MLP BASED PHONEME DETECTORS FOR AUTOMATIC SPEECH RECOGNITION

Samuel Thomas[1], Patrick Nguyen[2], Geoffrey Zweig[2] and Hynek Hermansky[1]

[1] The Johns Hopkins University, Baltimore, USA
[2] Microsoft Corporation, Redmond, USA

{samuel,hynek}@jhu.edu
{panguyen,gzweig}@microsoft.com

ABSTRACT

Phoneme posterior probabilities estimated using Multi-Layer Perceptrons (MLPs) are extensively used both as acoustic scores and features for speech recognition. In this paper we explore a different application of these posteriors - as phonetic event detectors for speech recognition. We show how these detectors can be built to reliably capture phonetic events in the acoustic signal by integrating both acoustic and phonetic information about sound classes. These event detectors are used along with Segmental Conditional Random Fields (SCRFs) to improve the performance of speech recognition systems on the Broadcast News task.

Index Terms— Phoneme Posteriors, Multi-layer Perceptrons, Segmental Conditional Random Fields

1. INTRODUCTION

Posterior probabilities of sound classes estimated using Multi-Layer Perceptrons (MLPs) are increasingly being used to improve the performance of Automatic Speech Recognition (ASR) systems [1, 2]. These posterior probabilities have been used mainly as either local acoustic scores or as acoustic features for ASR systems. In the Hybrid Hidden Markov Model/Artificial Neural Network (HMM-ANN) approach [3], posterior probabilities estimated using MLPs have been used as emission probabilities required for HMMs. Other examples of the use of posterior probabilities as scores for ASR include, for example, detection of Out-Of-Vocabulary (OOV) segments [4]. Posterior probabilities have also been used as features using the Tandem approach [5]. In Tandem, posteriors are used as features after first compressed using a logarithm operation and then decorrelated using the KLT transform. All these approaches of using posteriors take advantage of the discriminative training used to derive the posteriors.

In this paper, we present a new application of phoneme posteriors for ASR. We use MLP based phoneme posteriors to detect phonetic events in the acoustic signal. These phoneme detectors are then used along with Segmental Conditional Random Fields (SCRFs) [6] to represent the information in the underlying audio signal.

The reminder of the paper is organized as follows - we first describe how we build phoneme detectors using posterior probabilities. Information in the acoustic signal along with phonetic and lexical knowledge is integrated into these posteriors at different levels of training the MLPs. Section 3 talks about the SCARF toolkit [7] for SCRFs that is used to integrate the phoneme detectors. In section 4 experimental results which show the usefulness of these detectors in isolation and also with other event detectors on a large vocabulary speech recognition task are presented.

2. BUILDING PHONEME DETECTORS

Multilayer perceptrons are used to estimate the posterior probability of phonemes given the acoustic evidence. Each output unit of the MLP is associated with a particular HMM state to allow these probabilities to be used as emission probabilities of a HMM system [3]. The Viterbi algorithm is then applied on the hybrid system to decode phoneme sequences. Each time frame in the acoustic signal is associated with a phoneme in the decoded output. We use the output phonemes along with their corresponding time stamps as a collection of phoneme detections. A phoneme detection is registered at the mid-point of the time span in which a phoneme is present. These phoneme detections are subsequently used in the SCARF framework.

To derive reliable detections corresponding to the underlying acoustic signal, posterior probabilities of phonetic sound classes are estimated using a hierarchical configuration of MLPs. We use both short-term spectral and long-term modulation acoustic features as input along with the hierarchical configuration to identify phonetic events. These features are derived from sub-band temporal envelopes of speech using Frequency Domain Linear Prediction (FDLP). Spectral envelope features are obtained by the short-term integration of the sub-band envelopes, the modulation frequency components are derived from the long-term evolution of the sub-band envelopes [8].

2.1. Hierarchical estimation of posteriors

In our approach of estimating posterior probabilities we use two MLPs in a hierarchical fashion to estimate posterior probabilities of phonetic sound classes as shown in Figure 1. The first MLP transforms acoustic features with a context of 9 frames to regular posterior probabilities. The second MLP in the hierarchy is trained, on posterior outputs from the first MLP. By using a context of 23 frames, we allow the second MLP to learn temporal patterns in the posterior features. These patterns include phonetic confusions at the output of the first MLP as well as the phonotactics of the language as observed in the training data [9]. The posteriors at the output of the first MLP are hence enhanced with phonetic knowledge specific to the training data language. These enhanced posteriors are used to derive phonetic detectors.

3. INTEGRATING DETECTORS WITH SCARF

An important characteristic of the SCRF approach is that it allows a set of features from multiple information sources to be integrated together to model the probability of a word sequence using a log-linear

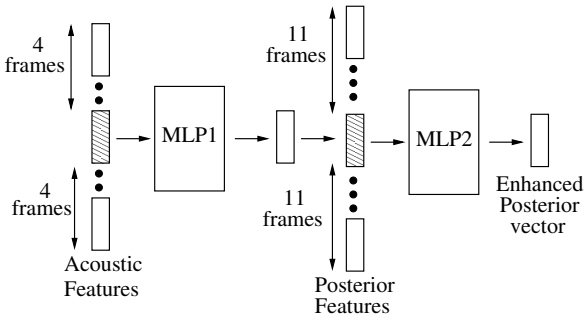

Fig. 1. Hierarchical estimation of posteriors

model. For speech recognition, SCARF uses the SCRF approach by building segment-level features that relate hypothesized words to the detections. The segment-level features are in turn, related to states of the SCRF. For automatic speech recognition these correspond to the states in an underlying finite state language model, for example a trigram LM. The set of segmentations of the observation stream are constrained to the set of possible alternatives found in the lattices generated by a conventional ASR system. SCARF uses four basic kinds of features to describe the events present in the observation stream to the words being hypothesized. These include - expectation features, levenshtein features, existence features, language model features and baseline features [7]. The expectation and levenshtein features measure the similarity between expected and observed phoneme strings, while the existence features indicate simple co-occurrence between words and phonemes. The baseline feature indicates (dis)agreement between the label on a lattice link, and the word which occurs in the same time span in a baseline decoding sequence.

The phoneme detections that we now include capture phonetic events that occur in the underlying acoustic signal. As shown in Figure 2, the phoneme detectors indicate which phonemes occur in the underlying acoustic waveform along with time stamps of when they occur. During the training process SCARF learns weights for each of the features. In the testing phase, SCARF uses the inputs from the detectors to search the constrained space of possible hypothesis.

4. EXPERIMENTS

4.1. Baseline system

We use the SCARF along with the earlier described features on the Broadcast News task. The acoustic models for this task are built using 430 hours of data based on [10]. [11] contains details of the datasets used for training both the acoustic and language models. We use the NIST dev04f set (22k words) as the development data and test on the NIST RT04f data (50k words). Table 1 summarizes the different acoustic modeling techniques that are used to build the baseline system. After training the models, IBM's Attila tool is used to produce lattices and the baseline features described in previous section. These lattices are used to build the baseline SCARF system along with baseline features and language model features (SCARF in Table 1).The improvement observed from retraining SCARF with the baseline feature could be because of the limited dynamic range of the baseline feature and the LM weight being discriminatively tuned on the training data.

For our experiments we train the MLP networks using a 2-fold cross validation also on 430 hours of broadcast news. Short-term spectral envelope (FDLP-S) and modulation frequency features

Table 1. Performance at different stages of baseline system

Stages	WER% on dev04f	WER% on RT04
Baseline system (LDA + MLLT+ VTLN + fMLLR + MLLR + fMMI + mMMI + wide beams)	16.3	15.7
SCARF with baseline features	16.0	15.4

(FDLP-M) derived from sub-band temporal envelopes of speech along with conventional PLP features are used to hierarchically estimate phoneme posterior probabilities. While the first MLP in the hierarchy is trained using 8000 hidden nodes, the second MLP uses a much simpler network with 800 hidden nodes. The MLP networks are trained using the standard back propagation algorithm with cross entropy error criteria. Both the networks use an output phoneset of 42 phones.

4.2. Oracle experiments to verify use of detectors

Before we use these phonetic detectors with SCARF, we perform a set of oracle experiments using syllable-like multi-phones [12] to verify the usefulness of acoustic event detectors with SCARF. In the first of these experiment, an oracle multi-phone detector is used with the SCARF framework. Table 2 shows the results of the oracle experiment. The experiment clearly shows that if correct acoustic detections are provided, SCARF can bring the WER down to the oracle WER.

Table 2. Performance with an oracle multi-phone detector

Setup	WER% on dev04f
SCARF	16.0
SCARF + Oracle Multi-phone Detector	11.8
Oracle	11.2

In a second oracle experiment we verify if SCARF can exploit complementary sources of information from different detection streams. We use the phoneme sequence derived from a decoding of the multi-phone units with the baseline system for this experiment. The baseline phoneme detector is used to generate new detectors based on two phonene sets. These sets are created by dividing the original phoneset of 42 phonemes into two. In the first detector, detections of phonemes in the first set are preserved. If a phoneme from the second subset occurs it is replaced by a random phoneme from among all phonemes. This procedure is reversed to create a second detector.

These corrupted streams are then used with the SCARF framework. SCARF is also trained on the original phoneme detector. The experiment uses a unigram LM and no baseline feature. Table 3 shows the results of this oracle experiment. As expected the WER increases when SCARF is trained individually on each of the streams. However when the framework is trained with both the streams together, the number drops back to the WER with the single uncorrupted stream. This experiment show that SCARF can effectively combine information from detectors that have complimentary information or in other words, SCARF recovers the baseline performance when errors in the detector streams are uncorrelated. A unigram LM is used in these experiments, accounting for the 16.9% initial WER.

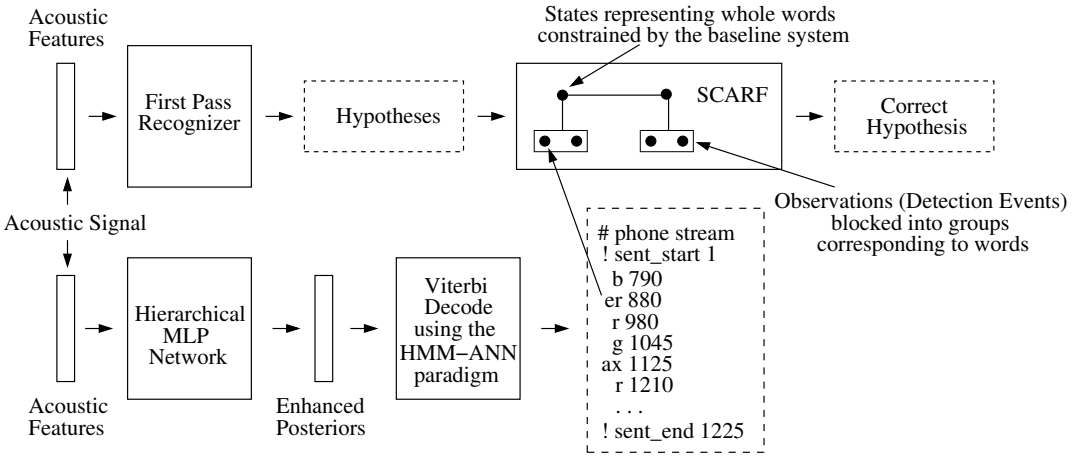

Fig. 2. Integrating MLP based phonetic detectors with SCARF

Table 3. Performance with artificially corrupted phoneme detectors

Setup	WER% on dev04f
SCARF + Uncorrupted Detector	16.9
SCARF + Detector with Phonemes in Set 1 corrupted	17.4
SCARF + Detector with Phonemes in Set 2 corrupted	17.5
SCARF + Both corrupted detectors	16.9

Table 4. Performance of phoneme detectors without baseline features

Setup	PER%	WER% on dev04f
SCARF without baseline features	12.8 (PER of baseline system)	17.9
Phoneme detector built with PLP features	32.5	17.2
Phoneme detector built with FDLP-S features	31.1	17.0
Phoneme detector built with FDLP-M features	28.9	16.9

4.3. Usefulness of individual phoneme detectors

In this set of experiments we show the usefulness of MLP based phoneme detectors when used without the baseline features. Removing the baseline features allow the phoneme detectors to be the sole source of acoustic information. These detectors are used in the SCARF framework along with the levenshtein, expectation and existence features on the dev04f development data. Table 4 shows how the phoneme detectors perform as sources of acoustic information. We observe that feature streams with lower Phoneme Error Rates (PER) provide better improvements. A trigram LM is used with SCARF for these experiments. The experiment shows that SCARF can effectively use information about phonetic events available in the phoneme detectors. It is interesting to observe that even though that the MLP based phoneme detectors have quite high PER when compared to the baseline system, they are able to provides additional information to improve recognition accuracies. Section 4.5 has a more detailed analysis of these streams.

4.4. Combining phoneme detectors with a word detector

In addition to these posterior detectors, we train a word detector stream from just the HUB4 transcribed data. The word detector has a WER of 20.4%. This stream is also used for our final experiments on the RT04 evaluation set. Table 5 shows the results of using the word detector stream along with all the phoneme detector streams in combination. In this experiment we observe further improvements with the phoneme detectors even after the word detectors have been used. Both the experiments clearly show that additional information in the underlying acoustic signal is being captured by the detectors and hence the further reduction in error rates. It should be noted that these improvements are on top of results using state-of-the-art recognition systems.

Table 5. Performance of phoneme detectors with a word detector

Setup	WER% on dev04f	WER% on RT04
Baseline Attila system	16.3	15.7
SCARF	16.0	15.4
+ Word Detector	15.3	14.5
+ Phoneme Detectors	15.1	14.3

4.5. When do phoneme detectors help the most?

In our final experiment we analyze when the proposed phoneme detectors contribute the most to reducing WER, especially when multiple detectors are used together. From the earlier oracle experiment with two corrupted phoneme streams, it is clear that evidences from multiple streams are useful if the streams are complimentary to each other. In this analysis, we first align the baseline phoneme stream with the reference sequence as shown for an example utterance in Figure 3. The phonemes sequence from the detector stream is then aligned with baseline phoneme stream. As shown in Figure 3, the following outcomes are possible when the detector streams interact

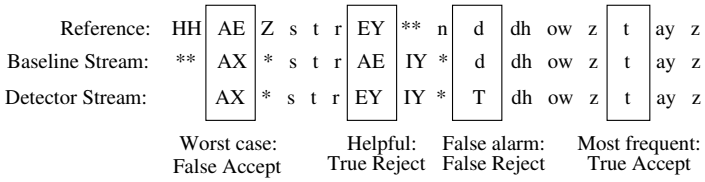

Fig. 3. Analyzing how detectors work in the SCARF framework

with the baseline stream -

- False accept cases where both the baseline stream and the detector stream agree and are wrong. In these kinds of correlated errors SCARF cannot recover.
- True reject cases where the baseline and the detector stream disagree and the detector stream is correct. These kinds of errors weaken the baseline and could result in SCARF recovering from the error.
- False alarm cases where the baseline is correct but the detector stream is wrong. Even though these kinds of errors weaken the baseline, SCARF could recover from the error using other available features like the language model feature.
- True accept cases where both the baseline and the detector are correct. These cases strengthen the correct baseline and are useful.

In Table 6 we measure these quantities for the three phoneme detectors we use. The PLP based detector stream, for example, has higher False Accepts/Alarms and lower True Accepts/Rejects. The differences between the two FDLP streams is very small; however, the PLP based detectors are worse by every measure, correlating well with its much smaller accuracy improvement in SCARF (see Table 4).

Table 6. Analysis of different phoneme detectors

Detector Stream	False Accept%	True Reject%	False Alarm%	True Accept%
PLP	26.55	73.45	26.55	72.25
FDLP-S	25.86	74.14	25.86	74.41
FDLP-M	25.98	74.02	25.98	77.15

5. CONCLUSIONS

In this paper we have explored a new application of posteriors derived using MLPs. We observe that these discriminatively trained posteriors are able to provide additional information about events in the underlying acoustic signal. Additional gains on the Broadcast News task are observed when these posterior detectors are combined with other kinds of detectors. It is evident that SCRFs can integrate multiple types of information, at different levels of granularity, and with varying degrees of quality, to improve on results from state-of-the-art speech recognition systems.

6. ACKNOWLEDGMENTS

The authors would like to thank Damianos Karakos for help with setting up the baseline recognizer. The research presented in this paper was partially funded by IARPA BEST program under contract Z857701 and DARPA RATS program under D10PC20015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IARPA or DARPA.

7. REFERENCES

[1] J. Park, F. Diehl, M.J.F. Gales, M. Tomalin and P.C. Woodland, "Training and Adapting MLP features for Arabic Speech Recognition," Proc. of IEEE ICASSP, 2009.

[2] C. Plahl, B. Hoffmeister, G. Heigold, J. Loof, R. Schluter and H. Ney, "Development of the GALE 2008 Mandarin LVCSR System," Proc. of ISCA Interspeech, 2009.

[3] H. Bourlard and N. Morgan, "Connectionist Speech Recognition A Hybrid Approach," Kluwer Academic Publishers, 1994.

[4] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C.M. White, S. Khudanpur, H. Hermansky, J. Cernocky, "Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs," Proc. of IEEE ICASSP, 2008.

[5] H. Hermansky, D. Ellis and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," Proc. of IEEE ICASSP, 2000.

[6] G. Zweig and P. Nguyen, "A Segmental CRF Approach to Large Vocabulary Continuous Speech Recognition," Proc.of IEEE ASRU, 2009.

[7] G. Zweig and P. Nguyen. "SCARF: A Segmental Conditional Random Field Toolkit for Speech Recognition," Proc. of ISCA Interspeech, 2010.

[8] S. Thomas, S. Ganapathy and H. Hermansky, "Phoneme Recognition using Spectral Envelope and Modulation Frequency Features," Proc. of IEEE ICASSP, 2009.

[9] J. Pinto, G.S.V.S. Sivaram, M. Magimai-Doss, H. Hermansky and H. Bourlard, "Analyzing MLP Based Hierarchical Phoneme Posterior Probability Estimator," IEEE Trans. on Audio, Speech, and Language Processing, 2010.

[10] S.F. Chen, B. Kingsbury, L.Mangu, D. Povey, H. Soltau, and G. Zweig, "Advances in Speech Transcription at IBM under the DARPA EARS program," IEEE Trans. on Audio, Speech, and Language Processing, vol.14, no. 5, 2006.

[11] G. Zweig, P. Nguyen et.al., "Speech Recognition with Segmental Conditional Random Fields: A Summary of the JHU CLSP Summer Workshop," Proc. of IEEE ICASSP, 2011.

[12] G. Zweig and P. Nguyen, "Maximum Mutual Information Multi-phone Units in Direct Modeling," in Proc. of ISCA Interspeech, 2009.