Speaker-Independent Phone Recognition Using BREF

Jean-Luc Gauvain and Lori F. Lamel

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE

ABSTRACT

A series of experiments on speaker-independent phone recognition of continuous speech have been carried out using the recently recorded BREF corpus. These experiments are the first to use this large corpus, and are meant to provide a baseline performance evaluation for vocabulary-independent phone recognition of French. The HMM-based recognizer was trained with hand-verified data from 43 speakers. Using 35 context-independent phone models, a baseline phone accuracy of 60% (no phone grammar) was obtained on an independent test set of 7635 phone segments from 19 new speakers. Including phone bigram probabilities as phonotactic constraints resulted in a performance of 63.5%. A phone accuracy of 68.6% was obtained with 428 context dependent models and the bigram phone language model. Vocabulary-independent word recognition results with no grammar are also reported for the same test data.

INTRODUCTION

This paper reports on a series of experiments for speaker-independent, continuous speech phone recognition of French, using the recently recorded BREF corpus[4, 6]. BREF was designed to provide speech data for the development of dictation machines, the evaluation of continuous speech recognition systems (both speaker-dependent and speaker-independent), and to provide a large corpus of continuous speech to study phonological variations. These experiments are the first to use this corpus, and are meant to provide a baseline performance evaluation for vocabulary-independent (VI) phone recognition, as well as the development of a procedure for automatic segmentation and labeling of the corpus.

First a brief description of BREF is given, along with the procedure for semi-automatic (verified) labeling and automatic segmentation of the speech data. The ability to accurately predict the phone labels from the text is assessed, as is the accuracy of the automatic segmentation. Next the phone recognition experiments performed using speech data from 62 speakers (43 for training, 19 for test) are described. A hidden Markov model (HMM) based recognizer has been evaluated with context-independent (CI) and context-dependent (CD) model sets, both with and without a duration model. Results are also given with and without the use of 1-gram and 2-gram statistics to provide phonotactic constraints. Preliminary VI word recognition results are presented with no grammar. The final section provides a discussion and summary, and a comparison of these results to the performance of other phone recognizers.

THE BREF CORPUS

BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers. 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[4]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train VI phonetic models. Hon and Lee[5] concluded that for VI recognition, the coverage of triphones is crucial. Separate text materials, with similar distributional properties were selected for training, development test, and evaluation purposes. The selected texts consist of 18 "all phoneme" sentences, and approximately 840 paragraphs, 3300 short sentences (12.4 words/sentence), and 3800 longer sentences (21 words/sentence). The distributional properties for the 3 sets of texts, and the combined total, are shown in Table 1. The sets are distributionally comparable in terms of their coverage of word and subword units and quite similar in their phone and diphone distributions. For comparison, the last column of the table gives the distributional properties for the original text of *Le Monde*.

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. The speakers, chosen from a subject pool of over 250 persons in the Paris area, were paid for their participation. Potential subjects were given a short reading test, containing selected sentences from *Le Monde* representative of the type of material to be recorded[6] and subjects judged to be incapable of the task were not recorded. The recordings were made in stereo in a sound-isolated room, and were monitored to assure the contents. Thus far, 80 training, 20 test, and 20 evaluation speakers have been recorded. The number of male and female speakers for each subcorpus is given in Table 2. The ages of the speakers range from 18 to 73 years, with 75% between the ages of 20 and 40 years. In these experiments only a subset of the training and development test data was used, reserving the evaluation data for future use.

Unit	*Train*	*Development*	*Evaluation*	*Total*	*Le Monde*
#sentences	3,877	3,624	3,501	11,002	167,359
#words (total)	55,760	50,946	49,040	115,746	4,244,810
#distinct words	14,089	12,803	12,280	20,055	92,185
#phonemic words	11,215	10,177	9,757	15,460	63,981
#syllables	3,339	3,040	2,976	3,977	9,571
#dissyllables	11,297	10,472	10,072	14,066	37,636
#phones (total)	252,369	230,102	222,250	726,988	16,416,738
#distinct phones	35	35	35	35	35
#diphones	1,107	1,092	1,082	1,115	1,160
#triphones	15,704	14,769	14,399	17,552	25,999

Table 1: Distributional properties of selected text subsets: training, development test, and evaluation, and of the original text.

	Number of Speakers		
Corpus	*Male*	*Female*	*Total*
training	37	43	80
development	9	11	20
evaluation	9	11	20
total	55	65	120

Table 2: Speakers in each corpus set

Phone	*Example*	*Phone*	*Example*
Vowels		Consonants	
i	lit	s	sot
e	blé	z	zèbre
E	sel	S	chat
y	suc	Z	jour
X	leur	f	fou
x	petit	v	vin
@	feu	m	mote
a	patte, pâte	n	note
c	sol	N	digne
o	seule	l	la
u	feu	r	rond
Nasal Vowels		p	pont
I	brin, brun	b	bon
A	chant	t	ton
O	bon	d	don
Semivowels		k	cou
h	lui	g	gond
w	oui	English phones	
j	yole	G	thing
·	silence	D	the
		T	Smith
		H	hot

Table 3: The 35 phone symbol set.

Labeling of BREF

In order to be used effectively for phonetic recognition, time-aligned phonetic transcriptions of the utterances in BREF are needed. Since hand-transcription of such a large amount of data is a formidable task, and inherently subjective, an automated procedure for labeling and segmentation is being investigated.

The procedure for providing a time-aligned broad phonetic transcription for an utterance has two steps. First, a text-to-phoneme module[10] generates the phone sequence from the text prompt. The 35 phones (including silence) used by the text-to-phoneme system are given in Table 3. Since the automatic phone sequence generation can not always accurately predict what the speaker said, the transcriptions must be verified. The most common errors in translation occur with foreign words and names, and acronyms. Other mispredictions arise in the reading of dates: for example the year "1972" may be spoken as "mille neuf cent soixante douze" or as "dix neuf cent soixante douze." In the second step, the phone sequence is aligned with the speech signal using Viterbi segmentation.

The training and test sentences used in these experiments have been processed automatically and manually verified prior to segmentation. The manual verification only corrected "blatant errors" and did not attempt to make fine-phonetic distinctions. Comparing the predicted and verified phone strings, 97.5% of the 38,397 phone labels[1] were assessed to be correct, with an accuracy of 96.6%. However, during verification about 67% of the automatically generated phone strings were modified. This indicates that verification is a necessary step for accurate labeling. The exception dictionary used by the text-to-phoneme system has been updated accordingly to correct some of the prediction errors, thereby reducing the work entailed in verification.

Table 4 summarizes the phone prediction accuracy of the text-to-phone translation. 86% of the errors are due to insertions and deletions by the text-to-phone system. Liaison and the pronunciation of mute-e account for about 70% of these. Liaison is almost always optional and thus hard to accurately predict. While most speakers are likely to pronounce mute-e before a pause, it is not always spoken. Whether or not mute-e is pronounced depends on the context in which it occurs and upon the dialect of the speaker. Substitutions account for

[1] Silence segments were disregarded.

Prediction	*Percent*
Correct	97.5
Substitutions	0.5
Deletions	0.9
Insertions	2.0
Accuracy	95.5

Table 4: Phone prediction errors.

only 14% of the errors, with the most common substitutions between /z/ and /s/, and between /e/ and /E/.

A problem that was unanticipated was that some of the French speakers actually pronounced the English words present in the text prompt using the correct English phonemes, phonemes that do not exist in French. These segments were transcribed using the "English phones" listed in Table 3, which were added to the 35 phone set. However, so few occurrences of these phones were observed that for training they were mapped to the "closest" French phone.

In addition, a few cases were found where what the speaker said did not agree with the prompt text, and the orthographic text needed to be modified. These variations were typically the insertion or deletion of a single word, and usually occurred when the text was almost, but not quite, a very common expression.

Validation of automatic segmentation

A subset of the training data (roughly 12 minutes of speech, from 20 of the training speakers) was manually segmented to bootstrap the training and segmentation procedures. In order to evaluate the Viterbi segmentation, the phone recognition accuracy using the manual segmentation for training was compared to the recognition accuracy obtained using Viterbi resegmentation (3 iterations) on the same subset of training data. For this comparison 35 context-independent phone models with 8 mixture components and no duration model, were used. The recognizer was tested on data from 11 speakers in the development test speaker set, and the averaged results are given in Table 5. The performance is estimated by the phone accuracy given by: 1 - (*subs* + *del* + *ins*) / *correct number of phones*. The recognition accuracies are seen to be comparable, indicating that, at least for the purposes of speech recognition, the Viterbi algorithm can be used to segment the BREF corpus once the segment labels have been verified. Including a duration model increases the phone accuracy to 58.0% with the Viterbi segmentation.

Condition	*Correct*	*Subs.*	*Del.*	*Ins.*	*Accuracy*
manual	60.4	27.3	12.3	3.8	56.7
Viterbi	61.8	27.7	10.5	5.0	56.8

Table 5: Training based on manual vs. Viterbi resegmentation

The segmentations determined by the Viterbi algorithm have been compared to the manual segmentations on a new independent set of test data. To do so the offset in number of frames was counted, using the manual segmentation as the reference. Silence segments were ignored. The test data consisted of 115 sentences from 10 speakers (4m/6f) and contained 6517 segments. 71% of the segment boundaries were found to be identical. 91% of the automatically found boundary locations were within 1 frame (96% within 2 frames) of the hand boundary location. The automatic boundaries were located later than the hand location for 23% of the segments, and earlier for 5% of the segments. This assymmetry may be due to the minimum duration imposed by the phone models.

PHONE RECOGNITION EXPERIMENTS

Phone Recognizer

The baseline phone recognizer uses a set of 35 CI phone models. Each model is a 3-state left-to-right HMM with Gaussian mixture observation densities. The 16 kHz speech was downsampled by 2 and a 26-dimensional feature vector was computed every 10 ms. The feature vector is composed of 13 cepstrum coefficients and 13 differential cepstrum coefficients. Duration is modeled with a gamma distribution per phone model. As proposed by Rabiner et al.[11], the HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. Maximum likelihood estimators were used for the HMM parameters and moment estimators for the gamma distributions.

Data

The training data consists of approximately 50 minutes of speech from 43 training speakers (21m/22f). There are 33,289 phone segments containing 5961 different triphones. Thirty-seven of the sentences are *"all-phone"* sentences in which the text was selected so as to contain all 35 phones[4]. These sentences are quite long, having on the order of 190 phones/sentence. The remaining sentences are taken from paragraph texts and have about 65 phones/sentence. The test data is comprised of 109 sentences spoken by 21 new speakers (10m/11f). There are a total of 7635 phone segments (70 phones/sentence) and 3270 distinct triphones in the test set.

Phonotactic constraints

Phone, diphone and triphone statistics, computed on the 5 million word original text, are used to provide phonotactic constraints. Table 6 gives the information stored in the Markov sources (1-gram to 3-gram) estimated from the occurrence frequencies on the original text in bits/phone[4]. For now only the 1-gram and 2-gram constraints have been incorporated in the model.

	#distinct	*entropy*	*model*
Unit/model	*units*	*(b/ph)*	*I(b/ph)*
phones/1-gram	35	4.72	0.40
diphones/2-gram	1,160	3.92	1.21
triphones/3-gram	25,999	3.40	1.72

Table 6: N-gram statistics computed on the 5 million word text and the information stored in Markov source models.

Condition	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
0-gram	62.4	25.4	12.3	3.2	59.2
0-gram+duration	63.5	25.3	11.3	3.5	60.0
1-gram	64.7	23.7	11.6	3.2	61.5
1-gram+duration	65.3	24.1	10.6	3.5	61.8
2-gram	65.9	22.8	11.3	3.3	62.7
2-gram+duration	67.2	22.6	10.2	3.7	63.5

Table 7: Phone recognition results for 35 CI models.

Results

Table 7 gives recognition results using 35 CI phone models with 16 mixture components. Silence segments were not included in the computation of the phone accuracy. Results are given for different phone language models, both with and without a duration model. The improvement obtained by including the duration model is relatively small, on the order of 0.3% to 0.8 %, probably in part due to the wide variation in phone durations across contexts and speakers. Each additional order in the language model adds about 2% to the phone accuracy. The best phone accuracy is 63.5% with the 2-gram language model and duration.

Condition	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
0-gram	69.5	21.7	8.8	4.3	65.2
0-gram+duration	70.8	21.4	7.8	4.7	66.1
1-gram	70.4	20.7	8.8	4.4	66.0
1-gram+duration	72.0	20.5	7.5	4.7	67.2
2-gram	72.1	20.1	7.8	4.6	67.5
2-gram+duration	73.3	20.0	6.7	4.7	68.6

Table 8: Phone recognition results for 428 CD models.

Table 8 gives recognition results using a set of 428 CD phone models[12] with 16 mixture components. The modeled contexts were automatically selected based on their frequencies in the training data. This model set is essentially composed of right-context phone models, with only one-fourth of the models being triphone models. Less than 2% of the triphones found in the training data can be modeled in full. In choosing to model right contexts over left contexts, a preference is given to modeling anticipatory coarticulation over perservatory coarticulation.

Including the duration models improves performance a little more than was observed for the CI models. The duration models are probably better estimates of the underlying distribution since the data has less variability due to context. The duration models give about a 1% improvement in accuracy when used with a 1-gram or 2-gram language model. The phonotactic constraints, however, have a larger effect with the CI models, presumably because the CD models already incorporate some to the phonotactic information.

The use of CD models reduces the errors by 14% (comparing the best CI and CD models), which is less than the 27% error reduction reported by Lee and Hon[7]. There are several factors that may account for this difference. Most importantly, Lee and Hon[7] compare 1450 right-CD models to 39 CI models, whereas in this study only 428 contexts were modeled. In addition, the baseline recognition accuracy reported by Lee and Hon is 53.3% with a bigram language model, compared to our baseline phone accuracy of 63.5%.

Confusion pair	*# Subs.*	*% Subs.*
e → E	64	4.2
E → e	58	3.8
a → E	31	4.2
E → a	27	1.8
n → m	27	1.8
y → i	27	1.8

Table 9: The most common substitutions with 428 models.

The most recognition errors occurred for the phones: /E/ 8.1%, /a/ 7.6%, /e/ 7.2%, /c/ 4.9%, /t/ 4.3%,and /x/ 4.2%, accounting for almost 40% of the substitution errors. Of these phones only /c/ and /E/ have high phone error rates of about 40%. Table 9 shows the most frequent substitutions made by the recognizer. The two most common confusions are reciprocal confusions between /e/ and /E/ and between /E/ and /a/. Together these account for 13% of the confusions. Many speakers do not make a clear distinction between the phones /E/ and /e/ when they occur word-internally, which may account for their high confusability. The high number of errors for /a/ are probably due to the large amount of variability of /a/ observed in different contexts.

14% of the insertions are /r/, followed by 11% for /l/. These two phones also are deleted the most: 13% of the deletions are /l/ and 11% /r/. Although /l/ and /r/ account for many of the insertion and deletion errors, the overall error rate for these phones are relatively low, 11% and 7%, respectively. Improved performance on these phones may be achieved by modeling more contexts and by improving their duration models.

Condition	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
CD 132	69.1	22.0	8.9	3.9	65.2

Table 10: Recognition results for phone class based CD models.

In Table 10 results are given for a set of 132 CD models. The models were selected so as to group phonetically similar contexts based on manner of articulation classes. This is similar to the approach taken by Deng et al.[2]. Taking into consideration that French is a syllable-based language, left-context models were defined for vowels and right-context models for consonants. The phone accuracy of 65.2% lies in between the recognition accuracies of the CI and CD models.

WORD RECOGNITION EXPERIMENTS

Two types of implementation are usually considered to recognize words based on phone models. In the first solution, which can be called *integretated approach*, an HMM is generated for each word by concatenating the phone models according to the phone transcriptions. The word models are put together to represent the entire lexicon with one large HMM. The recognition process is then performed for example by using the Viterbi decoding algorithm. The second solution uses the output of the phone recognizer as an intermediary level of coding such that the lexical decoding is derived only from this ouput. Phonological rules may be included in the lexical decoding, or alternatively may be represented directly in the lexical entries. The phone recognizer output is usually a phone treillis including phone hypotheses for each of the associated speech segments and their corresponding likelihoods. If the first approach appears to offer a more optimal solution to the decoding problem by avoiding an intermediary coding, the second approach greatly reduces the computional requirements of the acoustic level which is independent of the lexicon size and offers a solution to handle out of lexicon words.

Since our goal is to build a system capable of recognizing at least 20,000 words, the second solution is attractive since it allows us to develop and evaluate lexical and language models without interaction with the acoustic level. In particular, this approach is of interest as it permits us to more easily study problems like liaison which are specific to the French language. However, in order to obtain preliminary results on word recognition using BREF, we have chosen to use the integrated approach, primarily because the phone recognizer does not at this time provide a phone trellis. In doing so we have represented liaison in the lexicon by providing alternate pronunciations.

Lexicon	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
1K	73.4	20.9	5.8	4.2	69.2
3K	66.5	27.5	6.0	5.3	61.2
5K	61.4	32.0	6.6	5.9	55.6
10K	55.4	36.9	7.7	6.4	49.0

Table 11: VI word recognition results (no grammar).

Vocabulary-independent word recognition experiments were run using four different lexicons. The smaller lexicon (1K lexicon) contains 1139 orthographic words, only those words found in the test sentences. The 3K lexicon contains all the words found in the training and test sentences, a total of 2716 words. The 5K and 10K lexicons include all the words in the test data complemented with the most common words in the original text. These two lexicons contain respectively 4863 and 10511 words. Alternate pronunciations increase the number of phonemic forms in the lexicon by about 10%. The word recognition results with no grammar are given in Table 11. Since no grammar is used, single word homophone confusions are not counted as errors.

Homophones present a large problem for French. If the homophone errors are included the phone accuracies drop by about 10%. A lexical study with 300,000 words found that there can be over 30 words with the same pronunciation[1]. In the *Le Monde* text corpus of 4.2 million words, there were 92,185 orthographically distinct words, but only 63,981 phonemically distinct words, giving a homophone rate of about 30%. In the 1K and 3K lexicons the homophone rate is lower, on the order of 15%. The "worst-case" homophone in the 3K lexicon is for the phonemic word /sA/, which may correspond to any of the 7 following orthographic words: *100, cent, cents, s'en, sang, sans, sent.* For comparison, there are roughly 3% homophones in RM, less than 2% for TIMIT, and less than 5% in the MIT Pocket lexicon.

While the large number of word homophones in French presents its problems, more complicated homophone problems exist, where sequences of words form homophones. The example in Figure 1 shows some of the homophones for the phonetic sequence /parle/ for the words in the 3K lexicon. These multiple word homophones account for a few percent of the errors in Table 11. In fluent speech, the problems are more complicated as illustrated by Figure 2. While nominally the phonetic transcription of the word "adolescence" is /adclEsAs/, the realized pronunciation is /adxlEsAs/, having the given homophones.

phonetic transcription: p a r l e
word candidates: parler
par les
part les
parle es
parlent es
parle et
parlent et

Figure 1: An example of a multiple word homophone.

phonetic transcription: a d x l E s A s
word candidates: adolescence
a de les sans
a de les sens

Figure 2: An example of a homophone caused by vowel reduction.

The examples given in Figures 1 and 2 do not consider syntactic or semantic constraints. Figure 3 gives an example of the possible analyses of the phrase "un murmure de mécontentement". This example taken from [1] illustrates both the complexity of the problem and the power of the syntactic constraints. Lexical access using a full-form lexicon with over 300,000 entries yields 340 possible word segmentations. This expands to over 2 million possible phrases when all the combinations are considered. Syntactic constraints including form agreement reduce the set to 6 possibilities, all of which are semantically plausible.

text: un murmure de mécontentement
phones: /ImyrmyrdxmekOtAtmA/
lexical access: 340 possible word segmentations
2,419,620 phrases
syntactic analysis: 6 possible phrases
- un murmure de mécontentement
- un murmure de mécontentes ment
- un murmure de mes contentements
- un mur mûr de mécontentement
- un mur mûr de mécontentes ment
- un mur mûr de mes contentements

Figure 3: Lexical hypotheses from a phonemic transcription.

DISCUSSION AND SUMMARY

These preliminary experiments have set a baseline performance for phone recognition using BREF. The preliminary results are somewhat comparable to those obtained for English using the TIMIT corpus. Lee and Hon[7] report 53% accuracy (66% correct) for 39 CI models and 66% accuracy (74% correct) using 1450 right-CD models. Digalakis et al.[3] report 64% (70% correct) accuracy using CI models with a 39-phone symbol set. Levinson et al.[8] report 52% phone recognition with 12% insertions, and do not specify the number of deletions. Phone recognition rates reported for French by Merialdo[9] for speaker-dependent (4 speakers) recognition of isolated syllables were 80.6% accuracy (84.4% correct).

We have taken a first step at vocabulary-independent word recognition using 1K to 10K word lexicons with no grammar. The word accuracy falls from 69% to 49% when the lexicon size increases from 1K to 10K. While these experiments are preliminary, they have given us insight into the problems encountered in lexical access, particularly the difficulties found with single-word and multiple-word homophones, and with liaison.

A procedure for automatic segmentation and labeling of the BREF corpus is being developed. The preliminary investigations indicate that the main problems lie in predicting the phone string, and that while the segmentation is not exact, the vast majority of segment boundaries are located within the same frame as a hand-segmentation. However, it is expected that more accurate segmentations will be obtained by using CD models for segmentation. In addition, a smaller frame step will be used to provide a finer segmentation.

Text-to-phone prediction can be improved by including difficult items, such as foreign words and acronyums, in the exception dictionary. This will not, however, eliminate the need for verification, as it will not handle alternate pronunciations. One option is to have the text-to-phoneme system propose alternate pronunciations for dates and acronyms, and to allow liaison and the pronunciation of mute-e to be optional. In addition, providing a means of flagging poor matches would greatly ease the process of verification.

An HMM-based recognizer has been used for a baseline performance evaluation and verification of the data. In the future better acoustic phone models and duration models will be used. The improvement observed using the sets of CD models indicates, at least with these preliminary experiments, that the improvement appears to be related to the number of CD models that can be trained. We expect to obtain improved phone recognition performance by using more of the training data as only a small portion of the BREF corpus has been used.

REFERENCES

[1] G. Adda (1992), personal communication.

[2] L. Deng, V. Gupta, M. Lennig, P. Kenny, P. Mermelstein (1990), "Acoustic Recognition Component of an 86,000-word Speech Recognizer," *Proc. IEEE ICASSP-90*, pp. 741-744.

[3] V. Digalakis, M. Ostendorf, J.R. Rohkicek (1990), "Fast Search Algorithms for Connected Phone Recognition Using the Stochastic Segment Model," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990, pp. 173-178.

[4] J.-L. Gauvain, L.F. Lamel, M. Eskénazi (1990), "Design Considerations and Text Selection for BREF, a large French read-speech corpus," *Proc. ICSLP-90*, pp. 1097-2000.

[5] H.-W. Hon and K.-F. Lee (1990),"On Vocabulary-Independent Speech Modeling," *Proc. ICASSP-90*, pp. 725-728.

[6] L.F. Lamel, J.-L. Gauvain, M. Eskénazi (1991), "BREF, a Large Vocabulary Spoken Corpus for French," *Proc. EUROSPEECH-91*, pp. 505-508.

[7] K.-F. Lee, H.-W. Hon (1989), "Speaker-Independent Phone Recognition Using Hidden Markov Models," *Proc. IEEE Trans. ASSP*, Vol. 37, No. 11, pp. 1641-1989.

[8] S.E. Levinson, M.Y. Liberman,A. Ljolje, L.G. Miller (1989), "Speaker Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition," *Proc. IEEE ICASSP-89*, pp. 441-444.

[9] B. Merialdo (1988), "Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training," *Proc. IEEE ICASSP-88*, pp. 111-114.

[10] B. Prouts (1980),"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.

[11] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M. Sondhi (1985), "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Technical Journal*, 64(6), pp. 1211-1233, July-Aug. 1985.

[12] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, J. Makhoul (1985), "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," *Proc. ICASSP-85*, pp. 1205-1208.