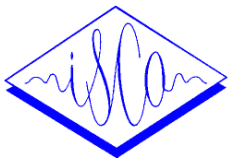

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

Speaker-Independent Continuous Speech Dictation

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

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

ABSTRACT

In this paper we report progress made at LIMSI in speaker-independent large vocabulary speech dictation using newspaper speech corpora. 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. Acoustic modeling uses cepstrum-based features, context-dependent phone models (intra and interword), phone duration models, and sex-dependent models. Two corpora of read speech have been used to carry out the experiments: the DARPA Wall Street Journal-based CSR corpus and the BREF corpus containing recordings of texts from the French newspaper *Le Monde*. For both corpora experiments were carried out with up to 20K word lexicons. Experimental results are also given for the DARPA RM task which has been widely used to evaluate and compare systems.

INTRODUCTION

Our speech recognition work focuses on developing recognizers that are task-, speaker- and vocabulary-independent so as to be easily adapted to various applications. These characteristics advocate an approach based on sub-word units, such as phone-like units. The primary research axes are a dictation task and a dialog project. In this paper we report on recent efforts in large vocabulary, speaker-independent continuous speech recognition for English and French. Three corpora have been used to carry out the experiments: the DARPA Resource Management corpus (RM)[23], the DARPA Wall Street Journal-based CSR corpus (WSJ0)[20], and the BREF-*Le Monde* corpus[17]. All three corpora contain large amounts of read speech material from a large number of speakers, recorded under similar conditions (8kHz bandwidth, close-talking microphone, read-speech). WSJ and BREF also have associated text materials which can be used as a source for statistical language modeling. For these two corpora, experiments with comparable size lexicons and test perplexities have been carried out to enable language-dependent performance issues to be addressed. The recognizer has been evaluated in the September 1992 DARPA continuous speech recognition evaluation on the 1000-word Resource Managment task[21] and also in the DARPA Wall Street Journal evaluation in November 1992[22].

This paper is organised as follows. In the next section the recognizer is described, with an emphasis on the characteristics which are different from other HMM-based systems. The following three sections present an evaluation of the system on each of the RM, WSJ, and BREF corpora, including description of the corpus and task specific details. The final section discusses some of the problems in the dictation task, highlights some language-dependent differences, as well as indicates directions for future development.

RECOGNIZER OVERVIEW

The recognizer uses a time-synchronous graph-search strategy as opposed to some recently developed multi-pass approaches for large vocabulary continuous speech recognition[25, 1, 19]. Our experiments show that the time-synchronous approach is still 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, gender-dependent models, and a bigram language model[14, 5]. 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 the phone models for each word, according to its phone transcription as found in the lexicon.

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 cosinus transform [3].

Acoustic models: The acoustic models are sets of context-dependent (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. The HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. Maximum a posteriori (MAP) estimators are used for the HMM parameters[10] and moment estimators for the gamma distributions. Separate male and female models are used to more accurately model the speech data.

Lexicon: The lexicon is represented phonemically, with different lexicons for each task. 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 some of the phonological variations observed in fluent speech.

Language Model: Bigram-backoff[13] language models are estimated on the training text material for WSJ and BREF. The backoff mechanism has been efficiently implemented using a tree. The LM size can be arbitrarily reduced by relying more on the backoff. For RM, the standard deterministic word-pair grammar was used.

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. Decoding consists of Viterbi search, a one pass beam search. The male and female models are run in parallel, and the output with the highest likelihood is chosen.

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. 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. The mechanism for the phonological rules allows the potential for generalization and extension. A pronunciation graph is generated for each word from the baseform transcription to which word internal phonological rules are applied. In forming the word network, word boundary phonological rules are applied at the phone level to take into account interword phonological variations such as palatalization, voicing assimilation, or glide insertion for English. The same mechanism has been used to handle liaisons, mute-e, and final consonant cluster reduction for French.

Much of system development has been carried out by performing phone recognition instead of word recognition in order to reduce the computational requirements and speed up the development process. We have shown that improvements in phone accuracy are directly indicative of improvements in word accuracy when the same phone models are used for recognition[5, 16]. This has allowed us to evaluate many alternatives for the front-end and the acoustic models. Phone recognition provides the added benefit that the recognized phone string can be used to understand errors in word recognition, and problems with the lexical representation.

EXPERIMENTS USING RM

The DARPA RM speech corpus[23] is a corpus of read speech designed to provide speech data for evaluation of continuous speech recognizers with medium size vocabulary (1000 words) and has been widely used in comparative evaluations. The standard set of 3990 sentences (SI-109) has been used to train two sets of 2300 CD phone models, from the male and female speakers data. The standard word-pair grammar (perplexity 60) was used.

The JUN88, FEB89, and OCT89 SI test sets were used as development data to evaluate various alternatives for the front end, for the representation of the lexicon, phonological rules, and to estimate some parameter values such as the word insertion penalty. The data sets were then complemented with SD-DEV and SD-EVAL data (for a total of 2700 sentences) that were used to analyse the most common errors, add alternate pronunciations to the baseline lexicon and create some task-specific phonological rules. This error analysis was not only based on the word recognizer output but also on the phone recognizer output[14, 16]. The FEB91 test data was reserved for evaluation at the end of each development cycle.

The RM lexicon is represented with a set of 36 phones. This reduced set was used primarily to eliminate infrequent phones for which there was insufficient training data, and to provide a means of better sharing contexts. In doing so, more data is available to train the remaining models, and the number of potential triphone contexts is reduced. The infrequent phones /Z, U/ were eliminated and replaced by another "close" phone. The diphthongs /Y, O, W/ were represented by a sequence of phones. Allophonic distinctions such as the syllabics, the context-dependent difference between the two schwas /x, |/, and the stress difference between /X, R/, were no longer made. Care was taken to ensure that these changes did not create any new homophones in the lexicon. Reducing the phone set gave an improvement of about 10% on the 3 development tests.

The lexicon was also expanded to provide alternate pronunciations for about 10% of the words, and to allow some phones to be optional. For example, the word MONTICELLO has the pronunciations /mantxsElo/ and /mantxtSElo/, and the /t/ in COUNTED (/kawn{t}xd/) is optional. Intra- and inter-word phonological rules are optionally applied during training and recognition. The use of phonological rules for the RM task has been previously reported by SRI[2] and AT&T[11]. In the case of AT&T, phonological rules were used only with CI phone models. A single speaker may mark phonetic distinctions in different ways even in similar phonetic environments. This means that the use of CD phones as they are typically defined, combines allophones which can be acoustically very different. The use of phonological rules during training should result in purer acoustic models, and thus improve the system performance.

DARPA test	*Corr.*	*Subs.*	*Del.*	*Ins.*	*WErr.*
JUN88	97.1	2.5	0.4	0.4	3.3
FEB89	97.7	1.7	0.5	0.2	2.5
OCT89	97.0	2.2	0.9	0.3	3.3
FEB91	97.7	1.9	0.4	0.3	2.6
SEP92*	96.0	2.9	1.2	0.4	4.4

Table 1: Word recognition results on the DARPA-RM-SI corpus with a WP grammar of perplexity 60. (*official DARPA SEP92 evaluation results)

Some examples of the phonological rules used for the RM task are given in [14]. These include general rules for well known variants such as palatalization, glide insertion and gemination, as well as rules to handle allophonic variation, using only the reduced phone set. Since the CD models are position independent, instead of having a syllable- or word-final allophones for the voiceless stops, they are optionally allowed to be replaced with their voiced counterparts. Some more specific rules allow the deletion of the offglide /w/ in the phone sequence /aw/, in certain contexts. While this is a fairly general phenomenon, in the context of RM this rule becomes very specific for the word sequences "how much" and "how many."

The developmental changes based on the error analysis provided an 18% reduction on the word error rate measured on the development data [14]. Results on the last 5 DARPA tests are reported in Table 1. After the Sep92 DARPA test, the contribution of the system components to the performance on the Sep92 test data was assessed indicated that the interword phonological rules and the sex-dependent models had the largest influence in reducing the word error[14].

EXPERIMENTS USING WSJ

The DARPA WSJ corpus[20] 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. The standard WSJ0 SI-84 training data include 7240 sentences from 84 speakers. The language model is a bigram-backoff estimated on the 33 million word standardized WSJ text provided by Lincoln Labs[20]. The lexicon is represented using a set of 46 phones. The pronunciations were obtained from various existing lexicons (TIMIT, Pocket and Moby), missing forms were generated by rule when possible, or added by hand. Some of the missing proper names were transcribed by the ORATOR system of Bellcore. In manually verifying the transcriptions, optional and/or alternate phonemes were added.

Phonological rules are optionally applied during training and test. 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.

This system was evaluated in the Nov92 DARPA evaluation test for the 5k-closed vocabulary using the standard bigram language models[20]. The official reported results are given in the first two lines of Table 2 using 493 CD models for the VP and NVP condition, without the second derivative of the cepstral coefficients. Increasing the number of CD models and the number of features, reduced the error rate by about 20% over the system used for the Nov92 evaluation. Results are also given in Table 2 for the Nov92 NVP 64k test data using both open and closed 20k vocabularies. (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[20]). It can be seen that the error rate is doubled when the vocabulary size

WSJ - Conditions	Corr.	Subs.	Del.	Ins.	Err.
493m, 32f, 5k, VP*	93.6	5.5	0.9	1.4	7.8
493m, 32f, 5k, NVP*	91.8	6.9	1.3	1.5	9.7
884m, 48f, 5k, VP	94.5	4.7	0.7	1.1	6.5
884m, 48f, 5k, NVP	94.1	5.2	0.7	1.0	6.9
884m, 48f, 20k, NVP	88.3	10.1	1.5	2.0	13.6
884m, 48f, 20k+, NVP	86.8	11.7	1.5	2.7	15.9

Table 2: Word recognition results on the WSJ0 corpus with a probabilistic grammar (2-grams) estimated on WSJ text data. (5k: 5000 word lexicon, 20k: 20,000 word lexicon, 20k+: 20,000 word lexicon with open test, VP: verbalized punctuation, NVP: non verbalized punctuation, *official DARPA NOV92 evaluation results).

goes from 5k to 20k, whereas the test perplexity goes from 111 to 244 (NVP tests). The higher error rate with the 20k+ open lexicon can be contributed to the out-of-vocabulary words, which account for almost 2% of the words in the test sentences.

One problem using the bigram-backoff LM is that the number of connections is very large. We investigated the effects of reducing the size of the bigram model by relying more on the backoff. Using a count threshold of 4 occurrences, reduces the bigram size by 53% and gives a word error of 7.2% on the 5k-nvp test. This is only a slight increase in the error compared to the 6.9% obtained with a threshold of 1 (baseline bigram[20]).

EXPERIMENTS WITH BREF

BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[17]. 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 text material was read without verbalized punctuation. All the data used for the experiements reported in this paper comes from the BREF80 sub-corpus (2 CDs). 2770 sentences from 57 speakers were used for training. Phonetic transcriptions of the training data were automatically derived and manually verified[4]. A bigram-backoff language model was estimated on about 4 million words of normalized text material from *Le Monde*. The base lexicon, represented with 35 phones, was obtained using text-to-phoneme rules[24], and was extended to annotate potential liaisons and pronunciation variants.

Normalization of the text material entailed a processing rather different from the pre-treatment of the WSJ texts[20]. The main two differences are in the treatment of upper and lower case, compound words and abbreviations. In BREF the distinction between the cases is kept when the upper case 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 determined if a transformation to lower case was needed. The symbols for hyphen, quote, and period can lead to ambiguous separations. For example the hyphen in compound words like "beaux-arts" and "au-dessus" is treated as word-internal. In other cases it may be associated with the first word as in "ex-", or "anti-", or with the second word as in "-là" or "-né". Finally, the hyphen may appear in the text but is not part of 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"). The period may be part of a word, for instance, "L.A.", "sec." (secondes), "p." (page), or not part of any word.

As for the WSJ task, two vocabularies have been used for the recognition experiments, corresponding to the 5k and 20k most common words in the *Le Monde* texts. It should be noted that the word coverage for French: 5k (86%), 20k (95%) is significantly smaller than for the same size lexicons for WSJ: 5k (92%), 20k (98%). For French, the lexicon size must be doubled to obtain the same coverage as in English. The test data consist of 100 sentences for each vocabulary size, with perplexities of 122 for the 5k sentences and 205 for the 20k sentences.

BREF - Conditions	Corr.	Subs.	Del.	Ins.	Err.
500m, 48p, 5k, NVP	87.1	10.3	2.6	1.7	14.5
500m, 48p, 20k, NVP	84.6	12.8	2.6	2.9	18.3
500m, 48p, 5k-H, NVP	90.7	7.2	2.6	1.7	11.5
500m, 48p, 20k-H, NVP	89.4	8.0	2.6	3.0	13.5

Table 3: Word recognition results on the BREF80 corpus with a probabilistic grammar (2-grams) estimated on *Le Monde* text data. (5k: 5000 word lexicon, 20k: 20,000 word lexicon, 5k-H: 5k word lexicon with homophone errors not counted, NVP: non verbalized punctuation).

Word recognition results using 500 CD models and the bigram-backoff language model estimated on the normalized text material from *Le Monde* are shown in Table 3. The word error is 14.5% for the 5k lexicon and 18.3% for the 20k lexicon. The last two entries give the results when the recognizer output is scored without counting homophone errors. These results are provided because French has more homophones than English arising primarily from verb conjugation, the mark of plurals (an -s at the end of the word), as well as the mark of the feminine form (-e at the end of the word) that are often not pronounced. Some complex homophones examples are *multiple word* homophones such as "leur coût" (sing.) and "leurs coûts" (pl.), or *multiword* homophones such as "a mis" and "amis" or "c'est" and "ces". The difference in the results scored with and without homophones points out the need for better language modeling.

DISCUSSION AND SUMMARY

Even though better phone recognition accuracies are obtained for BREF than for WSJ[15, 16], word recognition in English is better. This may be due in part to the higher lexical ambiguity for French. Table 4 gives homophone rates for BREF and WSJ, counted on the training lexicon and texts, where homophone rate is defined to be the number of words which are homophones (words having the same pronunciation), divided by the total number of words. 35% of the words in the 10,311-word BREF training lexicon are homophones, compared to 6% in 8996-word WSJ training lexicon. In the WSJ training texts, 1 out of 5 words is ambiguous, given a perfect phonemic transcription. For BREF, over half the words in the training text are ambiguous. In the right part of Table 4 is shown the number of homophone classes of size k, where a homophone class is the set of graphemic words with a given phonemic transcription. For the WSJ lexicon, the largest homophone class has 4 entries: *B., Bea, bee, and be.* In the BREF lexicon there are 3 homophone classes each having 7 orthographic words, as in *100, cent, cents, san, sang, sans, sent.*

	Homophone rate		Homophone class size (k)			
Corpus	Lexicon	Text	1	2	3	≥ 4
BREF	35%	57%	6686	1329	215	73
WSJ	6%	18%	8453	237	22	1

Table 4: Left: Single word homophones in BREF and WSJ. Right: Table entries correspond to the number of homophone classes with k graphemic forms in the class.

In French, there can also be a relatively large number of pronunciations for a given word. For the word "autres" the following transcriptions are allowed: /ot/, /otrx/, /otr/, /otrxz/, each of which is 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.

A large number of errors for French and English involve short words of one or two phonemes. While there are relatively few of these words, they are very frequent, accounting for about 50% and 30% of all word

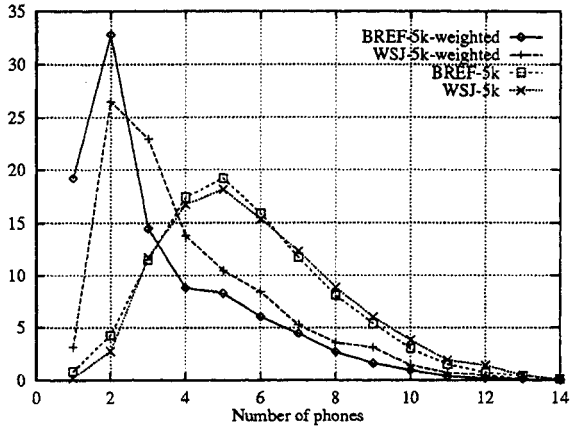

Figure 1: Word distribution in WSJ and *Le Monde* texts for the 5K lexicons as a function of the word length in phones.

occurrences in French and English respectively. Figure 1 shows the distribution of words in the 5k lexicons for BREF and WSJ, as a function of the word length in phones. The curves labeled "weighted" reflect the word occurrences in the training text materials. While the distributions in the lexicons are seen to be quite similar, there is a large disparity in the number of monophone words for the running text (almost 20% for *Le Monde* compared to 3% for WSJ).

In English these short function word errors account for about 70% of the deletions and 40% of the insertions, and 15% of the substitutions. In French, most of the insertions are of mute-e, or of a monophone word that is the same as one of the surrounding phonemes. The latter kind of error is difficult to handle as on the acoustic level it requires refined duration models, and on the LM level a longer span model than a bigram is needed. Deletion problems also involve mostly monophone words, where the reasons for deletion are similar than those for insertion.

For English, we have also observed errors involving inflected forms of verbs such as "finishing→finish in" or "expect it→expected". These are almost multiword homophones. In the first case the error seems to arrive from acoustic causes, where the "ng" is misrecognized as "n", and in the second case the cause is the language model in that "expect it" has a higher probability than "expected".

A problem that was alluded to earlier is that for French, the bigram is less effective than for English, as the amount of training text material is about 1/8 of that for WSJ. This implies that the most frequent words (in particular the monophone words, as shown in Figure 1) have better backoff LM scores, and thus appear easily in place of acoustically similar words which had fewer observations in the text. This problem is also observed for long words with low counts in the training corpus: they are often recognized as a sequence of small words with identical phonemic transcriptions (multiword homophones).

The use of N-class language models (as opposed to N-grams) can be helpful for French, where the number of different graphemic forms for a given root form is much higher than for English. The use of N larger than 2 is also very important, in order to better account for gender and number homophones, and to deal with compound words. In French many compound words are formed by sequences of three words, the frequencies of which can not be estimated by bigrams.

In summary, we have presented our ongoing efforts in developing a task-, speaker- and vocabulary-independent recognizer that can be easily adapted to various applications. The recognizer uses a time-synchronous graph-search strategy which includes intra- and inter-word context-dependent phone models, intra- and inter-word phonological rules, phone duration models, gender-dependent models, and a bigram language model. The recognizer has been evaluated on the DARPA RM and WSJ corpora in English, and the BREF corpus in French, with vocabularies containing up to 20K words.

ACKNOWLEDGEMENT

The authors express their thanks to Murray Spiegel (Bellcore) for providing ORATOR phonetisizations for a subset of the WSJ lexicon.

REFERENCES

[1] F. Alleva, X. Huang, M.-Y. Hwang, "An Improved Search Algorithm Using Incremental Knowledge for Continuous Speech Recognition," *ICASSP-93*.

[2] M. Cohen, *Phonological Structures for Speech Recognition*, PhD Thesis, U. Ca. Berkeley, 1989.

[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.F. Lamel, "Speaker-Independent Phone Recognition Using BREF," *DARPA Sp. & Nat. Lang. Workshop*, Feb. 1992.

[5] J.L. Gauvain, L.F. Lamel, G. Adda, "LIMSI Nov92 WSJ Evaluation," presented at the *DARPA Spoken Language Systems Technology Workshop*, Cambridge, MA, Jan. 1993.

[6] J.L. Gauvain et al, "Speech-to-Text Conversion in French," to appear in *Int. J. Pat. Rec. & A.I.*, 1993.

[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 of Gaussian Mixture Densities for Hidden Markov Models," *DARPA Sp. & Nat. Lang. Workshop*, Feb. 1991.

[9] J.L. Gauvain, C.H. Lee, "MAP Estimation of Continuous Density HMM: Theory and Applications," *DARPA Sp. & Nat. Lang. Workshop*, Feb. 1992.

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

[11] E. Giachin, A.E. Rosenberg, C.H. Lee, "Word Juncture Modeling using Phonological Rules for HMM-based Continuous Speech Recognition," *Computer Speech & Language*, **5**, 1991.

[12] B.H. Juang, "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains", *AT&T Tech. J.*, **64**(6), 1985.

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

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

[15] L.F. Lamel, J.L. Gauvain, "Cross-lingual Experiments with Phone Recognition," *ICASSP-93*.

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

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

[18] C.H. Lee et al, "Acoustic modeling for large vocabulary speech recognition," *Comp. Sp. & Lang.*, **4**, 1990.

[19] H. Murveit et al, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *ICASSP-93*.

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

[21] D.S. Pallett, J.G. Fiscus, "Resource Management Corpus - Continuous Speech Recognition - September 1992 Test Set Benchmark Test Results," Final review of the *DARPA ANNT Speech Program*, Sep. 1992.

[22] D.S. Pallett et al, "Benchmark Tests for the DARPA Spoken Language Program," *ARPA Workshop on Human Language Technology*, Mar. 1993.

[23] P. Price et al, "The DARPA 1000-word Resource Management Database for Continuous Speech Recognition," *ICASSP-88*.

[24] 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.

[25] R. Schwartz et al, "New uses for N-Best Sentence Hypothesis Within the BYBLOS Speech Recognition System," *ICASSP-92*.