Improving Context Modeling for Phoneme Recognition

Daniel Vásquez[1,2], Guillermo Aradilla[1], Rainer Gruhn[1,2], Wolfgang Minker[2]
[1] *Harman Becker Automotive Systems, Speech Dialog Systems, Ulm, Germany*
[2] *University of Ulm, Institute of Information Technology, Ulm, Germany*
daniel.vasquez@uni-ulm.de

Abstract

In this paper we investigate a novel technique for modeling long temporal context in phoneme recognition. Conventional techniques take a global context consisting of consecutive concatenated acoustic features as input for a phoneme classifier. Thus, the classifier must learn different variations corresponding to different parts of the global context. In contrast to conventional techniques, our scheme decomposes the temporal context of each phoneme into a set of slices. Each temporal slice is input for a non-linear classifier given by a Multilayered Perceptron (MLP). Every MLP is trained using the central label of the global context. Thus, the assumption that the current phoneme occupies the global context is preserved. Furthermore, each MLP can robustly model different temporal context inside each phoneme. The outputs of the classifiers are combined to estimate posterior probabilities of phoneme-level classes. These posteriors are employed in a hybrid HMM/MLP framework. Experiments have shown an absolute phoneme error reduction of 3.6% compared to a baseline classifier with the same context length.

Introduction

Phoneme recognition has received much attention in the field of automatic speech recognition (ASR). A phoneme is defined as the minimal unit of speech sound in a language that can serve to distinguish meaning. Phoneme recognition is highly utilized for improving speech recognition [1]. Some further applications of phoneme recognition are found in speaker recognition [2], language identification [3] and keyword spotting [4]. For this reason, this module is required to be highly accurate.

A common and successful approach for phoneme recognition is based on Hybrid Hidden Markov Models - Multilayered Perceptrons (HMM/MLP) [5]. In a hybrid system, the MLP outputs are used as HMM state output probabilities. This method has the considerable advantage that the MLP is trained to discriminatively classify phonemes. In addition, the MLP can easily incorporate a long temporal context without making explicit assumptions. The latter property is particularly important because the characteristics of a phoneme can be spread on a long temporal context [6]. However, the amount of temporal information given to the MLP is limited by the quantity of training data and parameters of the classifier.

Different approaches have been proposed aiming to exploit the context information under the constraint of sparse training data. A general approach consists of dividing the classification task with several specialized classifiers, followed by a combination of all of them [7, 8, 9]. In this paper we present a novel approach for exploiting the temporal context of input patterns. We train several MLPs, where each one is specialized in a particular context of a phoneme. Then, the MLPs are combined by a merger. Our approach differs from other hitherto existing systems [9] since our scheme introduces overlapped slices. We will show that the introduction of overlapped slices out-performs the non-overlapped technique since the transition is better modeled.

This paper is organized as follows: next section describes the experimental setup. An explanation and detailed evaluation of the proposed approach is given in the section *context extension*. Finally, conclusions and a proposal for future work are presented in the last section.

Experimental setup

Our work is based on the TIMIT corpus [10] without the SA dialect sentences. The whole database is divided in three parts. The training data set consists of 3346 utterances spoken by 419 speakers, the cross-validation data set with 350 utterances spoken by 44 speakers and the standard test data set consisting of 1344 utterances spoken by 168 speakers. We use the 39 phoneme set given in [11] with the difference that closures are merged to the regarding burst as it was performed in [12].

We calculate feature vectors of dimensionality 39, consisting of 13 MFCCs, including log-energy, delta and double delta coefficients. The 39-dimensionality feature vectors are under global mean and variance normalization. Each feature vector is extracted from a window of 25ms of speech with a shift of 10ms.

We trained the MLPs with the Quicknet software tool [13]. Three layer perceptrons were implemented with 1000 hidden units. The number of output units corresponds to 39 and 117, for 1-state and 3-state models respectively with the softmax nonlinearity function at the output. The standard back-propagation algorithm with cross entropy error criteria is used for training the neural network. The learning rate reduction and stop training criteria are controlled by the frame error rate in cross-validation to avoid overtraining. The phoneme insertion penalty has been set to the one giving maximum phoneme accuracy in the cross-validation.

TIMIT hand labels are utilized for the initial training of the neural networks. MLPs are then re-trained em-

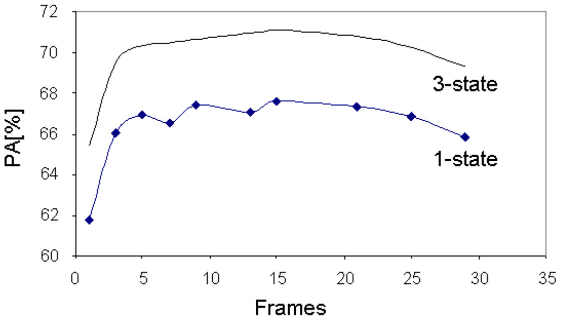


Figure 1: Varying frame window size at the input of a single MLP classifier for 1-state and 3-state models.

ploying labels from force alignment. For 3-state models, the phoneme hand labels are uniformly distributed into states before starting the iterative process.

A Viterbi decoder has been implemented with a minimum duration of three states per phoneme. Furthermore, it is assumed that all phonemes and states are equally distributed. No language model is used, unless otherwise stated. The silence is discarded for evaluation. Finally, the phoneme accuracy (PA) and frame accuracy (FA) are used as a measure of performance.

Context extension

Motivation

The main motivation of this work is to exploit as much as possible the context information, under the constraint of finite training data. A conventional way for increasing the context is performed by concatenating several feature vectors, resulting in a multi-feature vector. Thus, the multi-feature vector is given at the input of the MLP phoneme classifier. Figure 1 shows results when different feature vectors are concatenated, forming a window of different frame lengths. For a single MLP classifier, we can observe that the simple approach of increasing the number of frames of the window has an optimal point ($\sim 15 frames$) where the performance is maximized. If the number of frames is moved away from this optimal point, either by decreasing or increasing frames, the performance decreases.

To estimate how much information can be contained in a window and how many phonemes it can involve, we calculated the average number of consecutive frames per phoneme from the training data. In Figure 2 it is shown that the phoneme /oy/ in average stretches over the longest time interval, with an average number of 16.9 frames ($\sim 170ms$). The shortest phoneme according to its average number of frames is /dx/ with 2.9 frames($\sim 30ms$). The dashed line marks the average number of frames that any phoneme may occupy: 8.9 frames($\sim 90ms$).

The average number of frames per phoneme are coherent with the results given in Table 1. Having 15 frames in a window, the context encloses entirely all possible phonemes, optimizing in this way the use of a reduced training data set. If the window is highly enlarged, information of other phonemes may start to appear in the context currently considered. The neighboring phonemes cause specific coarticulation effects. This information is useful for improving phoneme classification. In contrast, several combinations of phonemes must be observed during training, requiring more training data.

In the next section we will introduce an approach in which, the context can be highly augmented, decomposing a large context into several slices without the necessity of requiring more training data.

Proposed approach

To augment the context information, the conventional method increases the window size at the input of a single MLP. In contrast we propose to introduce several MLPs which are placed at different points in time. Each MLP is trained using the same label belonging to the central frame of the entire context. Thus, the assumption that the current phoneme occupies the global context is preserved. In addition, each MLP has a fixed number of input frames and the input window of different MLPs may overlap each other. Figure 3a shows this approach. Here, three windows or slices corresponding to the input of three different MLP are overlapped. Each slice covers a total of 9 consecutive frames, i.e. 9 concatenated MFCCs feature vectors. This number was selected based on the results given in Figure 1 under constraints of reduced number of parameters and optimum performance. In addition, we chose an overlap of 4 frames between different slices. Therefore, Figure 3a covers a total of 19 frames.

In order to continue expanding the context, we derived the scheme shown in Figure 3b from Figure 3a, where two overlapped slices were added at the corners. A total of 29 frames are covered by Figure 3b. Finally, in order to test our scheme with a non-overlapped technique, we removed the overlapped slices from Figure 3b deriving the scheme shown in Figure 3c. The total number of frames covered by the last approach is 27, since there is no gap-frame between different slices.

Based on the proposed schemes, we aim to train a classifier with a large context more robustly, compared to the conventional single classifier which involves the same number of frames. In Figure 3 each classifier is based on reduced slices of the global context. This fact implies a decrease in the required amount of training data, obtaining a global classifier which is better estimated.

In order to combine the output of each classifier, we used another MLP as a merger. The input of the MLP merger consists of a concatenation of the posterior feature vectors from the classifiers to be combined. For training, the merger uses the central frame labels of the central classifier as labels.

Table 1 shows the phoneme accuracy of each classifier for the proposed approach given in Figure 3b. As it was expected, the central classifier performs the best since it is assumed that the most prominent information is in

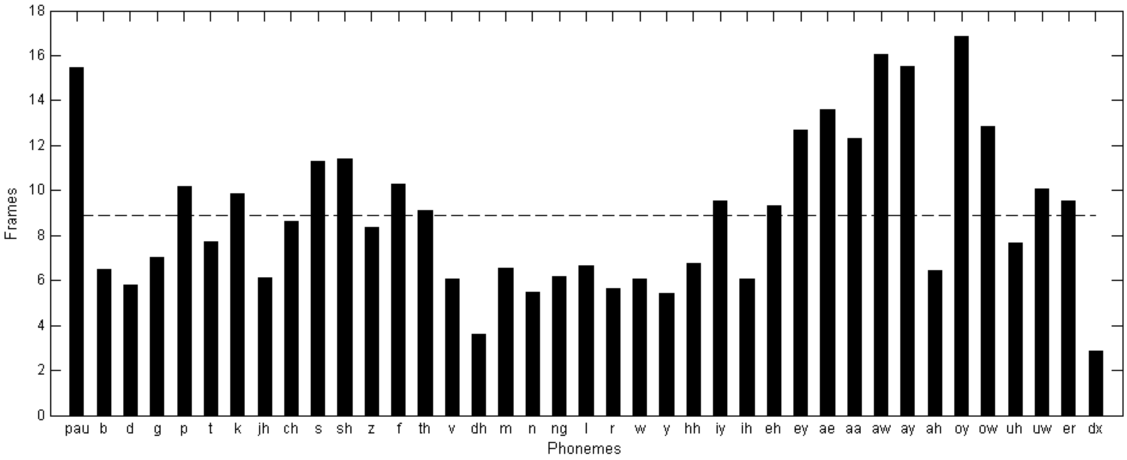
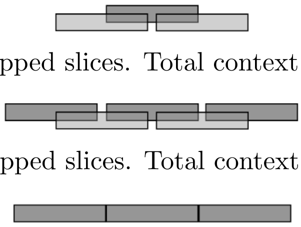


Figure 2: Average length in frames per phoneme.



(a) 3 overlapped slices. Total context: 19 frames.

(b) 5 overlapped slices. Total context: 29 frames.

(c) 3 non-overlapped slices. Total context: 27 frames.

Figure 3: Proposed context extension technique. Each slice consists of 9 concatenated MFCCs feature vectors. For the case of two different overlapped slices, they have in common 4 overlapped MFCCs

the middle of the global context. On the other hand, it can be observed that the classifiers situated at the left are slightly better than the classifiers situated at the right. Thus we can conclude that a phoneme is better characterized at its beginning rather than at its end.

Table 1: Phoneme Accuracy of each classifier given in Figure 3b.

Classifier	1-state	3-state
leftmost	55.80	60.04
left-overlapped	65.60	69.05
middle	67.37	70.64
right-overlapped	65.00	68.31
rightmost	54.32	58.91

Table 2 shows the results for the proposed approach when 1 and 3-state models are employed. In the first two rows, results for a *single classifier* with a window length of 9 and 29 frames are given, as shown in Figure 1. The following rows show the results of the proposed approaches when a MLP was used as a merger of different classifiers. Hence, the next two rows show the results when the context was further expanded from 9 frames to 19 frames and 29 frames as it is indicated in Figure 3a and Figure 3b respectively. The last row show the approach when there is non-overlapped slice which it is illustrated by Figure 3c.

Table 2: Phoneme Accuracy for the proposed temporal decomposition.

system	total frames	1-state	3-state
Single classifier	9	67.37	70.64
Single classifier	29	65.82	69.31
3 overlapped slices	19	69.20	71.78
5 overlapped slices	29	**70.10**	**72.89**
3 non-overlapped slices	27	69.37	72.39

By comparing the last two rows of Table 2, we can observe that the overlapped scheme out-performs the non-overlapped technique. The reason is that by introducing one classifier in the intersection of two classifiers, overlapping both of them, a better transition between classifiers can be modeled. Hence additional helpful information can be extracted, further improving performance. When extending the proposed approach from three overlapped slices to five overlapped slices, a further considerable improvement is achieved. Therefore, we can conclude that there is still useful information included in a larger context, which is worth considering.

Finally, comparing the results of the five overlapped slices with a single classifier covering the same number of frames (29 frames), an absolute improvement of 4.28%(1-state) and 3.58%(3-state) is obtained. It is possible to enhance the recognition accuracy with a bigram language model. For Setup B with 3-states it increases from 72.89% to 73.42%.

Further analysis of the proposed approach was conducted on the longest and shortest phoneme. Figure 2 shows that the longest phoneme is /oy/ with an average number of 16.9 frames, and the shortest is /dx/ with 2.9 frames. Figure 4 shows the frame accuracy (FA) of both phonemes when the scheme of 5 overlapped slices was tested. This figure shows the FA of each single classifier corresponding to the five different overlapped slices. In addition, it shows the FA when a MLP has been utilized

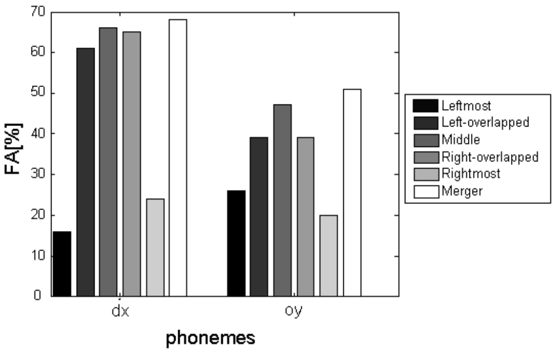


Figure 4: Frame Accuracy of each classifier given in Figure 3b. In addition, the FA of the combination of all five classifiers by using an MLP as a merger is also shown.

as a merger of the five classifiers.

As it was expected, for the case of /dx/ the leftmost and rightmost classifiers have a very low FA compared to the central classifier. The reason is that frames situated far away from the center of the global context, contain very little information relevant to the current central phoneme. In contrast, for the phoneme /oy/ the leftmost and rightmost classifiers have a considerably high performance compared to the central classifier. Finally we can see that, after applying the merger, both long and short phonemes benefit from the proposed approach by out-performing the FA of all five classifiers.

Conclusions and future work

In this work we proposed a novel approach of temporal decomposition for context expansion. We have trained several classifiers specialized in different slices of the global context. This method yields more stable phoneme classifiers in spite of sparse training data, which is clearly superior to single classifier based methods.

Several temporal decomposition approaches were evaluated, together with different classifier combiners. The best system obtained consists of a mixture of five overlapping classifiers and an MLP classifier as a merger. This approach out-performed a single classifier with the same context length with an absolute improvement of 3.6%. This shows clearly that our approach exploits contextual information more effectively than hitherto existing systems.

For comparison with the work described in [14], our proposed approach can be classified as context modeling at the feature level. In the same paper, context modeling at the posterior level (hierarchical approach) was also introduced. This technique consists of taking a window of posteriors features, generated at the feature level, as input to another MLP building a hierarchical system. In [14] a single MLP was applied at the feature level. A suggestion for feature work is to introduce our *context extension* scheme at the feature level and verify if our proposed scheme is still fruitful in the hierarchical framework.

References

[1] R. Gruhn, K. Markov, and S. Nakamura, "A statistical lexicon for non-native speech recognition," in *Proc. of Interspeech*, 2004, pp. 1497–1500.

[2] E.F.M.F. Badran and H. Selim, "Speaker recognition using artificial neural networks based on vowelphonemes," in *Proc. WCCC-ICSP*, 2000, vol. 2, pp. 796–802.

[3] M.A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," in *Proc. ICASSP*, 1995, vol. 5, pp. 3503–3506.

[4] I. Szöke, P. Schwarz, L. Burget, M. Karafiát, and J. Cernocký, "Phoneme based acoustics keyword spotting in informal continuous speech," in *Radioelektronika*, 2005, pp. 195–198.

[5] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[6] H.H. Yang, S. Van Vuuren, S. Sharma, and H Hermansky, "Relevancy of time-frequency features for phonetic classification measured by mutual information," in *Proc. ICASSP*, 1999, pp. 225–228.

[7] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. ICASSP*, 1999, vol. 1, pp. 289–292.

[8] P. Schwarz, P. Matejka, and J. Cernocký, "Towards lower error rates in phoneme recognition," in *Proc. TSD2004*, 2004, pp. 465–472.

[9] P. Schwarz, P. Matejka, and J. Cernocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, 2006, pp. 325–328.

[10] L.F. Lamel, R.H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recogn. Workshop.*, 1986, pp. 100–109.

[11] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," in *Proc. IEEE Trans. ASSP*, 1989, vol. 37, pp. 1641–1648.

[12] P. Schwarz, P. Matejka, and J. Cernocký, "Recognition of phoneme strings using trap technique," in *EUROSPEECH*, 2003, pp. 825–828.

[13] "The SPRACHcore software packages," `www.icsi.berkeley.edu/dpwe/projects/sprach`.

[14] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, "Exploiting contextual information for improved phoneme recognition," in *Proc. ICASSP*, 2008, pp. 4449–4452.