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

High Performance Speaker-Independent Phone Recognition Using CDHMM

L.F. Lamel and J.L. Gauvain

Speech Communication Group
LIMSI - CNRS, B.P. 133
91403 Orsay, France
Email: {lamel,gauvain}@limsi.fr

ABSTRACT

In this paper we report high phone accuracies on three corpora: WSJ0, BREF and TIMIT. The main characteristics of the phone recognizer are: high dimensional feature vector (48), context- and gender-dependent phone models with duration distribution, continuous density HMM with Gaussian mixtures, and n-gram probabilities for the phonotatic constraints. These models are trained on speech data that have either phonetic or orthographic transcriptions using maximum likelihood and maximum a posteriori estimation techniques. On the WSJ0 corpus with a 46 phone set we obtain phone accuracies of 72.4% and 74.4% using 500 and 1600 CD phone units, respectively. Accuracy on BREF with 35 phones is as high as 78.7% with only 428 CD phone units. On TIMIT using the 61 phone symbols and only 500 CD phone units, we obtain a phone accuracy of 67.2% which correspond to 73.4% when the recognizer output is mapped to the commonly used 39 phone set. Making reference to our work on large vocabulary CSR, we show that it is worthwhile to perform phone recognition experiments as opposed to only focusing attention on word recognition results.

INTRODUCTION

This paper presents some of our recent research on speaker-independent continuous phone recognition using continuous density HMM (CDHMM) context-dependent phone models trained with maximum likelihood (MLE) and maximum a posteriori (MAP) estimation techniques. The phone accuracy is assessed on the Wall Street Journal (WSJ)[17] and TIMIT[2] corpora for English, and on the BREF[4, 13] corpus for French. WSJ and BREF are similar style corpora in that both contain spoken material from read newspaper text recorded under similar conditions (8kHz bandwidth, close-talking microphone) . Results are given for TIMIT in order to allow comparison with other researchers' work, as TIMIT has been widely used to evaluate phone recognition[14, 11, 19, 20, 15].

While in recent years speech recognizer evaluation has focused on word error rates, we believe that evaluating recognition at the phone level is important for several reasons. With increased interest in portable speech recognition components, there is a demand for vocabulary-independent (VI), speaker-independent (SI), continuous speech recognition which typically implies an approach based on phone-like units. The better these phone models (or acoustic models) are, the better the performance of the entire system will be. Only considering word or sentence recognition performance, particularly when word-based n-gram constraints are used, can mask problems that arise from the acoustic level.

In performing detailed error analysis of word recognition errors the phonetic recognition results are often used to understand the source of the errors. In fact, our recent experiments on the WSJ task show that improvements in phone accuracy directly led to improvements in word accuracy when the same phone models were used for recognition[10]. Phone recognition is also useful in determining pronunciation errors in the lexicon and identifying alternate pronunciations that need to be included. A related research area has been the identification of non-linguistic speech features[12], which uses phone-based acoustic likelihoods. This approach has been shown to be effective for French/English language identification, and speaker and sex identification in both languages.

In this paper, we show that high phone recognition accuracies can be obtained using relatively small sets of continuous density context-dependent (CD) phone models. The paper is organized as follows. First the recognizer and the training procedure are described. Then experimental results are reported for the three corpora. Finally, by illustrating the link in performance between phone recognition and word recognition using the same phone model sets, the importance of assessing the speech recognizer at the phone level is demonstrated.

RECOGNIZER DESCRIPTION

The speech signal is converted in a sequence of feature vectors with a fixed 30 ms frame and a frame rate of 10 ms. The feature set includes cepstral coefficents derived from LPC or DFT based cepstra with their first and second order derivatives (Δ and $\Delta\Delta$ cepstrum). The log-energy and its first and second derivatives are also included in the same high dimensional feature vector. LPCC analysis is used for the 4kHz bandwidth and MFCC analysis is used for the 8kHz bandwidth. For the MFCC analysis, the channel Bark power spectrum is obtained by applying triangular windows to the DFT and the cepstrum coefficients are then computed using a cosinus transform[1]. For a 4kHz bandwidth no significant differences were observed using LPC or DFT based cepstra[9]. For an 8kHz bandwidth, this analysis was found to outperform an LPC-based analysis even with frequency warping.

To model the sequence of feature vectors, the recognizer uses a set of CD phone models, where each model is a three-state first-order left-to-right CDHMM with Gaussian mixture observation densities. The phone contexts to be modeled are automatically selected based on their frequencies in the training data. The models may be triphone models, right-context phone models, left-context phone models, or context-independent phone models. The covariance matrices of all the Gaussian components are diagonal. Since phone duration is not adequately modeled with a three state Markov chain, a separate duration density is associated with each phone model. Duration is thus modeled with a gamma distribution per state. As proposed by Rabiner et al.[18], the HMM and duration parameters are estimated separately and combined in the recognition process during the Viterbi search.

The main advantage continuous density modeling offers over discrete or semi-continuous (or tied-mixture) observation density is that the number of parameters used to modelize an HMM observation distribution can easily be adapted to the amount of available training data associated to this state (This can be estimated by use of the Viterbi algorithm, for example). So as a consequence, high precision modeling can be achieved for highly frequented states without the explicit need of smoothing techniques for the densities of less frequented states. Discrete and semi-continuous modeling use a fixed number of parameters to represent a given observation density and therefore cannot achieve high precision without the use of smoothing techniques. This problem can be alleviated by tying some states of the Markov models in order to have more training data to estimate each state distribution[20, 6]. This kind of tying requires careful design, some a priori assumptions, and results in a more complex training procedure. However, these techniques are of interest, particularly in situations where the training data is limited and cannot easily be increased.

When a bigram phone model is used, the overall Markov chain is obtained by connecting the phone HMMs through null states representing all the possible diphones. These null states, which do not emit any observation, are used to merge all the transitions corresponding to the same diphone, thus reducing the number of connections to a more manageable value (i.e., the fourth order becomes a cubic form). With 500 CD models for the WSJ corpus, the resulting HMM includes 1656 non-null states and has about 4 million parameters. Gender-dependent phone models are used to more accurately model the speech signal. This is done by doubling the number of CD phone models, one set being trained on female data and the other on male data. When using gender-dependent models, duplicate Markov chains are built (one for each sex) and the initial and final states of the two chains are merged. The duration models are gender independent.

Phone n-gram probabilities (2-grams or 3-grams) computed on the training data are used to provide phonotactic constraints. These phonotactic constraints correspond to the between phone model transition probabilities (in the case of the 2-gram the transition probabilities are tied to be associated to the n^2 2-gram values, where n is the number of phones).

Maximum likelihood estimators are used for the speaker-independent HMM parameters[7] and moment estimators for the gamma distributions. Training for TIMIT and BREF makes use of the associated phone labels. For WSJ, training uses the orthographic transcriptions, since corresponding phone transcriptions do not exist. In this case, a Markov chain corresponding to all the possible phone strings for the given sentence is generated based on phone transcriptions in an associated lexicon. The network is then modified to incorporate alternate pronunciations in the lexicon, and phonological rules are applied to hypothesize possible realizations, in an attempt to account for some of the phonological variations observed in fluent speech. Using these optional phonological rules during training results in better acoustic models, as they are less "polluted" by "wrong" transcriptions. To build gender-dependent models, MAP estimation is used to perform adaptative training of the speaker-independent models with the sex-specific data[5].

The phone decoding is carried out by determining the most likely Markov state sequence using a one pass Viterbi beam search. When using gender-dependent models, the two Markov chains are processed in parallel. In practice we have found that the Markov chain corresponding to the wrong sex is rapidly removed from the search space.

EXPERIMENTS WITH WSJ0

The DARPA Wall Street Journal Corpus[17] was designed to provide general-purpose speech data with large vocabularies. This corpus serves as the focus for the DARPA continuous speech recognition evaluations[16]. In these experiments the standard SI-84 training material, containing 7240 sentences from 84 speakers (42m/42f) is used to build the phone models. The non-verbalized (nvp) and verbalized (vp) punctuation Feb92 pilot evaluation material are used for test, containing 200 sentences from 10 speakers (6m/4f) for each condition. Since there are no associated phone transcriptions for this data, a phone transcription was determined by performing segmentation as described above. A set of 46 phones are used for WSJ consisting of 21 vowels, 24 consonants, and silence. Phonotactic constraints are provided by a phone bigram estimated on automatically generated phone labels of the training data. The phone perplexity of the nvp test data is 17.5.

WSJ0	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
nvp, 4k, Δ, 500m	71.3	21.4	7.3	5.2	**33.9**
nvp, 8k, Δ, 500m	74.8	18.7	6.5	4.9	**30.1**
nvp, 8k, ΔΔ, 500m	77.0	17.1	5.9	4.6	**27.6**
nvp, 8k, ΔΔ, 900m	78.9	16.2	4.9	4.8	**25.9**
nvp, 8k, ΔΔ, 1600m	79.3	16.2	4.5	5.0	**25.7**
vp, 8k, ΔΔ, 1600m	82.3	13.7	3.9	4.3	**22.0**

Table 1: Phone recognition results for WSJ0 using 46 phones and phone bigram.

Experimental results for the Feb92 pilot test data are given in Table 1 where silences have been removed prior to scoring. Two sets of CD models are used, one for each gender. The first two entries compare the phone accuracy for a 4kHz and 8kHz bandwidth, using 500 CD models and the Δ cepstrum. An absolute reduction in the phone error of almost 4% is obtained with the larger bandwidth. Increasing the size of the feature vector to include the ΔΔ cepstrum gave an additional absolute error reduction of 2.5% with the same model set. The next entry shows that when the number of CD models is increased to 900, the absolute error is reduced by 1.7%. Increasing the number of models to 1600 gives only a small reduction of 0.2%. The final entry in the table shows the phone accuracy on the Feb92 vp test data using the same set of 1600 CD models. The improved accuracy can be attributed to the frequent occurence of the phones in the punctuation words (particularly *period*, and *comma*), which are both well-modeled and well-recognized.

These results are all based on a comparison of the recognized phone string with what we have designated the "reference" phone transcription - that determined by the recognizer given the text string and allowing for alternate pronunciations and the application of phonological rules. We may pose the question that this produces optimistic accuracies in that there a bias in favor of the recognizer, which has provided the reference. In order to assess the order of magnitude of such a bias, the phone transcriptions of the 200 nvp sentences were manually verified and corrected, at the phonemic level. On average about 3.5% of the phone labels were modified. The recognizer output was rescored using these corrected transcriptions, and the overall phone error increased by 0.3%. However, in rescoring with the modified transcriptions, we introduce a new bias – this time against the recognizer, since it was trained under different conditions (i.e., uncorrected transcriptions). Given that the difference in phone accuracy relative to the overall error rate is small, we believe that the overall quality of the phone models can be assessed using automatically generated reference labels.

EXPERIMENTS WITH BREF

BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[13]. 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 phone models. The text material was read without verbalized punctuation. The training material used in these experiments consists of 2770 sentences from 57 speakers (28m/29f). This represents approximately 4.3 h of speech data. 93 sentences from 8 speakers (4m/4f) were used for test. The test text material is distinct from the training texts and has a phone perplexity of 16.1. Phone transcriptions of these utterances were automatically generated and manually verified[3] using a set of 35 phones including 14 vowels, 20 consonants, and silence[8]. The phone bigram was trained on automatically generated phoneme transcriptions of the training text material in the *Le Monde* corpus.

BREF	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
4kHz, 16g, Δ	79.4	15.0	5.6	3.2	**23.8**
8kHz, 16g, Δ	79.9	15.2	4.8	3.5	**23.6**
8kHz, 32g, $\Delta\Delta$	81.7	13.7	4.6	3.0	**21.3**

Table 2: Phone recognition results for BREF using 35 phones and phone bigram.

Phone recognition results for BREF are given in Table 2 using gender-specific sets of 428 CD models. Silences were removed prior to scoring. Comparing the first two entries it can be seen that increasing the bandwidth to 8kHz from 4kHz gives only a minimal reduction in the phone error (0.2%) which is certainly not significant. In contrast to the observation for WSJ, increasing the bandwidth for French is not particularly useful. The last table entry increases the maximum number of Gaussians per state to 32 (from 16) and also includes the $\Delta\Delta$ cepstrum in the feature vector. The resulting phone error is 21.3%. This high performance in phone recognition has also been obtained on other test data from the BREF corpus.

Model set	*#features*	*TIMIT (61)*	*TIMIT (39)*
4k LPCC	26	39.3	32.8
8k MFCC	32	37.2	30.9

Table 3: Phone error for TIMIT coretest with 500 CD models and phone bigram.

TIMIT	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
61, bg, 8k, $\Delta\Delta$	71.2	23.2	5.6	4.6	**33.4**
39, bg, 8k, $\Delta\Delta$	77.5	17.1	5.3	4.7	**27.1**
61, tg, 8k, $\Delta\Delta$	72.1	22.8	5.2	4.9	**32.8**
39, tg, 8k, $\Delta\Delta$	78.3	16.7	4.9	4.9	**26.6**

Table 4: Phone recognition results for TIMIT complete test, with 500 phone models, $\Delta\Delta$ cepstrum, and phone bigram (bg) or trigam (tg).

EXPERIMENTS WITH TIMIT

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[2] is a corpus of read speech designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the U.S. For these experiments we have used the training/test subdivision of TIMIT as specified on the TIMIT CDROM[2] which ensures that there is no overlap in the text materials. The subdivision provides 10 sentences from each of 462 speakers for training. The coretest set consists of 192 sentences, 8 from each of 24 speakers (2m/1f from each dialect region) and the complete test test is comprised of 1344 sentences from 168 speakers (112m/56f). All of the utterances in TIMIT have associated time-aligned phonetic transcriptions. The standard set of 61 phone symbols is used for TIMIT[2]. The phone language models were trained on the training texts after removing the 2 SA (speaker calibration sentences) that were spoken by each speaker. The test data has a phone perplexity of 17.7 with this phone set and 14.6 with a commonly used reduced 39 phone set[14].

Table 3 gives results comparing 4k LPCC and 8k MFCC analysis with Δ cepstrum, one set of 500 speaker-independent CD models and a phone bigram for the coretest. Silences are included as transcribed following the commonly adapted scoring practice for TIMIT. Scoring without the sentence initial and final silences increases the phone error by about 1.5%. With both phone sets an error reduction of about 2% is observed using an 8kHz bandwidth instead of 4kHz.

Phone recognition results with a larger feature vector which includes the $\Delta\Delta$ cepstrum are shown in Table 4 for the complete test. The final entries show the recognition results when a phone trigram is used to provide phonotactic constraints in place of the bigram. The improvement is seen to be small, only about 0.5%. This is to be expected as the test set 61-phone perplexity with the trigram is 17.0, which is only slightly lower than with the bigram. This can certainly be attributed to the limited training data available to estimate the trigram.

LINK WITH WORD RECOGNITION

Our recent development work on the WSJ task has pointed out the importance of phone recognition for assessing the word recognizer. Evaluating phone recognition enables us to assess the quality of the phone models without constraints imposed by lexical information. It is also extremely important in that it is much easier and faster to test out ideas using phone recognition, than word recognition. In Table 5 results are given for phone accuracies on development data from WSJ0 (Feb92-5k-si-nvp) and corresponding word accuracies on the Nov92 5k-nvp test data. Improvements in speaker-independent phone accuracy on the development data are seen to yield improvements in word accuracy on independent test data. These results provide direct evidence that it is worthwhile to run phone recognition experiments in order to improve acoustic modeling.

Condition	*Phone Accuracy* *Feb92-5k-si-nvp*	*Word Accuracy* *Nov92-5k-si-nvp*
500 models, Δ	69.9	90.3
500 models, $\Delta\Delta$	72.4	91.7
900 models, $\Delta\Delta$	74.1	93.1

Table 5: Phone accuracy and word accuracy for WSJ .

Our analysis of the word recognition errors for the Resource Management[9] and Wall Street Journal[10] tasks made reference to the results of phone recognition in order to understand the acoustic source of the errors, and to propose potential solutions. This approach allowed us to discover alternate pronunciations that need to be added to the lexicon, as well as errors in the existing pronunciations. The analysis also led to the revision of phonological rules as well as the discovery of new phonological rules to be incorporated in the system.

SUMMARY

Our recent work focuses on developing phone-based recognizers that are task-, speaker- and vocabulary-independent so as to be easily adapted to various applications. In this paper we have described a phone recognizer based on CDHMMs and showed that high performance is obtained with a small number of CD phone models. The recognition experiments have compared phone recognition accuracies for different feature sets and for model sets of different sizes. The experiments indicate that a 4kHz bandwidth is sufficient to recognize French, but that an 8kHz bandwidth improves the phone accuracy for English. It also appears that French is easier to recognize at the phone level (accuracy: BREF 78.7% vs. WSJ 74.3%) for test sets with comparable perplexities. It may be simply that the phonetic structure of French is easier to recognize than that of English. French has fewer consonant clusters than English, and has a more regular consonant-vowel alternation. French vowels are also acoustically relatively stable when compared to American English vowels whose spectral characteristics vary more within the segment. The phone accuracy on TIMIT with a phone trigram is 73.4% when scored remapping the phone labels to 39 symbols. This result is slightly lower then the results reported by Robinson [19] on TIMIT (75.0%) by using a recurrent neural network. There is certainly space for improvements in our TIMIT recognizer, in particular by increasing the number CD phone units associated with tying techniques as used in [20].

We advocate the use of phone accuracy to assess the quality of the acoustic modeling. Used to aid the analysis of word recognition errors, the phone transcription can help locate poor acoustic models, modifications necessary to the lexicon, as well as lead to the discovery of additional phonological rules. We have illustrated that improvements in *phone accuracy* have led to improvements in *word accuracy* when the word recognizer uses the same phone models.

REFERENCES

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

[2] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM" NTIS order number PB91-100354, 1993.

[3] J.L. Gauvain, L.F. Lamel, "Speaker-Independent Phone Recognition Using BREF," *DARPA Speech & Nat. Lang. Workshop*, Arden House, NY, Feb. 1992.

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

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

[6] M.Y. Hwang, X. Huang, "Subphonetic Modeling with Markov States - Senone," *ICASSP-92.*

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

[8] L.F. Lamel, J.L. Gauvain, "Experiments on Speaker-Independent Phone Recognition Using BREF, *ICASSP-92.*

[9] L.F. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," *DARPA Continuous Speech Recognition Workshop*, Stanford, CA, Sep. 1992.

[10] L.F. Lamel, J.L. Gauvain, G. Adda "LIMSI Nov92 WSJ Evaluation," presented at the *DARPA Speech and Natural Language Workshop*, Cambrigde, MA, Jan. 1993.

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

[12] L.F. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *EUROSPEECH-93.*

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

[14] K.F. Lee, H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. ASSP*, 37(11), 1989.

[15] A. Ljolje, "High Accuracy Phone Recognition Using Context Clustering and Quasi-triphonic Models," submitted to *Computer Speech & Language.*

[16] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, "Benchmark Tests for the DARPA Spoken Language Program," *ARPA Workshop on Human Language Technology*, Mar. 1993.

[17] D. Paul, J. Baker, "The Design for the Wall Street Journal-based CSR Corpus" *DARPA Speech & Nat. Lang. Workshop*, Arden House, NY, Feb. 1992.

[18] 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 Tech. J.*, 64(6), 1985.

[19] T. Robinson, "Several improvements to a recurrent error propagation phone recognition system," *Tech. Rep. CUED/TINFENG/TR.82*, 1991.

[20] S.J. Young, P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *EUROSPEECH-93.*