A HIERARCHICAL BROAD-CLASS CLASSIFICATION TO ENHANCE PHONEME RECOGNITION

Carla Lopes[1,2] and Fernando Perdigão[1,3]

[1]Instituto de Telecomunicações, [2]Instituto Politécnico de Leiria-ESTG, [3]Universidade de Coimbra - DEEC
Pólo II, P-3030-290 Coimbra, Portugal
phone: + 351 239 79 62 36, fax: + 351 239 79 62 93, email:{calopes, fp}@co.it.pt
web: www.it.pt

ABSTRACT

In this paper a hierarchical classification of different levels of phonetic information is proposed in order to improve phone recognition. In this paradigm several intermediate classifiers give posterior probability predictions for broad phonetic classes, achieving phone detail in the last layer. Class membership probabilities are weighted and combined in order to get a more robust phoneme prediction. A method for finding the best set of weights is also proposed based on discriminative training in a hybrid MLP/HMM system. Experiments show that the use of broad-class information enhances phone recognition. Relative improvements of 8% in Correctness and 5% in Accuracy were achieved in phoneme recognition on the TIMIT database compared to a baseline system.

1. INTRODUCTION

Speech sounds share acoustic, articulatory, phonological or other properties, with other speech sounds. Different types of speech features can be used to represent the speech signal and a speech recognition system can use different categories of speech features in order to introduce heterogeneous information into the existing Automatic Speech Recognition (ASR) system. According to the perspective adopted, the features can be phonological, [1],[2], broad phonetic groups, [3], attributes, [4], events, [5],[6], articulatory [7], etc. This additional information aims at correcting the errors made by an existing recognizer. Usually this information is captured by independent systems (Artificial Neural Networks (ANNs) [1] – [3], Hidden Markov Models (HMMs) [4] or Support Vector Machines [4]) and enters the system as additional features [1],[2],[4],[8].

The combination of different levels of phonetic detail was already investigated in several works, e.g. [3], where the outputs of four broad phonetic group classifiers are combined in order to correct or enhance a phone classifier. In this paper a similar approach is carried out with a Multi Layer Perceptron (MLP), however, with an hierarchical structure, from broad to fine phonetic detail. The class predictions form earlier classifiers are fed to the next ones in order to enhance the class discrimination in the current classifier. This serial arrangement did not beat our one-hidden layer, well-trained baseline system, with about the same number of training parameters. However, if we use the class membership predictions as priors to a final decision, then, a better phone classifier is achieved. We did this by weighting the broad-class predictions in several different ways. We demonstrate in this paper that the multi-output MLP combination approach can benefit from a trained weighted combination rule, where the weights are trained as a function of the accuracy of each phoneme in a hybrid MLP/HMM phoneme recognizer.

This paper is organized as follows. In Section 2, the hierarchical MLP classifier is described. Section 3 outlines the proposed MLP combination approach, where each phoneme is predicted by scaling 4 broad-class outputs associated with each phoneme, with 4 weights, which may be different or equal for each phoneme. In Section 4 we discuss the performance of our proposal and present the experimental results. The paper ends with some concluding comments about the proposal.

2. HIERARCHICAL MLP DESCRIPTION

An MLP network was trained for both phoneme frame classification and broad-class frame classification. Speech is analyzed every 5ms with a 15ms Hamming window. Thirty-nine parameters were used as standard input features representing 12 MFCC, plus energy, and its first and second derivatives. An additional set of temporal and spectral speech features (described in [9],[10]) were also used, resulting in 49 input features. Experiments have shown that this additional feature set has new information and actually contributes to the discrimination between classes. Two systems, one with the standard 39 input features and another with 49, with a similar number of parameters (about 124 k) were compared. Figure 1 shows the frame error rate (FER) with the two systems. In all training iterations the MLP with 49 features got the lower frame error rate. FER is about 1.3% (3% relative) lower than if we use only the standard 39 MFCC features.

The context window used is 85ms (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 centre of the context window (temporal information of past and future is included). Figure 2 illustrates the procedure. The white squares represent the frames discarded and the grey ones are the ones considered. This enlargement of the temporal

information included in training performs better than the typical context window where the used context looks only to past frames.

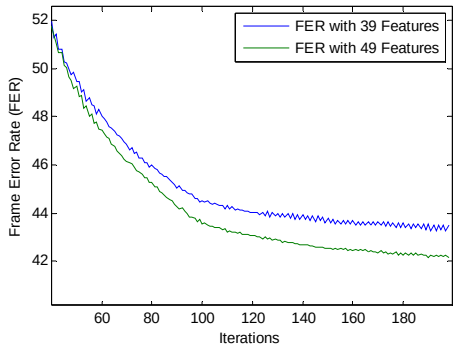

Figure 1 – FER comparison of training results when using 39 or 49 input features

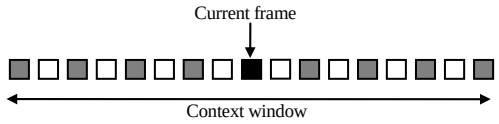

Figure 2 – Acoustic context window widening used.

The softmax function was used as the activation function of the layers with outputs/targets, so that the output values can be interpreted as posterior probabilities. The other layers (hidden layers) use a sigmoid activation function. All the network weights and bias are adjusted using batch training with a resilient back-propagation (RP) algorithm [11] 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 [12], which was later clarified by Dunne [13]. It states that the softmax activation function should couple with the cross-entropy penalty function.

The proposed system consists of 10 layers as depicted in Figure 3. The neural net has about 85k parameters and the number of nodes in the layers is (in numerical order): 50-3-50-5-50-12-50-34-50-61). The network is trained as a function of both the 61 TIMIT [14] phonemes and 4 additional sets of broad phonetic groups consisting of 3, 5, 12 and 34 broad phonetic classes. Section 2.2 describes these sets. The last layer performs a 1-to-61 classification over the set of phonemes. All layers are trained concurrently so that in training mode targets were presented at all even layers: layer 2, 4, 6, 8 and 10.

2.1 Speech Data

Both training and testing were carried out using the TIMIT database[14], and its original 61 phoneme set. This database is commonly used in phone recognition benchmarking, e.g. [1][3][15]. The training set consisted of all `si` and `sx` sentences of the original training 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.

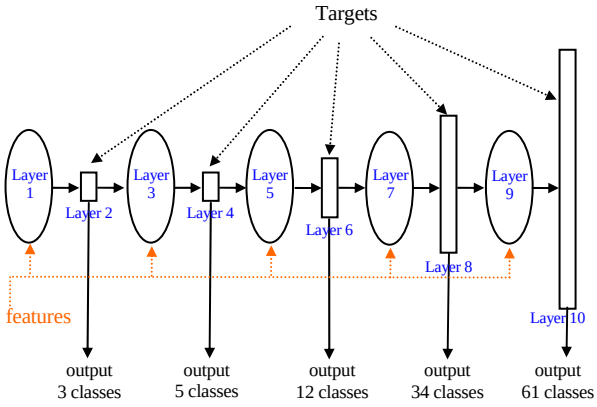

Figure 3 – Neural net architecture for phoneme classification with a hierarchical set of broad-class phonetic classifiers. Targets are defined for 5 output layers. Input features are presented to odd hidden layers.

Although the neural network is tailored to discriminate 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 as proposed by Lee and Hon [16].

2.2 Broad-Classes Description

As stated above, the proposed MLP system is trained as a function of the 61 phones and 4 additional sets of broad phonetic classes, consisting of 3, 5, 12 and 34 TIMIT phone sets. The first group classifies the signal into 3 classes: voiced, unvoiced and silence, according to the division proposed by [17]. The other sets were grouped according to the division presented in Table 1.

All the broad classes show strong agreement within some phonetic, articulatory and/or acoustic properties, each of which provides different information about the speech signal. For example, the consonant [n] shares nasality with [m], as well as complete oral closure within the set [ptk], etc. These common properties allow us to group the phonemes into broad classes, which not only describe a set of specific properties but also contrast with the remaining classes. The network has been trained and applied to a hybrid MLP/HMM system. The phone recognition results obtained using only the phone posteriors were 67.0% for Correctness and 65.6% for Accuracy, about 1% less than our baseline system, as indicated in Table 2. The baseline system has only one hidden layer, the same input layer as in Figure 2 and about the same number of weights. A possible explanation for this result is the concurrent training and a non-optimal class subdivision.

5 Classes	12 Classes	34 Classes	TIMIT Phonemes
Vowel	Vowel	v1	iy
		v2	uh uw ux
		v3	ax ax-h ah
		v4	ix ih
		v5	aa ao
		v6	eh
		v7	ae
	Dipthongs	d1	ey
		d2	aw
		d3	ay
		d4	oy
		d5	ow
	Semi-vowel	sv1	r w y
		sv2	l el
		sv3	er axr
Stop	Stop-V	stV	b d g
	Stop-uV	stuV	p t k
	Affricate	afr	jh ch
Fricative	Fricative-V	fV1	z
		fV2	zh
		fV3	v dh
	Fricative-uV	fuV1	s
		fuV2	sh
		fuV3	f th
	Whisper	w	hh hv
Nasal	Nasal	n1	en n nx
		n2	m em
		n3	ng eng
Silence	Silence	sil1	h#
		sil2	pau epi
	Closures	vcl	bcl dcl gcl
		uvcl	pcl tcl kcl
		cl1	dx
		cl2	q

Table 1 – Broad-Classes description.

3. HIERARCHICAL MLP COMBINATION APPROACH

The goal of the proposed combination approach is to take advantage of the broad-class posteriors along with the phone posteriors, in order to condition and improve the global phoneme recognition performance. Two approaches were tested. In both approaches the outputs of layer 2 were not used. One approach considers that each phone is predicted by combining 4 broad-class outputs associated with each phone, with weights different for each phone. These weights are found by means of a discriminative training method. In the other approach the broad-class posteriors are scaled by fixed weights, equal for all phones.

3.1 Discriminative Weight Training

In this approach, a weight will be assigned to the logarithm of each network output and to each phoneme. The global phoneme posteriors are found by combining the corresponding outputs of layers 4, 6, 8 and 10. The proposed combination rule is expressed in equation (1),

$$\hat{P}\left(p_k \mid \mathbf{x}\right) = \frac{1}{Z}\exp\left(\sum_{l=1}^{N_L} \alpha_{c_k}^{(l)} \log(y_{c_k}^{(l)})\right). \quad (1)$$

$\hat{P}\left(p_k \mid \mathbf{x}\right)$ is the phone probability prediction, given the observation vector $\mathbf{x}$ and broad-class predictions, $k \in \{1,..,61\}$, where k is the phoneme index and $N_L = 4$ is number of taken layers in the weighting (layers 4, 6, 8 and 10). $y_{c_k}^{(l)}$ and $\alpha_{c_k}^{(l)}$ are the network output and corresponding weight of layer l and index c_k, denoting the broad-class index to which the phoneme k belongs. Each phoneme is predicted weighting the 4 class outputs associated with the phoneme k, which are different for each phoneme. For example, for the phone [v], the 4 weights are applied to the outputs of phone *v*, class "fV3", voiced fricative class and fricative class. In equation (1) Z is a *softmax* normalization factor in order to the predictor $\hat{P}\left(p_k \mid \mathbf{x}\right)$ for the 61 phonemes sum up to one.

We tested also a common set of 4 weights for all phonemes. If all weights are 1, this corresponds to the product of the 4 outputs and gives a better result than using only the phoneme layer outputs (weights {0, 0, 0, 1}).

3.1.1 Cost Function

The best set of weights is the one which gives the highest phoneme accuracy according to our hybrid MLP/HMM recognition system. Consequently, an iterative training method based on the paradigm of discriminative training is appropriate. Every kind of error should be considered: substitutions, insertions and deletions. Since these errors are found by the *Levenshtein* distance, the objective function should include a minimization of this function. However, we used a simple 1-best discriminative function, thereby avoiding the error counting, but is better than applying the phone targets to the network output layer as is usually done.

The *Levenshtein* distance aligns two label sequences. One is the correct sequence, W_{lab}, and the other is the best decoding hypothesis given by the recognizer, W_{rec}. Using the Viterbi algorithm, we define an error function as:

$$d\left(W_{rec}, W_{lab}\right) = g(W_{rec}) - g(W_{lab}) \quad (2)$$

where $g(W_{lab})$ and $g(W_{rec})$ represent the reference and best acoustic likelihood of the observation sequence according to the Viterbi algorithm. This difference is always greater than

zero, and is only zero if the two transcriptions are exactly the same (labels and time alignments).
If N_{BD} is the total number of training utterances, the global cost is then given by:

$$E = \sum_{n=1}^{N_{BD}} d(W_{rec}^{(n)}, W_{lab}^{(n)}) \quad (3)$$

In the hybrid MLP/HMM approach the a priori probability function $b_s(\mathbf{x})$ is the likelihood of observing $\mathbf{x}$ in the HMM state s, is transformed in the posterior probability predicted by equation (1).
In order to find the appropriate set of weights ($\alpha_k^{(l)}$) a gradient descendent method is applied. In this case, it can be shown that the error gradient has terms of the form:

$$\frac{\partial}{\partial \alpha_k^{(l)}} \log \hat{P}(p_k \mid \mathbf{x}) = \log(y_k^{(l)})(1 - 1/Z) \quad (4)$$

We used the resilient back propagation algorithm in order to accelerate the convergence to a solution.

3.2 Simple Weighting

We also tested a simple weighting model that corresponds to use only 4 weights, common to all phonemes. The phoneme probability prediction is given by

$$\hat{P}(p_k \mid \mathbf{x}) = \exp\left(\sum_{l=1}^{N_L} \alpha_l \log(y_k^{(l)}) \right) \quad (5)$$

In this case the fastest way to get a solution is to test the system with a grid of weight values. Surprisingly, this simple method turns out to be as good as or even better than a set of weights per phoneme. Results for both methods are given in the next section.

4. EXPERIMENTAL RESULTS

Experiments on phoneme recognition were carried out using a hybrid MLP/HMM system, where a priori state likelihoods are replaced by posterior probabilities, according to equation (1). The HMM models were built for all phonemes by using HTK 3.4 [18] with some changes in order to replace the usual Gaussian mixture models by the outputs of the MLP. Each phoneme was modelled by a three-state left-to-right HMM and each state shares the same MLP output. The performance of the hybrid system was evaluated by means of Correctness (Corr) and Accuracy (Acc) using the HTK evaluation tool `HResults`.
Results are presented in Table 2. The first line shows the baseline system which corresponds to a network with a single hidden layer. The second line shows the results if only the last layer outputs, representing phones, is used. The third line corresponds to the product of the broad-class and phoneme posteriors, which can be seen as a joint probability of the classes. The simple inclusion of the broad-class predictions leads to a relative improvement of 9.6% in Correctness and 4.1% in Accuracy. This shows the importance of combining broad-class posteriors with the phoneme posteriors for enhanced phoneme recognition. The fourth line represents the result achieved with empirical fixed weights. The weights show that different importance should be given to the different layers. This is in agreement with the results presented in the fifth line, for which the weights were obtained with a broad-class confidence measure of the predictions. The confidence measure is computed based on the difference between the best and second best class prediction values. The weights reflect the mean values of this confidence measure across all frames.

Weights				**% Corr**	**% Acc**	**% Improvement**	
Layer 4	Layer 6	Layer 8	Layer 10			Corr	Acc
baseline				68.3	66.7		
0	**0**	**0**	**1**	67.0	65.6	-	-
1	**1**	**1**	**1**	73.5	68.3	9.6	4.1
0.6	**0.6**	**0.4**	**1**	72.2	68.9	7.7	5.2
0.82	**0.59**	**0.48**	**1**	72.4	68.9	8.1	5.1
Discriminative training				72.4	68.9	8.1	5.1

Table 2 – Phone recognition results.

The last line shows the results with the discriminative training of the weights, $\alpha_{c_k}^{(l)}$. We achieved a Correctness rate of 72.4% and an Accuracy rate of 68.9%. This is a relative improvement of 8.1% and 5.1% for Correctness and Accuracy, respectively. As required, the increment of the recognition rates was accomplished by an error decrement. As long as the weights are being updated, the Viterbi alignment converges to the reference alignment, which means that the discriminative function $g(W_{rec})$ approaches $g(W_{lab})$.

4.1 Discussion

The results are very close to those presented in [4] and [1], where the best phoneme accuracy achieved was 69.52% in the first and 70.10% in the second. Our results are not comparable with those presented in [3], because those authors evaluate their system by means of *phoneme classification* and not *phoneme recognition*, (which is a harder problem). Even though the presented results did not surpass the results presented in [1] and [4], this work shows that the use of an appropriate set of weights on the broad-class posteriors along with the phone posteriors enhances phoneme recognition. Furthermore, the results presented here may not be fully optimized, as line 4 and 5 of Table 2 may suggest. The fact that the weights do not increase with the layer hierarchy may indicate that the proposed class division is not optimized for phoneme recognition. It seems that there is no real hierarchy among the classes, as has been supposed. Future work will tackle the question of refinement of the class division.

5. CONCLUSIONS

This paper describes a hierarchical multi-layer perceptron architecture to improve phoneme recognition rate and the outcome of tests performed on it. It follows the idea that middle representations between the speech signal and the corresponding phonetic units may help phoneme recognition. Prior to recognizing phonemes, this hierarchical system recognizes several broad phonetic classes. The system is based on membership predictions in 4 layers, with 5, 12, 34 and 61 outputs. The last one corresponds to the TIMIT phonemes and the others to broad phonetic classes. The information provided by the layers was combined, and relative improvements of 8% in correctness and 5% accuracy, were achieved, compared with a system with a single hidden layer. Results do show, therefore, that the use of a suitable set of weights on the broad-class posteriors, along with the phone posteriors, enhances phoneme recognition.

6. ACKNOWLEDGEMENTS

Carla Lopes would like to thank the Portuguese foundation: Fundação para a Ciência e a Tecnologia for the PhD Grant (SFRH/BD/27966/2006).

REFERENCES

[1] J. Morris and E. Fosler-Lussier, "Further experiments with detector-based conditional random fields in phonetic recognition," in Proc of ICASSP2007, April, 2007.

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

[3] 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) , pp 803-812, March 2007.

[4] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Mattews, A. Moreno-Daniel, J. Morris, S. Siniscalchi, Y. Tsao, and Y. Wang, "Detection-based ASR in the automatic speech attribute transcription project," in Proc. of Interspeech2007, pp. 1829-1832, 2007.

[5] J. Li and C. H. Lee, "On Designing and Evaluating Speech Event Detectors," in Proc. of Interspeech 2005, Lisbon, 2005.

[6] S. Prasanna, "Event based analysis of speech," in Dept. of Computer Science and Engineering, Ph.D. Thesis: Indian Institute of Technology Madras, India, 2004.

[7] K.Y. Leung, and M. Siu, "Speech Recognition Using Combined Acoustic and Articulatory Information with Retraining of Acoustic Model Parameters," in Proc. of ICSLP2002, volume 3, pages 2117-2120, Sep, 2002.

[8] Ö. Çetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel and K. Livescu, “An articulatory feature-based tandem approach and factored observation modeling”. In Proc. of ICASSP, April 2007.

[9] C. Lopes, F. Perdigão, F., " Phonetic Recognition Improvements through Input Feature Set Combination and Acoustic Context Window Widening", 7th Conference on Telecommunications, Conftele 2009, Sta Maria da Feira, Portugal, v. 1. pp 449-452, May 2009.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] M. Mahajan, A. Gunawardana, A. Acero, "Training algorithms for hidden conditional random fields". In: Proc. ICASSP, Toulouse, France, 2006

[16] 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.

[17] T. Abu-Amer and J. Carson-Berndsen, Multi-linear HMM based system for articulatory feature extraction. In Proc. ICASSP'03, volume 2, 21–24, Hong Kong, 2003.

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

.