Experiments on Speaker-Independent Phone Recognition Using BREF

Lori F. Lamel and Jean-Luc Gauvain

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE

ABSTRACT

A series of experiments for speaker-independent, continuous speech phone recognition have been carried out using the recently recorded BREF corpus. Our experiments are the first to use this database, and are meant to provide a baseline performance evaluation for vocabulary independent phone recognition. The system was trained using hand-verified data from 43 speakers. Using 35 context-independent phone models, a baseline phone accuracy of 60% (no phone grammar) has been obtained on an independent test set of 7635 phone segments from 19 speakers. Including phone bigram probabilities as phonotactic constraints results in a performance of 63.5%. A phone accuracy of 68.6% (73.3 % correct) was obtained with 428 context dependent models.

INTRODUCTION

We report on a series of experiments for speaker-independent, continuous speech phone recognition of French, using the recently recorded BREF corpus[3, 4]. 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. Our experiments are the first to use this database, and are meant to provide a baseline performance evaluation for vocabulary independent phone recognition, as well as the basis for development of a procedure for automatic segmentation and labeling of the corpus.

In the first section we give a brief description of BREF, 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. The next section describes the phone recognition experiments performed using speech data from 62 speakers (43 for training, 19 for test) that had been manually verified. An HMM based recognizer with context-independent (CI) and context-dependent (CD) model sets were evaluated, 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.

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[3]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train vocabulary-independent phonetic 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. 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. In these experiments we use only a subset of the training and development test data, reserving the evaluation data for future use.

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-transcriptions of such a large amount of data is a formidable task, and inherently subjective, we are investigating an automated procedure for labeling and segmentation. Initially, however, the labeling is manually verified prior to segmentation.

We are investigating a procedure to provide time-aligned broad phonetic transcriptions by generating the phone sequence with a text-to-phoneme system, and aligning these phones with Viterbi segmentation. The 35 phones (including silence) used by the text-to-phoneme system are given in Table 1. Since the automatic phone sequence generation can not always accurately predict what the speaker said, the transcriptions must be verified. The most common mispronunciations occur with foreign words and names, and acronyms. Other mispredictions are 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."

The training and test sentences used in these experiments have been processed automatically and manually verified. 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.

[1] Silence segments were disregarded.

Phone	*Example*	*Phone*	*Example*
Vowels		Consonants	
i	l**i**t	s	**s**ot
e	bl**é**	z	**z**èbre
E	s**e**l	S	**ch**at
y	s**u**c	Z	**j**our
X	l**eu**r	f	**f**ou
x	p**e**tit	v	**v**in
@	f**eu**	m	**m**ote
a	p**a**tte, p**â**te	n	**n**ote
c	s**o**l	N	di**gn**e
o	s**au**le	l	**l**a
u	f**ou**	r	**r**ond
Nasal Vowels		p	**p**ont
I	br**in**, br**un**	b	**b**on
A	ch**an**t	t	**t**on
O	b**on**	d	**d**on
Semivowels		k	**c**ou
h	l**ui**	g	**g**ond
w	**oui**	English phones	
j	**y**ole	G	thi**ng**
·	silence	D	**th**e
		T	Smi**th**
		H	**h**ot

Table 1: The 35 phone symbol set.

This indicates that verification is a necessary step for accurate labeling. The exception dictionary has been updated accordingly to correct some of the prediction errors, thereby reducing the work entailed in verification.

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

Table 2: Phone prediction errors.

Table 2 summarizes the phone predictions. Substitutions account for 14% of the errors, with the most common substitutions between /z/ and /s/, and between /e/ and /E/. 60% of the errors are insertions and 26% are deletions by the text-to-phoneme system. Liaison and the pronunciation of mute-e account for about 70% of the insertions and deletions. 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.

A problem that we did not anticipate was that some of the French speakers actually pronounced the English words using the correct English phonemes, phonemes that do not exist in French. These segments were transcribed using the "English phones" listed in Table 1 which have been added to the 35 phone set. However, so few occurrences of these phones were observed that for training purposes they were mapped to the "closest" French phone.

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 3: Training based on manual vs. Viterbi resegmentation

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

The segmentations determined by the Viterbi algorithm have been compared to the manual segmentations on an 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, 4 male and 6 female, 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 they were located earlier for 5% of the segments. This assymmetry may be due to the minimum duration imposed by the phone models.

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 3. The performance is estimated by the phone accuracy given by: 1 - (*substitutions* + *deletions* + *insertions*) / *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.

RECOGNITION EXPERIMENTS

Phone Recognizer

Our baseline phone recognizer uses a set of 35 phone models. Each model is a 3-state left-to-right hidden Markov model (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.

Unit/model	*#distinct units*	*entropy (b/ph) (b/ph)*	*model 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 4: 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 5: Phone recognition results for 35 CI models.

As proposed by Rabiner et al.[9], the HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. We used maximum likelihood estimators for the HMM parameters and moment estimators for the gamma distributions.

Data

The training data consists of about 50 minutes of speech from 43 training speakers (21 male, 22 female). 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[3]. 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 (10 male, 11 female). There are a total of 7635 phone segments (70 segments per sentence) and 3270 distinct triphones.

Phonotactic constraints

Phone, diphone and triphone statistics, computed on the 5 million word original text, are used to provide phonotactic constraints. Table 4 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[3]. For now only the 1-gram and 2-gram constraints have been incorporated in the model.

Results

Table 5 gives recognition results using 35 CI phone models with 16 mixture components. Silence segments were not included in the computation of the phone accuracy because we did not want to artificially inflate the scores, since silence is frequent and has a high recognition rate. 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 6: Phone recognition results for 428 CD models.

Table 6 gives recognition results using a set of 428 CD phone models[10] 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. We are able to model less than 2% of the triphones found in the training data. In choosing to model right contexts over left contexts, we have selected to model anticipatory coarticulation more than 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[5]. There are several factors that may account for this difference. Most importantly, Lee and Hon[5] compare 1450 right-CD models to 39 CI models, whereas we only model 428 contexts. 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 7: 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. However, of these phones only /c/ and /E/ have high phone error rates of about 40%. /E/ and /e/ are highly confusable as can be seen in Table 7 which shows the most frequent substitutions made by the recognizer. The two most common confusions are reciprocal confusions between /e/ and

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

Table 8: Phone recognition results for phone class based CD models.

/E/ and between /E/ and /a/. Together these account for 13% of the confusions. The high number of errors for /a/ are probably due to the large amount of variability of /a/ observed in different contexts. Our models include only 12 triphone models and 18 right-context models for /a/, with which we are not able to capture enough of the variation. 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.

14% of the insertions are /r/, followed by 11% for /l/. These two phones also are deletion 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. We are looking into ways to improve the performance on these phones by modeling more contexts and by improving the duration model.

In Table 8 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.[1]. Taking into consideration that French is a syllable-based language, we defined left-context models 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. Although our preliminary attempt to expand the 428 CD model set using a measure of phonetic similiarity has not been successful, we intend to investigate this approach further.

DISCUSSION AND SUMMARY

These preliminary experiments have set a baseline performance for phone recognition using BREF. Our preliminary results are somewhat comparable to those obtained for English using the TIMIT corpus. Lee and Hon[5] report 66% accuracy (74% correct) using CD models for and Digalakis et al.[2] report 64% (70% correct) accuracy using CI models and a 39-phone symbol set. Levinson et al.[6] report 52% phone recognition with 12% insertions, and do not specify the number of deletions. Phone recognition rates reported for French by Merialdo[8] for speaker-dependent (4 speakers) recognition of isolated syllables were 80.6% accuracy (84.4% correct). We are encouraged by our results and expect to obtain improved phone recognition performace by using more of the training data.

We are developing a procedure for automatic segmentation and labeling of the BREF corpus. Our 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, we expect that more accurate segmentations will be obtained by using CD models for segmentation. We also plan to use a smaller step for a finer segmentation.

We plan to improve text-to-phone prediction 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 to propose alternate pronunciations for dates and acronyms, and to allow liaison and mute-e to be optional. In addition, providing a means of flagging poor matches would greatly to ease process of verification.

We have used a relatively simple HMM to do our baseline performance evaluation and verification of the data. In the future we plan to use better acoustic phone models with variable numbers of states or to allow skips. 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 can train. We are encouraged by our results and expect to obtain improved phone recognition performance by using more of the training data as we have used only a small portion of the BREF corpus. We also plan to experiment with other CD phones sets based on phonetic similarity.

REFERENCES

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

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

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

[4] L.F. Lamel, J.-L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *Proc. EUROSPEECH-91*, 1991.

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

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

[7] B. Merialdo, A.-M. Derouault, S. Soudoplatoff, "Phoneme Classification using Markov Models," *Proc. IEEE ICASSP-86*, pp. 2759-2762, 1986.

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

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

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