IDIAP RESEARCH REPORT

Using RASTA in task independent TANDEM feature extraction

Guillermo Aradilla [a] John Dines [a]
Sunil Sivadas [a] [b]

IDIAP–RR 04-22

April 2004

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP
[b] OGI School of Science and Engineering

IDIAP Research Report 04-22

Using RASTA in task independent TANDEM feature extraction

Guillermo Aradilla John Dines Sunil Sivadas

April 2004

Abstract. In this work, we investigate the use of RASTA filter in the TANDEM feature extraction method when trained with a task independent data. RASTA filter removes the linear distortion introduced by the communication channel which is demonstrated in a 18% relative improvement on the Numbers 95 task. Also, studies yielded a relative improvement of 35% over the basic PLP features by combining TANDEM features and conventional PLP features.

1 Introduction

Automatic Speech Recognition (ASR) systems are basically formed by two main subsystems: feature extraction and pattern classification. Feature extraction obtains a representation of speech (feature vectors) which must carry enough information for the pattern classification subsystem to be able to differentiate the different sub-word units, e.g. phones. The classifier uses a Hidden Markov Model (HMM) where the mapping from the hidden states to the acoustic observations, i.e. feature vectors, is typically modeled by a Gaussian Mixture Model (GMM) [1] or a Multi Layer Perceptron (MLP) [2]. Traditionally, feature vectors have been based on short-term spectrum, but other kinds of feature vectors have also been investigated which try to emphasize temporal properties of speech [3].

One of these new feature extraction methods is the TANDEM approach [4], where a MLP is used to estimate context-independent phone posterior probabilities. TANDEM features are, in principle, capable of extracting and using speech-specific but task-independent knowledge from the development data. The development data for training of the TANDEM probability estimator does not have to be directly related to the recognition task on which TANDEM is to be applied. However, since the TANDEM module is trained on the separate developement data, it acquires (as any other classifier would) peculiarities of this data. So the more similar the development data and the target application data are, the better the TANDEM approach performs.

It would be desirable to have a general, i.e. task-independent, version of the neural-net stage because of the effort required to build it in terms of time and hardware resources. In order to obtain a TANDEM feature extractor as general as possible, we can use development data that contains all of the expected sources of the anticipated nonlinguistic variability or we can pre-process the development data in order to at least alleviate the harmful variability between the development data and the target application.

In this work, we investigate the second option by applying the RASTA filter [5] to the basic features. This approach is particularly effective when the development data and the task-specific data are recorded in different communication channels since RASTA filter removes linear distortion produced by the recording environment.

As others have shown [6], merging multiple feature vectors extracted from different context lengths can be beneficial. In this work, we also combine TANDEM features obtained from our MLP trained with task independent data and PLP conventional features obtaining improved accuracy over the TANDEM system alone.

We will begin in Section 2 with a description of RASTA filter followed by a brief overview of the TANDEM approach in Section 3. In Section 4 we discuss combination of acoustic features. Then, Section 5 describes the experiment setup. Section 6 presents the results obtained. Finally, some conclusions are given in Section 7.

2 RASTA filter

PLP[7], like all the conventional feature extraction methods, is based on the short-term spectrum of speech. Such features are highly vulnerable to modification of the spectrum by the frequency response of the communication channel. RASTA filter [5] replaces the common short-term spectrum by a spectral estimate in which each frequency channel is band-pass filtered across time by a filter with sharp spectral zero at the zero frequency. In this way, it is possible to remove the linear distortion which is usually introduced by the recording environment.

RASTA filtering seems to be a good strategy for normalizing databases recorded in different environments. In this work, we use the TIMIT corpus as task independent training data, which has been recorded on microphone channel and Numbers corpus as task-specific data, which has been recorded on telephone channel. In this way, knowledge obtained by the MLP from the TIMIT corpus will be more compatible with the Numbers corpus.

3 TANDEM approach

In the TANDEM approach [4], basic features are provided as input to the MLP and its processed output is used as input for a GMM/HMM based classifier. In this way two kinds of acoustic models are used in sequence (MLP and GMM). The MLP is trained to estimate posterior probabilities using Maximum A Posteriori (MAP) criterion hence, providing more discriminative features to the GMM/HMM system which are trained on the Maximum Likelihood (ML) criterion. The MLP can also extract more information about the temporal properties of speech because it takes a larger context of frames as input and because of the non-linear transformation it performs, which is more general than the linear weighting used to compute conventional dynamic features e.g. delta features [8].

It is necessary to take the logarithm followed by a PCA decorrelation of the output of the MLP. This ensures compatibility of the feature vectors with the GMM/HMM classifier which makes assumptions of decorrelated and Gaussian-like features.

The goal of this work is to build a TANDEM feature extractor which is independent from the task of the system. The MLP is trained with a database that is not specific to any task, but contains the variability that is encountered in the test condition. We have chosen to use the TIMIT database [9] for this purpose which has the added advantage of accurate phonetic transcriptions for the training of the MLP. This database is used to train the MLP, and Numbers corpus is used to train and test the GMM/HMM system. By using a different corpus to train the MLP than that used to train the GMM classifier we are adding more information to the system. In order to minimize differences between TIMIT and Numbers corpora, the RASTA filter is applied in the implementation of the PLP feature extraction. Figure 1 shows the block diagram of our TANDEM feature extraction scheme.

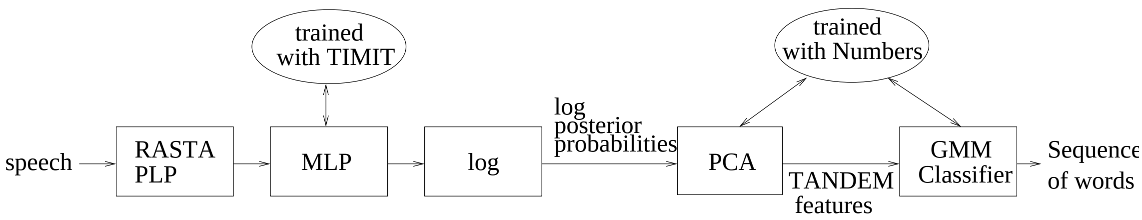

Figure 1: Block diagram of the feature extraction scheme used in this work.

4 Feature Combination

As the first stage in any speech recognition system, features are critical to the overall system performance. The ideal features reflect the relevant information in the speech signal, in our case it is the phonetic variation, while minimizing or eliminating irrelevant information, such as speaker identity or background conditions. A wide variety of features has been proposed and employed, each with different strengths and weaknesses.

There are three basic approaches of combination in speech recognition systems: feature combination (e.g. [10]), posterior combination (e.g. [11]) and hypothesis combination (e.g. [12]). In this work, we apply the first approach by combining basic PLP features and log posterior probabilities. These two feature extraction methods use the same speech signal but they extract information in a different manner so they may be suitable to be combined. The former features are based on short-term spectrum while the latter use a larger context and perform a non-linear transformation to obtain posterior probabilities. Stream combination is a technique which attempts to capitalize upon the differences in information carried by feature streams. The basic argument is that if the recognition errors of systems using the individual streams occur at different points, there is at least a chance that the combined system will be able to correct some of these errors by reference to other streams.

Also, the MLP used for the TANDEM features has been trained with a different and more general corpus so TANDEM features incorporates information that is not contained in the basic features. The method of combination that is applied in this work is the concatenation of features.

Instead of applying the PCA transform to the log posterior probabilities, it is applied to the concatenated feature vector as can be seen in Figure 2.

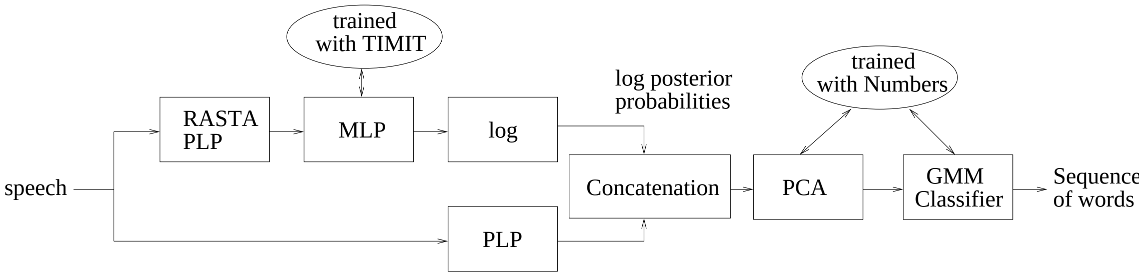

Figure 2: Block diagram of the feature extraction for the TANDEM system. Note that the MLP is trained with TIMIT but PCA and the classifier are trained with Number corpus. We use the RASTA filter with PLP because of the different corpora used with this system.

5 Experiment Description

PLP and RASTA-PLP feature vectors with 13 dimensions are extracted using the algorithm presented in the original papers [7] [5] (RASTA filter has been applied with a pole at $z = 0.98$). Their delta features are concatenated to form a 26-dimension vector.

We train the MLP from the TIMIT database using cross-entropy error criterion on 41 context-independent phones. We use 3696 training files and a tenth part of which is used as validation set. The MLP has one hidden layer with 500 units. The input consists of 6 frames left and right context (13 x 26 = 338 input units). The output is a 41-dimension vector (each output unit corresponds to a context-independent phone). For compatibility with the Numbers corpus, the TIMIT recordings are downsampled to 8KHz.

The PCA and the GMM/HMM classifier are trained from Numbers corpus. We use 6049 files to train and 2061 files to test. PCA is computed without any dimensionality reduction. The Numbers corpus contains 31 different words.

The GMM/HMM classifier has been implemented with HTK [13] using a HMM for each context-dependent phone with 3 emitting states and 12 mixtures per state.

The following experiments have been carried out:

- System 1: PLP feature vectors with 26 dimensions are directly used as inputs for the GMM/HMM classifier.
- System 2: The MLP is fed with 13 PLP feature vectors and its output is used as input for the GMM/HMM classifier.
- System 3: It is similar to the previous experiment except that RASTA-PLP implementation is used instead of PLP.

6 Results

We conducted the experiments described in Section 5. The results are presented in Table 1.

Experiments	Dimension	WER
System 1	26	6.8%
System 2	41	6.6%
System 3	41	5.4%

Table 1: WER of the experimental systems to observe the effect of RASTA channel normalization. The column Dimension indicates the number of elements contained in each feature vector.

We can see in Table 1 that RASTA filter is effectively performing displaying a relative improvemen of 18% in WER of System 3 over System 2. There is also an improvement over System 1, which uses conventional PLP features.

We have also investigated different combinations between TANDEM features and PLP features, varying the option of applying the RASTA filter. The following experiments were carried out:

- System 4: Concatenation of the log posterior probabilities derived from the output of the MLP with PLP as input features and PLP features.
- System 5: Concatenation of the log posterior probabilities derived from the output of the MLP with PLP as input features and RASTA-PLP features.
- System 6: Concatenation of the log posterior probabilities derived from the output of the MLP with RASTA-PLP as input features and RASTA-PLP features.
- System 5: Concatenation of the log posterior probabilities derived from the output of the MLP with RASTA-PLP as input features and PLP features (Figure 2).

Features	Dimension	WER
System 4	67	6.0%
System 5	67	6.0%
System 6	67	4.9%
System 7	67	4.4%

Table 2: WER of the experimental systems to test the different combination strategies. Again, the column Dimension indicates the number of parameters contained in the feature vector, in this case all have 67 dimensions (41 + 26).

As Table 2 shows, TANDEM features can work well when concatenated with conventional short-term based PLP features. The use of RASTA seems to be beneficial only in those cases where channel normalization is necessary, thus, it does not improve accuracy when it is used with task specific training data. Consequently, the best combination is TANDEM features using RASTA-PLP features combined with PLP features, showing a 35% of relative improvement regarding the basic system formed with PLP features (System 1).

7 Conclusions

In this paper we present a method for normalizing different databases in order to use them for obtaining a task independent TANDEM feature vector. RASTA filter appears to be very successful for channel normalization of features before input to the MLP obtaining a 18% relative improvement of the WER. Also, we show that the combination of TANDEM features and PLP features results in a further increase in accuracy, obtaining a 35% of relative improvement. Though RASTA works well when used with

TANDEM because of its capability of channel normalization, it does not seem to achieve a good performance in those cases where a channel normalization is not necessary, i.e. when there is no interaction between different databases.

Future work should focus on the relationship between the TANDEM feature extractor and the features with which it is trained and the use of task independent and independent training data.

References

[1] L. R. Rabiner and H. W. Juang, *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[2] N. Morgan and H. Bourlard, "An introduction to hybrid HMM/connectionist continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12(3), pp. 25–42, May 1995.

[3] H. Hermansky and S. Sharma, "Temporal Patterns (TRAPS) in ASR of Noisy Speech," *Proceedings of ICASSP*, 1999.

[4] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Proceedings ICASSP*, June 2000.

[5] H. Hermansky, "RASTA Processing of the speech," *IEEE Transactions on Speech and Audio Processing*, 1994.

[6] S. Wu, B. Kingsbury, N. Morgan, and S. Greenberg, "Performance improvements through combining phone- and syllable-length information in automatic speech recognition," *Proceedings of ICSLP*, pp. 854–857, 1998.

[7] H. Hermansky, "Perceptual Linear Predictive analysis of speech," *Journal of the Acoustic Society of America*, 1989.

[8] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *Proceedings of the IEEE*, 1986.

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

[10] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, 2001.

[11] A. Janin, D. Ellis, and N. Morgan, "Multi-stream speech recognition: Ready for prime time," *Proceedings Eurospeech*, 1999.

[12] G. Evermann and P. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," *Proceedings of the NIST Speech Transcription Workshop*, 2000.

[13] S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," tech. rep., Cambridge University, 1993.