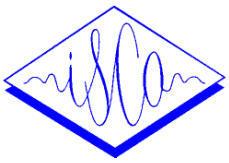

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Continuous Speech Dictation in French †

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 research activity at LIMSI is multilingual, speaker-independent, large vocabulary speech dictation. In this paper we report on efforts in large vocabulary, speaker-independent continuous speech recognition of French using the BREF corpus. 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 38 million words of newspaper text from *Le Monde* for language modeling. The recognizer uses a time-synchronous graph-search strategy. When a bigram language model is used, recognition is carried out in a single forward pass. A second forward pass, which makes use of a word graph generated with the bigram language model, incorporates a trigram language model. Acoustic modeling uses cepstrum-based features, context-dependent phone models and phone duration models. An average phone accuracy of 86% was achieved. A word accuracy of 84% has been obtained for an unrestricted vocabulary test and 95% for a 5k vocabulary test.

INTRODUCTION

Our current efforts focus on speech-to-text conversion of continuously spoken sentences, from any speaker, for very large vocabularies (eventually, unlimited). Because of the ambitiousness of the task, the acoustic models should be both independent of the speaker and the vocabulary. To this extent, a phone-based approach is being used, where phone-like units are trained with data from a large number of speakers. The applicability of speech recognition techniques for different languages is of particular importance in Europe, and multilingual speech recognition is one of LIMSI's active research areas. For French this research heavily relies on the BREF speech corpus[2, 8] for acoustic model training and 50 million words of text from the French newspaper *Le Monde* for the language model training material.[1]

In this paper we address some of the primary issues in large vocabulary, speaker-independent, continuous speech dictation including acoustic modeling, language modeling, lexical representation, and search strategy. 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. The recognizer uses a time-synchronous graph-search strategy[12] for a first pass with a bigram back-off language model (LM)[7]. A trigram LM is used in a second acoustic decoding pass which makes use of the word graph generated using the bigram LM[4]. Experimental results are reported for vocabularies of 5k and 20k words and for two training conditions.

SPEECH AND TEXT CORPUS

BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[8]. 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[2]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train vocabulary-independent phone models. Separate text materials, with similar distributional properties were selected for training, development test, and evaluation purposes. Each of 80 speakers read approximately 10,000 words (about 650 sentences) of text, and an additional 40 speakers each read about half that amount. Two channel recordings (Shure SM10, Crown PCC160) were made in a sound-isolated room, and were monitored to assure their contents. The text material was read without verbalized punctuation.

We have previously reported results using a small portion of the available training material for BREF, 2770 sentences from 57 speakers (si-3k)[3]. In this paper experimental results are also reported using a much larger amount of training data, 38,550 utterances from 80 speakers (si-38k).

The *Le Monde* text materials that are at our disposal comprise a total of about 38M words of text. This includes the 4M words of text material used to select prompts for spoken corpus (Sep89, Oct89, Jan90) as well as 34M words from the years 1992 and 1993 which are available on CDROM. The text materials were preprocessed and normalized for further use. This preprocessing concerned mainly compound words, abbreviations, and case. The symbols for hyphen, quote, and period may be part of a compound word, may be associated with one of the words or may appear in the text even if it is not part of any word. Since the case distinction is kept only when it designates a distinctive graphemic feature, the first word of each sentence was semi-automatically verified to determine if a transformation to lower case was needed.

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

[1] Most of our LV, CSR research in English focuses on the ARPA Wall Street Journal task[5, 4].

10.21437/ICSLP.1994-534

Le Monde texts	
Training text size	37.7M
#distinct words	259k (280)
5k coverage	85.5% (85.2)
20k coverage	94.9% (94.7)

Table 1: Some characteristics of *Le Monde* text material. The numbers in parentheses are when upper and lower case are distinguished.

Some characteristics of the text materials are given in Table 1. There are 259k distinct words in the 37.7M-word training texts without distinguishing case and 280k distinct words if case is kept when it is graphemically distinctive. The word coverage of the training text material is 85.2% for the 5k most frequent words and 94.7% for the 20k most frequent words.

RECOGNIZER DESCRIPTION

Acoustic-Phonetic Modeling

A feature vector containing 16 Bark-frequency scale cepstrum coefficients (8kHz bandwidth) and their first and second order derivatives is computed every 10 ms.

Sets of context-dependent(CD), position independent phone models are trained, where the contexts (intra-word and cross-word) are automatically selected based on their frequencies in the training data. The models include triphone models, right- and left-context phone models, and context-independent phone models. Each phone model is a 3-state, left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). 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.

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. A backoff mechanism[7] 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. An added advantage of the backoff mechanism is that LM size can be 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 first bigram decoding pass to reduce computational requirements. The trigram langage model is used in the second pass of the decoding process. These bigram and trigram language models were estimated on the 38M word training texts.

```
les             le(C.) lez(V)
mon             mO mOn(V)
ma              ma(C.)
soixante-dix    swasA[tn]di(C) swasA[tn]dis
                swasA[tn]diz(V)
contenu         kOt{x}ny
autres          ot(C.) otrx otr(V) otrxz(V)
```

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

Lexical Representation

Lexicons containing the most frequent 5k and 20k words in the *Le Monde* training texts have been created. The base pronunciations were obtained using text-to-phoneme rules[13] and extended semi-automatically to annotate potential liaisons and pronunciation variants. Alternate pronunciations are given for about 10% of the words.[2] Some example lexical entries are given in Figure 1. The mechanism developed to handle word boundary phonological rules in English[9] is used to deal optional liaisons, mute-e, and final consonant cluster reduction for French.

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[12] which uses a dynamic programming procedure. This basic strategy has been recently extended by adding other features such as "fast match"[6, 1], N-best rescoring[14], and progressive search[11]. 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[5, 4]. This decoding strategy, which is has two forward passes can be implemented in a single forward pass. A two pass solution has been chosen because it is conceptually simpler and requires less memory.

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 is used to generate a word lattice. Two concerns are the size of the lattice generated (if it is too large there is no interest in a two pass approach) and the optimality of the generated lattice. We use two pruning thresholds 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 word graph then is generated from the word lattice by iteratively merging "similar" graph nodes to reduce the overall graph size and at the same time generalizing the

[2]This count does not include word final optional phonemes marking possible liaisons. Including these raises the number of entries with multiple transcriptions to about 40%.

word lattice.

To fix these ideas, let us consider some numbers for the 20k-closed vocabulary test data. 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 14,000 word hypotheses per sentence. The generated word graph before trigram expansion contains on average 1600 arcs. After expansion with the trigram backoff LM, there are on average 4700 word instanciations including silences which are treated the same way as words.

EXPERIMENTAL RESULTS

Recognition vocabularies containing the most frequent 5k and 20k words in the training text material are used. 200 test sentences (25 from each of 8 speakers) for each vocabulary were selected from the development test material (Feb94-dev) for a closed vocabulary test. The 5k bg perplexity is 106 and the 20k bg perplexity is 178. An additional 200 sentences from the development material were used for a 20k-open test set.

We make use of phone recognition to evaluate different acoustic model sets which are then used for word recognition. The phone errors rates are given in the first column of Tables 2 and 3 using a phone bigram to provide phonotactic constraints. The reduction in phone error rate is seen to be about 29% using the si-38k training data for both the 5k and 20k test data. We have previously reported that improvements in phone accuracy are directly indicative of improvements in word accuracy when the same phone models are used for recognition[10].

Word recognition results for the 5k Feb94-dev test are given in Table 2 with bigram and trigram LMs. With 428 CD models trained on the si-3K sentences, the word error is 12.6%. Using 1747 CD models trained on the si-38k data the word error with the bigram is reduced by 29% to 9.0%. The use of a trigram LM gives an additional 39% reduction of error to 5.5%

Results for the 20k Feb94-dev test are given in Table 3. For the closed vocabulary test, the si-38k model set gives an error reduction of 21% over the si-3K model set. The use of the trigram LM reduces the word error by an additional 26%. In the 20k open (20k+) test 3.9% of the words are out-of-vocabulary (OOV) and occur in 72 of the 200 sentences. For the open vocabulary test, not only are there more substitutions (11.1% vs 7.3%) but the insertion rate is much higher (4.3% vs 1.0%). Thus apparently the OOV words are not simply replaced by another word, but are more often replaced by a sequence of words. For example, the word *endeuillé*, which is not in the lexicon, was recognized as the sequence of words *en deuil et*, which has the sequence of phonemes. Due to the OOV words the use of a trigram LM only reduces the word error by 16% for the open vocabulary condition.

Since a word graph is used to provide information to the trigram pass, there are evidently errors that cannot be recovered. The graph error rate (ie. the correct solution was not in the graph) was 4.5% and 11% respectively for the 5k and 20k closed tests. For the 20k open test the graph error was 10% on the 128 sentences without OOV words.

BREF - 5k	*Phone Error*	*Word Error*
si-3k, bg	20.0	12.6
si-38k, bg	14.2	9.0
si-38k, tg	-	5.5

Table 2: Word recognition results on the Feb94-dev 5k test with bigram (bg) and trigram (tg) LMs estimated on *Le Monde* text data.

BREF - 20k	*Phone Error*	*Word Error*
20k, si-3k, bg	19.6	16.3
20k, si-38k, bg	13.8	12.9
20k, si-38k, tg	-	9.2
20k+, si-38k, bg	14.3	19.5
20k+, si-38k, tg	-	16.4

Table 3: Word recognition results on the Feb94-dev 20k test with bigram (bg) and trigram (tg) LMs estimated on *Le Monde* text data. 20k+ stands for open test with nearly 4% OOV words.

DISCUSSION AND SUMMARY

French is a language with a high lexical ambiguity which is in part due to the large number of homophones (words having the same pronunciation). 57% of the words in the 20k-word BREF training are homophones, and 3 out of 4 words in the training text have a homophone. Not only are there many words which are homophones, the size of the homophone class can be relatively large. For example, there are 8 words in 20k lexicon with the same pronunciation /sA/: *100, cent, cents, san, sang, sans, sens, sent.* Most of the homophones arise from verb conjugation, the mark of plurals (-s) and feminine form (-e) that are often not pronounced. If the most common single word homophone errors are not counted, then the word error of the trigram runs are reduced by over 40% for the closed vocabulary test. This difference in word error highlights the need for better language modeling.

Not only does one phonemic form correspond to different orthographic forms, there can also be a relatively large number of pronunciations for a given word. For example, in Figure 1 the word "contenu" may be pronounced with 2 or 3 syllables, and the word "autres" has 4 possible pronunciations: /ot/, /otrx/, /otr/, /otrxz/. These alternate pronunciations, which arise mainly from optional liaison consonants and optional word-final consonant cluster reduction are all possible, but not equally likely, depending on the speaker, the dialect, the neighboring phones and words, and sometimes on the semantics. Using probabilities for each transcription can be useful, but their automatic training is not straightforward and requires a lot of data.

In our analysis of recognition errors, we have consistently noted that a large number of errors involve short words of one or two phonemes. While there are relatively few of these words in the lexicon, they are very frequent in running text. Almost 20% of the words in the training text

are monophone words and about 50% of all word occurrences have at most two phonemes. The monophone words may exhibit high acoustic variability, have no intraword phonotactic constraints, and low LM costs as they are very frequent. We have observed that nearly any word sequence can be transcribed by a larger number of short, frequent words, resulting in multiword homophones. Some examples of recognition errors where the longer word has been split in a sequence of shorter words, with no or minor errors in the phonetic transcription are "*désengagement→des engagements*", "*couteaux→coûts taux*", and "*il laisse→il et se*". This arises because the most frequent words (in particular the monophone words) have better backoff LM scores, and thus appear easily in place of acoustically similar words which had fewer observations in the text.

Another major source of error involves the insertion or deletion of mute-e. If all possible optional word-final mute-e are permitted, adjacent monophone words may easily be deleted. In contrast, if the lexicon does not allow a mute-e at the end of a word, the system has the tendency to insert a short word when the mute-e is pronounced. For example, the pronunciation of the word *Bankok* in the lexicon does not have a final mute-e and when a speaker produced it, the system made an error and recognized *Bankok que*. This shows the importance of accurate phonemic transcriptions in the lexicon or the means of predicting such phonological variation using rules.

The use of a trigram LM improves the recognition accuracy by 20% to 30% over a bigram LM as it is better able to model agreement of gender and number, and negation. In French a negative form is usually made by surrounding the verb with "*ne* VERB *pas*". While with the bigram the "*ne*" can be easily deleted, the trigram is able capture this constraint. The use of N-class language models (as opposed to N-grams) can be helpful, as for French we often have a high number of different graphemic forms for a given root form.

Concerning the efficiency of the search, the high number of homophones, the large number of possible word endings and the large number of frequent monophone words results in large recognition graphs, increasing the search time compared to similar experiments in English[4].

In this paper we have described our speaker-independent, continuous speech recognition system, and given experimental results using the BREF corpus of French with vocabularies of up to 20k words. The recognizer uses a time-synchronous graph-search strategy which includes intra- and inter-word context-dependent phone models, phonological rules, and a bigram language model. When a trigram language model is used, it is incorporated in a second forward pass which makes use of a word graph generated with the bigram language model. 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. Increasing the amount of training utterances by an order of magnitude reduces the word error by about 30%. Using larger training text materials it is possible to train a trigram language model which is used in a second acoustic pass, achieving an additional error reduction of 20% to 30%. The combined error reduction is on the order of 50%. The resulting phone accuracy is 86%. For an unrestricted vocabulary test, the word accuracy is 84% and for a 5k vocabulary test, 94.5%.

REFERENCES

[1] L. Bahl, P. de Souza, P. Gopalakrishanan, D. Nahamoo, M. Picheny, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *ICASSP-92*.

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

[3] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Eurospeech-93*.

[4] 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," *ICASSP-94*.

[5] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System," *ARPA Workshop Human Language Technology*, 1994.

[6] L. Gillick, R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *DARPA Speech & Natural Language Workshop*, 1990.

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

[8] L. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *Eurospeech-91*.

[9] L.F. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," Final review of the *DARPA ANNT Speech Program*, Sep. 1992.

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

[11] H. Murveit, J. Butzberger, V. Digalakis, M. Weintraub, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *ICASSP-93*.

[12] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. ASSP*, **32**(2), April 1984.

[13] 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, U. Paris XI, Nov. 1980.

[14] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, G. Zavaliagkos,"New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System," *ICASSP-92*.