The LIMSI Continuous Speech Dictation 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

A major axis of research at LIMSI is directed at multilingual, speaker-independent, large vocabulary speech dictation. In this paper the LIMSI recognizer which was evaluated in the ARPA NOV93 CSR test is described, and experimental results on the WSJ and BREF corpora under closely matched conditions are reported. For both corpora word recognition experiments were carried out with vocabularies containing up to 20k words. 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 recognizer uses a time-synchronous graph-search strategy which is shown to still be viable with a 20k-word vocabulary when used with bigram back-off language models. A second forward 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.

INTRODUCTION

Speech recognition research at LIMSI aims to develop recognizers that are task-, speaker-, and vocabulary-independent so as to be easily adapted to a variety of applications. The applicability of speech recognition techniques used for one language to other languages is of particular importance in Europe. The multilingual aspects are in part carried out in the context of the LRE SQALE (Speech recognizer Quality Assessment for Linguistic Engineering) project, which is aimed at assessing language dependent issues in multilingual recognizer evaluation. In this project, the same system will be evaluated on comparable tasks in different languages (English, French and German) to determine cross-lingual differences, and different recognizers will be compared on the same language to compare advantages of different recognition strategies.

In this paper some of the primary issues in large vocabulary, speaker-independent, continuous speech recognition for dictation are addressed. These issues include language modeling, acoustic modeling, lexical representation, and search. Acoustic modeling makes use of continuous density HMM with Gaussian mixture of context-dependent phone models. For language modeling n-gram statistics are estimated on text material. To deal with phonological variability alternate pronunciations are included in the lexicon, and optional phonological rules are applied during training and recognition. The recognizer uses a time-synchronous graph-search strategy[16] for a first pass with a bigram back-off language model (LM)[10]. A trigram LM is used in a second acoustic decoding pass which makes use of the word graph generated using the bigram LM[6]. Experimental results are reported on the ARPA Wall Street Journal (WSJ)[19] and BREF[14] corpora, using for both corpora over 37k utterances for acoustic training and more than 37 million words of newspaper text for language model training. While the number of speakers is larger for WSJ, the total amount of acoustic training material is about the same (see Table 1). It is shown that for both corpora increasing the amount of training utterances by an order of magnitude reduces the word error by about 30%. The use of a trigram LM in a second pass also gives an error reduction of 20% to 30%. The combined error reduction is on the order of 50%.

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. In this work bigram and trigram language models are estimated on the training text material for each corpus. This data consists of 37M words of the WSJ[1] and 38M words of *Le Monde*. A backoff mechanism[10] 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 can be used in the first bigram decod-

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

[1] 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[19] as required by ARPA so as to be compatible with the other sites participating in the tests.

ing pass to reduce computational requirements. The trigram langage model is used in the second pass of the decoding process.

In order to be able to construct LMs for BREF, it was necessary to normalize the text material of *Le Monde* newpaper, which entailed a pre-treatment rather different from that used to normalize the WSJ texts[19]. The main differences are in the treatment of compound words, abbreviations, and case. In BREF the distinction between the cases is kept if it designates a distinctive graphemic feature, but not when the upper case is simply due to the fact that the word occurs at the beginning of the sentence. Thus, the first word of each sentence was semi-automatically verified to determine if a transformation to lower case was needed. Special treatment is also needed for the symbols hyphen (-), quote ('), and period (.) which can lead to ambiguous separations. For example, the hyphen in compound words like *beaux-arts* and *au-dessus* is considered word-internal. Alternatively the hyphen may be associated with the first word as in *ex-*, or *anti-*, or with the second word as in *-là* or *-né*. Finally, it may appear in the text even though it is not associated with any word. The quote can have two different separations: it can be word internal (*aujourd'hui*, *o'Donnel*, *hors-d'oeuvre*), or may be part of the first word (*l'ami*). Similarly the period may be part of a word, for instance, *L.A.*, *sec.* (secondes), *p.* (page), or simply an end-of-sentence mark.

Table 1 compares some characteristics of the WSJ and *Le Monde* text corpora. In the same size training texts, there are almost 60% more distinct words for *Le Monde* than for WSJ without taking case into account.[2] As a consequence, the lexical coverage for a given size lexicon is smaller for *Le Monde* than for WSJ. For example, the 20k WSJ lexicon accounts for 97.5% of word occurrences, but the 20k BREF lexicon only covers 94.9% of word occurrences in the training texts. For lexicons in the range of 5k to 40k words, the number of words must be doubled for *Le Monde* in order to obtain the same word coverage as for WSJ.

The lexical ambiguity is also higher for French than for English. The homophone rate (the number of words which have a homophone divided by the total number of words) in the 20k BREF lexicon is 57% compared to 9% in 20k-open WSJ lexicon. This effect is even greater if the word frequencies are taken into account. Given a perfect phonemic transcription, 23% of words in the WSJ training texts is ambiguous, whereas 75% of the words in the *Le Monde* training texts have an ambiguous phonemic transcription. Not only does one phonemic form correspond to different orthographic forms, there can also be a relatively large number of possible pronunciations for a given word. In French, the alternate pronunciations arise mainly from optional word-final phones, due to liaison and optional word-final consonant cluster reduction (see Figure 1). There are also a larger number of frequent, monophone words for *Le Monde* than for WSJ, accounting for about 17% and 3% of all word occurrences in the respective training texts.

Corpus	*WSJ*	*Le Monde*
# training speakers	284	80
# training utterances	37.5k	38.5k
Training text size	37.2M	37.7M
#distinct words	165k	259k (280)
5k coverage	90.6%	85.5% (85.2)
20k coverage	97.5%	94.9% (94.7)
Homophone rate 20k lexicon	9%	57%
Homophone rate 20k text	23%	75%
Monophone words (20k)	3%	17%

Table 1: Comparison of WSJ and BREF corpora.

ACOUSTIC-PHONETIC 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 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. In the experimental section we demonstrate the improvement in performance obtained on the same test data by simply using additional training material.

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 and 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 cosinus transform [2].

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. The models include triphone models, right- and

[2] If case is kept when distinctive, there are 280k words in the *Le Monde* training material.

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). 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 posteriori estimators are used for the HMM parameters[8] and moment estimators for the gamma distributions. Separate male and female models are used to more accurately model the speech data.

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[12]. 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.

LEXICAL REPRESENTATION

Lexicons containing 5k, 20k, and 64k words have been used in these experiments. The lexicons are represented phonemically, using language-specific sets of phonemes. Each lexicon has alternate pronunciations for some of the words, and allows some of the phones to be optional.[3] 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 WSJ lexicons are represented using a set of 46 phonemes, including 21 vowels, 24 consonants, and silence. 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. The BREF lexicons, corresponding to the 5k and 20k most common words in the *Le Monde* texts are represented with 35 phonemes including 14 vowels, 20 consonants, and silence[3]. The base pronunciations, obtained using text-to-phoneme rules[20], were extended to annotate potential liaisons and pronunciation variants. Some example lexical entries are given in Figure 1.

Word boundary phonological rules are applied in building the phone graph used by the recognizer so as to allow for some of the phonological variations observed in fluent speech[11]. 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. Using optional phonological rules during training results in better acoustic models, as they are less "polluted" by wrong transcriptions. Their use during recognition reduces the number of mismatches. For English, only well known phonological rules, such as glide insertion, stop deletion, homorganic stop insertion, palatalization, and voicing assimilation have been incorporated in the system. The same mechanism has been used to handle liaisons, mute-e, and final consonant cluster reduction for French.

Example entries for WSJ:

INTEREST	IntrIst In{t}XIst
EXCUSE	Ekskyu[sz]
CORP.	kcrp kcrpXeSxn
GAMBLING	g@mb[Ll]|G
AREA	[@e]rix $\xrightarrow{ph.rule}$ [@e]riyx

Example entries for BREF:

sont	sO sOt(V)
les	le(C.) lez(V)
mon	mO mOn(V)
ma	ma(C.)
autres	ot(C.) otrx otr(V) otrxz(V)

Figure 1: Example lexical entries for WSJ and BREF. Phones in {} are optional, phones in [ ] are alternates. () specify a context constraint and **V** stands for vowel, **C** for consonant and the period represents silence.

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 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 [16] which uses a dynamic programming procedure. This basic strategy has been recently extended by adding other features such as "fast match"[9, 1], N-best rescoring[21], and progressive search[15]. The two-pass approach used in our system is based on the idea of progressive search where the information between levels is transmitted via word graphs. Prior to word recognition, sex identification is performed for each sentence using phone-based ergodic HMMs[13]. The word recognizer is then run with a bigram LM using the acoustic model set corresponding to the identified sex.

The first pass 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, intra- and inter-word phonological rules, phone duration models, and gender-dependent models, generates a list of word hypotheses resulting in a word lattice. Two problems need to be considered

[3] About 10% of the lexical entries have multiple transcriptions, if the word final optional phonemes marking possible liaisons for BREF are not included. Including these raises the number of entries with multiple transcriptions to almost 40%.

at this level. 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 "optimal" lattice to be used with a trigram LM. For example, any given word in the lattice will have many possible ending points, but only a few starting points. This problem was in fact less severe than expected since the time information is not critical to generate an "optimal" word graph from the lattice, i.e. the multiple word endings 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 used to control the lattice size.

A description of the exact procedure used to generate the word graph from the word lattice is beyond the scope of this paper. The following steps give the key elements behind the procedure.[4] 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.

To fix these ideas, let us consider some numbers for the WSJ 5k-closed vocabulary. With the pruning threshold set at a level such that there are only a negligible number of search errors, the first pass generates a word lattice containing on average 10,000 word hypotheses per sentence. The generated word graph before trigram expansion contains on average 1400 arcs. After expansion with the 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 due to memory constraints.

EXPERIMENTAL RESULTS

WSJ: The ARPA WSJ corpus[19] was designed to provide general-purpose speech data with large vocabularies. Text materials were selected to provide training and test data for 5k and 20k word, closed and open vocabularies, and with both verbalized (VP) and non-verbalized (NVP) punctuation.

5k - WSJ	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
Nov92, si84, 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, si84, 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 2: 5k results - Word recognition results on the WSJ corpus with bigram/trigram (bg/tg) grammars estimated on WSJ text data.

20k - WSJ	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
Nov92, si84c, bg	88.3	10.1	1.5	2.0	13.6
Nov92+, si84c, 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 3: 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.

For testing purposes, the 20k closed vocabulary includes all the words in the test data whereas the 20k open vocabulary contains only the 20k most common words in the WSJ texts. The 20k open test is also referred to as a 64k test since all of the words in these sentences occur in the 63,495 most frequent words in the normalized WSJ text material[19]. Two sets of standard training material have been used for these experiments: The standard WSJ0 SI84 training data which include 7240 sentences from 84 speakers, and the standard set of 37,518 WSJ0/WSJ1 SI284 sentences from 284 speakers. Only the primary microphone data were used for training.

The WSJ corpus provides a wealth of material that can be used for system development. 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 of the word recognizer 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 WSJ system was evaluated in the Nov92 ARPA evaluation test[17] for the 5k-closed vocabulary and in the Nov93 ARPA evaluation test[18] for the 5k and 64k hubs. Except when explicitly stated otherwise, all of the results reported for WSJ use the standard language models[19]. Using a set of 1084 CD models trained with the WSJ0 si84 training data, the word error is 6.6% on the Nov92 5k test data and 9.4% on the Nov93 test data. Using the combined WSJ0/WSJ1 si284 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 an addition 35% on the Nov92 test and by 22% on the Nov93 test.

Results are given in the Table 3 for the Nov92 nvp 64K

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

test data using both closed and open 20k vocabularies. With si84 training (si84c, a slightly smaller model set than si84) 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 higher error rate with the 20k open lexicon can be largely attributed to the out-of-vocabulary (OOV) words, which account for almost 2% of the words in the test sentences. Processing the same test data with a system trained on the si284 training data, reduces the word error by 30%. The word error on the Nov93 20k test is 15.2% with the si284 system. Using the trigram LM reduces the error rate by 18% on the Nov92 test and 22% on the Nov93 test.

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 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 6% and 12% 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 an attempt to eliminate 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 hypothesis of the trigram 20k system, an A^* algorithm is used to generate a word graph using phone confusion statistics and the 64k lexicon. This word graph is then used by the recognizer with a 64k trigram LM trained on the standard WSJ training texts (37M words). Using this approach only about 30% of the errors due to OOV words on the Nov93 64k test are recovered, reducing the word error to 11.2% from 11.8%.

BREF: BREF[14] is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f). The text materials were selected verbatim from the French newspaper *Le Monde*, so as to provide a large vocabulary (over 20,000 words) and a wide range of phonetic environments[7]. The material in BREF was selected to maximize the number of different phonemic contexts.[5] Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train vocabulary-independent acoustic models. The text material was read without verbalized punctuation using the verbatim prompts.[6]

5k - BREF	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
Feb94, si57, bg	88.7	7.5	3.7	1.4	12.6
Feb94, si80, bg	92.0	5.9	2.1	1.1	9.1
Feb94, si80, tg	95.2	3.7	1.1	1.0	5.8

Table 4: 5k word recognition results on the Feb94 test data with bigram/trigram grammars estimated on *Le Monde* text data.

20k - BREF	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
Feb94, si57, bg	85.5	11.9	2.6	1.8	16.3
Feb94, si80, bg	88.6	9.7	1.7	1.6	13.0
Feb94, si80, tg	91.6	7.5	0.9	1.2	9.6
Feb94+, si80, bg	84.6	14.2	1.3	4.6	20.0
Feb94+, si80, tg	87.4	11.6	1.0	4.3	16.9

Table 5: 20k word recognition results on the Feb94 test data with bigram/trigram grammars estimated on *Le Monde* text data. +: 20k word lexicon with open test.

We have previously reported results using only a small portion (2770 sentences from 57 speakers) of the available training material for BREF[3, 5, 4]. In these experiments, the amount of training data has been extended to 38,550 sentences from 80 speakers. The amount of text material used for LM training has also been increased to 38M words, enabling us to estimate trigram LMs. Vocabularies containing the most frequent 5k and 20k words in the training material are used and bigram and trigram LMs were estimated for both vocabularies. 200 test sentences (25 from each of 8 speakers) for each vocabulary were selected from the development test material for a closed vocabulary test. The perplexity of all the within vocabulary sentences of the development test data using the 5k/20k LM is 106/178 (which can be compared to 96/196 for WSJ computed under the same conditions with the 5k/20k-open LM). An additional 200 sentences were used for a 20k-open test set. As ensured by the prompt selection process, the prompt texts were distinct from the training prompts.

Word recognition results for the 5k test are given in Table 4 with bigram and trigram LMs estimated on the 38M-word normalized text material from *Le Monde*. With 428 CD models trained on the si57 sentences, the word error is 12.6%. Using an order of magnitude more training data (si80) and 1747 CD models, the word error with the bigram is reduced by 28% to 9.1%. The use of a trigram LM gives an additional 36% reduction of error.

Results for the 20k test are given in Table 5 using the same acoustic model sets and LMs, for both closed and open vocabulary test sets. For the closed vocabulary test, the si80 training data gives an error reduction of 20% over the si57

[5] This is in contrast to the WSJ texts which were selected so as to contain only words in the most frequent 64,000 words in the original text material.

[6] Another difference between BREF and WSJ0 is that the prompts for WSJ0 were normalized, where for BREF the prompts were presented as they appeared in the original text. This latter approach has since been adopted for the recordings of WSJ1. However, while for WSJ1 orthographic transcriptions are provided, for BREF the only reference currently available is the prompt text.

training. The use of the trigram LM reduces the word error by an additional 26%. The 20k-open test results are given in the lower part of the table. 3.9% of the words are OOV and occur in 72 of the 200 sentences. We observe almost a 50% increase in word error, with a three-fold increase in the word insertions compared with the closed vocabulary test. Thus apparently the OOV words are not simply replaced by another word, but are more often replaced by a sequence of words. The trigram LM only reduces the word error by 15% on this test.

DISCUSSION AND SUMMARY

The recognizer has been evaluated on 5k and 20k test data for the English and French languages using similar style corpora. It should be pointed out however, that although the Nov92 5k WSJ test data and the BREF 5k test data were closed-vocabulary, the conditions are not quite the same. For WSJ, paragraphs were selected ensuring not more than one word was out of the 5.6k most frequent words[19], and these additional words were then included as part of the vocabulary. For BREF, a lexicon was first constructed containing the 5k/20k most frequent words, and sentences covered by this vocabulary were selected from the development test material. The situation was slightly different for the Nov93 5k test in that the prompt texts were not normalized, and therefore several OOV words (0.3%) occurred in the test data despite it being a closed-vocabulary test.

However, looking at the recognition results for individual speakers, it appears that interspeaker differences are much more important than differences in perplexity, and perhaps more than language differences. 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.

We have observed some language dependencies, such as the higher number of homophones in BREF, which has the effect of reducing the efficiency of the search and the large number of frequent monophone words which results in larger networks. At the same time, the phone accuracy for BREF is better than that for WSJ, which speeds up the search.

Improving the model accuracy, at the acoustic level and at the language model level, by taking advantage of the available training data, has led to better system performance. For both WSJ and BREF increasing the amount of training utterances by an order of magnitude reduces the word error by about 30%. By using larger training text materials it is possible to train a trigram LM which was incorporated in a second acoustic pass. The trigram pass gives an error rate reduction of 20% to 30%. The combined error reduction is on the order of 50%.

It remains a general problem of how to define comparable test conditions for different languages. This may even depend on the definition of a word in a given language, which is linked to the lexical coverage. A primary aim of the aforementioned LRE Sqale project is to address this issue.

REFERENCES

[1] L.R. Bahl et al, “A Fast Match for Continuous Speech Recognition Using Allophonic Models,” *ICASSP-92*.

[2] S.B. Davis, P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. ASSP*, **28**(4), 1980.

[3] J.L. Gauvain, L.F. Lamel, “Speaker-Independent Phone Recognition Using BREF,” *DARPA S&NL Workshop*, 1992.

[4] J.L. Gauvain, L.F. Lamel, G. Adda, J. Mariani, “Speech-to-Text Conversion in French,” *Int. J. Pat. Rec. & A.I.*, 1994.

[5] J.L. Gauvain et al., “ Speaker-Independent Continuous Speech Dictation,” *Eurospeech-93*.

[6] J.L. Gauvain et al., “The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task,” *ICASSP-94*.

[7] J.L. Gauvain, L.F. Lamel, M. Eskénazi, “Design considerations & text selection for BREF, a large French read-speech corpus,” *ICSLP-90*.

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

[9] L. Gillick, R. Roth, “A Rapid Match Algorithm for Continuous Speech Recognition,” *DARPA Sp&NL Workshop*, 1990.

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

[11] L. Lamel, J.L. Gauvain, “Continuous Speech Recognition at LIMSI,” Final review *DARPA ANNT Speech Prog.*, Sep. 1992.

[12] L. Lamel, J.L.Gauvain,“High Performance Speaker-Independent Phone Recognition Using CDHMM,” *Eurospeech-93*.

[13] L. Lamel, J.L. Gauvain, “Identifying Non-Linguistic Speech Features,” *Eurospeech-93*.

[14] L.F. Lamel, J.L. Gauvain, M. Eskénazi, “BREF, a Large Vocabulary Spoken Corpus for French,” *Eurospeech-91*.

[15] H. Murveit et al, “Large-Vocabulary Dictation using SRI’s Decipher Speech Recognition System: Progressive Search Techniques,” *ICASSP-93*.

[16] 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.

[17] D.S. Pallett et al., “Benchmark Tests for the DARPA Spoken Language Program,” *ARPA HLT Workshop*, 1993.

[18] D.S. Pallett et al., “1993 Benchmark Tests for the ARPA Spoken Language Program,” *ARPA HLT Workshop*, 1994.

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

[20] B. Prouts,“Contribution à la synthèse de la parole à partir du texte: Transcription graphème-phonème en temps réel sur microprocesseur", Thèse de docteur-ingénieur, Université Paris XI, Nov. 1980.

[21] R. Schwartz et al.,“New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System,” *ICASSP-92*.