Phonetic Recognition Improvements through Input Feature Set Combination and Acoustic Context Window Widening

Carla Lopes[1,2], Fernando Perdigão[1,3]
[1]Instituto de Telecomunicações, Coimbra, Portugal
[2]Instituto Politécnico de Leiria-ESTG, [3]Universidade de Coimbra-DEEC
e-mail:{calopes, fp}@co.it.pt

***Abstract* — This paper deals with phoneme recognition based on a hybrid Multi-Layer Perceptron (MLP)/ hidden Markov model system. The effects of the combination of multiple feature sets and the use of a new wide acoustic context procedure on the training of a MLP are investigated.**

Experimental results show that the contribution of specific features to phoneme recognition, when used in combination with standard MFCC features was about 1.3% of accuracy improvement. The proposed acoustic context window widening leads to FER relative improvements of 2.8%. Relative improvements of 3.3% and 8.5%, respectively, on accuracy and correctness rates, were obtained if both proposals are included in the training of the phoneme recognition system.

I. INTRODUCTION

Two of the most important aspects in developing a Multi-Layer Perceptron based speech recognition system are to extract the useful information from the speech signal and to find a reliable network architecture. This paper tackles both questions. With regard to the training data, two approaches are suggested and tested. 1) MEL-frequency cepstral coefficients (MFCC), Perceptual Linear Prediction, and variations, are the most used features in current speech recognition systems. These parameterizations have been found to retain the most important acoustic information needed for accurate speech recognition. However, there have also been attempts to use other kind of speech parameterizations in the acoustic front-end, also leading to good recognition results,[1],[2]. In this paper we investigate the contribution of specific features to phoneme recognition accuracy, when used in combination with standard MFCC features. Combining the strengths of MFCC features with a set of features with some meaningful physical interpretation, like voicing, spectral flatness, etc., is a way of explicitly incorporating knowledge of important details of human speech production in the recognition process. 2) On the other hand, better recognition results are achieved when the system incorporates both short-time information and information over longer periods of time, [2],[5],[14]. Nevertheless, the resulting improvement in recognition accuracy is also related to a proportional increase in the number of training parameters. In order to overcome this problem, we propose a widening of the acoustic context at zero additional parameter, by using frames alternately. The number of training parameters is maintained, but a greater temporal context is incorporated.

Carla Lopes would like to thank the Portuguese foundation: Fundação para a Ciência e a Tecnologia for the PhD Grant.

II. FRAME ERROR REDUCTION THROUGH INPUT FEATURE SET COMBINATION

Neural Networks are capable of incorporating all kinds of input features and adjust itself in such a way that the optimal combination of these features is found for classification. Exploiting this potentiality, input features derived from two different parameterization algorithms are combined: standard MFCC and an additional set. Different parameterizations of the speech signal may potentially extract additional information useful to increasing the discrimination between confusable sound classes. As Li pointed out in [7], MFCC based spectral features, work well to classify some attributes, but fail in other cases where temporal features may be more discriminative. This motivated us to explore the contribution that other kinds of temporal and spectral speech features could make to phoneme discrimination. Table 1 shows the set of features used. These 10 features have already proved to be suitable for the identification of broad classes of events, [3].

Table I
Acoustic Feature Set used in combination with standard MFCC Features

Number	Feature description
1	35 ms log-energy
2	Max amplitude
3	Spectral Flatness Measure (SFM)
4	Spectral Centroid
5	Log of energy ratio at high/low frequencies
6	Median of energy in a HF band
7	Log energy for f<500Hz
8	Log energy for 500< f <1500Hz
9	Voice evidence (from a pitch detector)
10	Peakiness

The system used in the experiments consists of an MLP, which trains the original 61 phoneme set of the TIMIT database [6]. Speech is analyzed every 10ms with a 25ms Hamming window. Thirty-nine parameters were used as

standard input features representing 12 MFCC, plus energy, and its first and second derivatives.

A context window of 9 frames was considered for training. The MLP performance is evaluated by means of frame error rate (FER). In FER calculations boundary frames between two adjacent models are not considered. Both systems have similar number of parameters (about 124k). This is a result of an alteration of the number of hidden nodes. The MLP system that uses 39 features have 300 hidden nodes while the MLP of 49 features have 250.

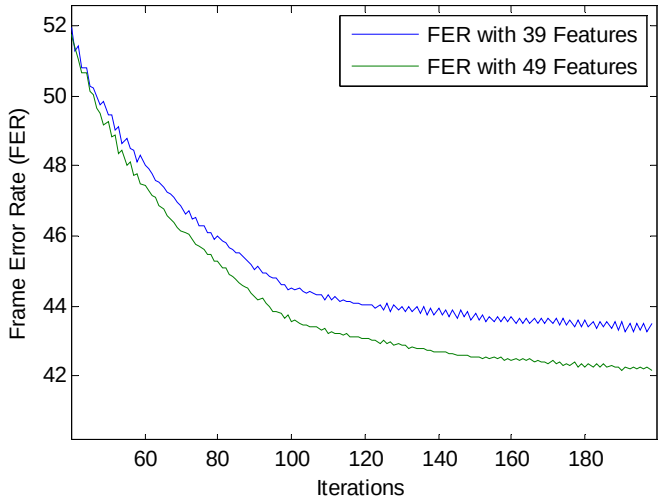

Fig. 1. FER comparison of training results when using 39 or 49 input features.

Results are displayed in Figure 1. In all training iterations, the MLP with 49 features (39 MFCC plus the 10 from Table I got, the best accuracy. Accuracy is about 1.3% higher (3% relative) than if we use only the standard 39 MFCC features, which means that the new set of features actually contributes to the discrimination between classes.

This effect is more evident for vowels, but silences, stops, fricatives and nasal classes also saw improvements. Furthermore, the best improvement was 8.6% for a nasal phoneme (/nx/). Figure 2 shows some examples of the improvements achieved when comparing the MLP with 49 features (dark gray) with the MFCC baseline system.

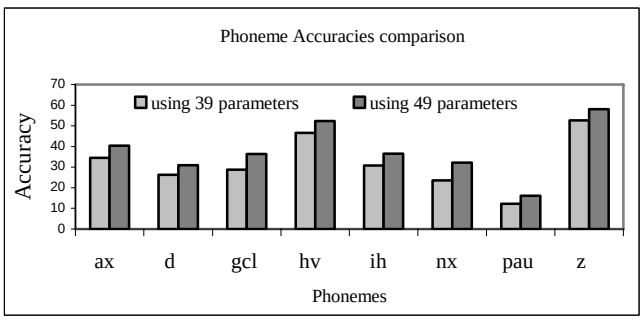

Fig. 2. Frame error rate major improvements due to the use of 49 input features instead of the traditional 39.

III. FRAME ERROR REDUCTION THROUGH ACOUSTIC CONTEXT WINDOW WIDENING

Short-time speech representations are widely used in current speech recognition systems, and have already proved to be suitable to represent the most important acoustic information needed for accurate speech recognition. However, better recognition results are achieved when the system also incorporates information over longer periods of time, [2],[5],[14].

In neural networks a context window spanning several input frames, enables the system to learn, within certain limits, the temporal patterns of speech units. This context window typically stops at about 9 frames. It is usually determined in order to balance the trade-off between the number of parameters and recognition accuracy.

To test the usefulness of incorporating a larger temporal context window, an MLP was trained doubling the context window (but also using 9 frame features) and its performance was compared with two standard 9 frame context window. One is centered in the middle frame and the other looks only to past frames. Figure 3 illustrates the procedure. The context window is 170ms (equivalent to 17 frames) but only 9 frame features were used, one every other. The unused frame features are used in the next window analysis. The current frame is in the center of the context window (temporal information of past and future is included). Regarding Figure 3, the white squares represent the frames discarded and the gray ones are the ones considered. In this way the number of training parameters was maintained with a larger temporal context.

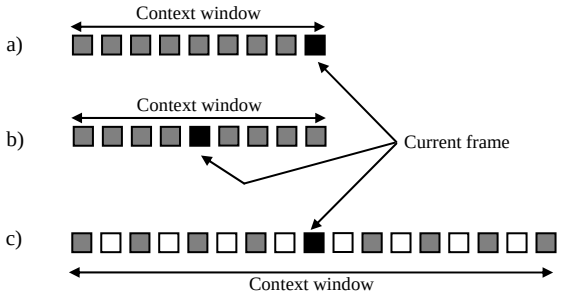

Fig. 3. a) Typical context window extension where the used context looks only to past frames. b) Context window extension where current frame is in the center of the context window. c) Proposed context window extension and position.

The results, in terms of FER are shown in Figure 4. The MLP trained with the proposed structure has a better FER in all iterations. The FER relative improvements are about 2.8%, comparing to the typical context window, which means that it is advantageous to use the proposed widening of the acoustic context, even though the used information is only a rude description.

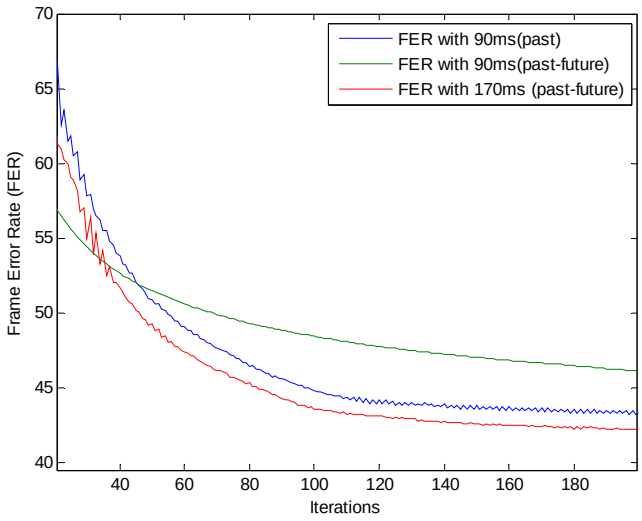

Fig. 4. FER comparison of training results when using 170ms vs. 90ms of context using only past frames and 90ms of context using past and future frames (in all tests the 49 features were used).

IV. PHONEME RECOGNITION SYSTEM

An MLP network consisting of an input layer and an output layer was trained for phoneme frame classification. Both training and testing were carried out with the TIMIT database [6], using the original 61 phoneme set. The training set consisted of all si and sx sentences of the original train-ing set (3698 utterances) and the test set consisted of all si and sx sentences from the complete 168-speaker test set (1344 utterances). The targets derive from the phoneme boundaries provided by the TIMIT database. The 49 parameters described in Section 2 were used as standard input features, and the context window described in Section 3 (170ms using 9 frames) was considered for training. The softmax function was used as the activation function of the output layer, so that the output values are interpreted as a posterior probability of phoneme. All the weights and bias of the network are adjusted using batch training with a resilient back-propagation (RP) algorithm [9] so as to minimize the minimum-cross-entropy error between network output and the target values. The choice of the error function followed Bishop's suggestion [4], which was later clarified by Dunne [11]. It states that the softmax activation function should couple with the cross-entropy penalty function.

Besides the neural network discriminating between the full 61 TIMIT phonemes, these symbols are sometimes considered a too narrow description for practical use, and for evaluation we collapsed the 61 TIMIT labels into the standard 39 phonemes proposed by Lee and Hon [8].

A. Frame Based Detection

For frame error rate results we simply chose the unit with the largest class prediction value from all the classes' output values ("*winner-takes-all*"). In the output layer we managed 25.62% of FER among the 39 classes. The results achieved are depicted in the first line of Table II. It was compared with another work which, in addition to being based on different methods and architectures, also evaluates FER on TIMIT. It was proposed by Scanlon, Ellis and Reilly, [10] and it is based on a modular architecture in which information about broad phonetic groups' membership is 'patched' into a baseline classifier, Their work also presents frame level accuracies, but in terms of four broad phonetic groups (vowels, stops, fricatives, nasals). Using TIMIT data, and the broad phonetic class description they provide in Table 1 of their paper, we also computed FER in terms of 5 broad classes (vowels, stops, fricatives, nasals and silence).

We transformed the 61 outputs of the network into 5 by summing the outputs that belong to the same class. This procedure differs from the one proposed in [10]. In their work, 4 networks work in parallel (frame belongs/does not belong to the class) while we have a single network. Apart from this difference, we obtained very promising results.

The results are depicted in Table II. Except in the stop class our results outperform the others. For example, for vowels we achieved a relative improvement of 90%!

Table II
Frame Error Rate comparison: overall FER evaluation of 39 TIMIT phonemes and four broad phonetic groups.

	Frame Error Rate (%)				
	39 phonemes	vowels	stops	fricatives	nasals
Our Proposal	**25.62**	**4.2**	23.5	**14.3**	**21.1**
Scanlon, Ellis & Reilly	—	40.2	**16.9**	18.6	24.1

B. Segment Based Detection

Besides a good frame error rate performance, another goal of speech recognition systems is segment based detection. For that reason, we propose a hybrid system that combines the time warping abilities of hidden Markov model (HMM) with the discrimination capabilities of Neural Networks.

The proposed system performs utterance segmentation in terms of the 39 TIMIT phonemes.

It uses a Markov process to temporally model the speech signal, but instead of using *a priori* state-dependent observation probabilities defined by Gaussian mixtures, it uses *a posteriori* probabilities estimated by the MLP, keeping the overall HMM topology unchanged.

In the proposed hybrid approach we considered that the output predictions of the MLP correspond to phoneme posterior probabilities for the input features, $P(p_k \mid \mathbf{X})$, with p_k representing the k^{th} phoneme and **X** the feature observation vector. We use them as local probabilities in HMM, avoiding the use of Gaussian mixtures.

The HMM acoustic models were built for all phonemes by using HTK 3.4 [13]. Each phoneme was modeled by a three-state left-to-right HMM and each state shared the same MLP output. We used HTK, for testing with some changes in order to replace the usual Gaussian mixture models by the normalized MLP outputs values.

The performance of the hybrid system was evaluated by means of Correctness (Corr) and Accuracy (Acc) using HTK evaluation tool (`HResults`).

Table III presents the results (see the "Single Layer" row). Also the results of a HMM baseline system (using standard Gaussian mixtures to model state-dependent observation probabilities) are presented for comparison. It was found that the proposed system outperforms the HMM system, using one Gaussian mixture in relation to both correctness and accuracy.

Relative improvements of 3.3% and 8.5%, respectively were achieved. This means that the proposed system has both good frame classification performance and good segmentation performance.

Table III
Phoneme recognition results, in TIMIT 39 phonemes

	Corr	**Acc**	**Details**
HMM	59.84	56.21	1 mixture
Hybrid ANN/HMM	61.79	61.00	Single layer

V. CONCLUSIONS

In this paper, a multi-layer perceptron architecture selection to improve phoneme recognition rate are described and tested. Combining the strengths of MFCC features with a set of features with some meaningful physical interpretation and combining both short and long time information lead to 1.3% of phoneme accuracy improvement. A new strategy related to the enlargement of the temporal information included in training resulted in relative frame error rate reduction of 2.8%. Using both the input feature set combination and the acoustic context window widening in the training of the MLP resulted on an improvement of the performance of the hybrid MLP/HMM phoneme recognition system. Experimental results show relative improvements of 3.3% and 8.5%, on accuracy and correctness rates, respectively.

REFERENCES

[1] A. Ali, J. der Spiegel, P. Mueller, G. Haentjens and J. Berman, "An Acoustic-Phonetic Feature-Based System For Automatic Phoneme Recognition In Continuous Speech", in Proc. ISCAS'99-Vol.3, Florida, June 1999, pp.118-121.

[2] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in Proc. ICSLP 2004, Jeju Island, KR, Oct. 2004.

[3] C. Lopes, F. Perdigão, "Speech Event Detection By Non Negative Matrix Deconvolution", in Proc. EUSIPCO-2007, Poznan, Poland, v. 1. pp 1280-1284, September 2007.

[4] C. Bishop, Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[5] F. Grézl, J. Cernocký, "TRAP-based Techniques for Recognition of Noisy Speech", Lecture Notes in Computer Science, c. 9, DE, s. 270-277, ISBN 978-3-540-74627-0, ISSN 0302-9743, 2007.

[6] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett and N. Dahlgren. DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology, 1990.

[7] J. Li, Y. Tsao and C.-H. Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," in Proc. ICASSP05, Philadelphia, 2005.

[8] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov models", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37 (11), November 1989, pp. 1642-1648.

[9] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in Proc. ICNN, San Francisco, CA, 1993, pp. 586–591.

[10] P. Scanlon, D. Ellis and R. Reilly, "Using Broad Phonetic Group Experts for Improved Speech Recognition", IEEE Transactions on Audio, Speech and Language Processing, vol.15 (3), March 2007, pp 803-812.

[11] R. Dunne and N. Campbell, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function", in Proc Eighth Australasian Conf. on Neural Networks, pp. 181-185, 1997.

[12] R.C. Rose and P. Momayyez, "Integration of multiple feature sets for reducing ambiguity in ASR", in Proc ICASSP2007, April 2007.

[13] S. Young, et all, The HTK book. Revised for HTK version 3.4, Cambridge University Engineering Department, Cambridge, December 2006.

[14] S.L. Wu, B. Kingsbury, N. Morgan and S. Greenberg, "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition", in Proc. ICASSP98.