independent Continuous Speech Dictation," *Eurospeech-93*.
5. J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System," *ARPA Workshop Human Language Technology*, 1994.
6. J.L. Gauvain, L. Lamel, M. Eskénazi, "Design considerations & text selection for BREF, a large French read-speech corpus," *ICSLP-90*.
7. J.L. Gauvain, C.H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, **11**(2-3), 1992.
8. L. Gillick, R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *DARPA Speech & NL Wshop*, 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, "Continuous Speech Recognition at LIMSI," Final review *DARPA ANNT Speech Prog.*, Sep. 1992.
11. L. Lamel, J.L. Gauvain,"High Performance Speaker-Independent Phone Recognition Using CDHMM," *Eurospeech-93*.
12. L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Eurospeech-93*.
13. L. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *Eurospeech-91*.
14. H. Murveit et al, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *ICASSP-93*.
15. 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.
16. D.S. Pallett et al., "Benchmark Tests for the DARPA Spoken Language Program," *ARPA Wshop Human Lang. Tech.*, 1993.
17. D.S. Pallett et al., "1993 Benchmark Tests for the DRPA Spoken Language Program," *ARPA Wshop Human Lang. Tech.*, 1994.
18. D.B. Paul and J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP-92*.
19. R. Schwartz et al.,"New uses for N-Best Sentence Hypothesis within the BYBLOS Speech Recognition System," *ICASSP-92*.