Institut Eurécom
Department of Multi-Media
2229, route des Crètes
B.P. 193
06904 Sophia-Antipolis
FRANCE

Research Report RR-06-166

Confidence Measures for Tandem Connectionist Feature Extraction

Mohamed Faouzi BenZeghiba, Christian Wellekens

Tel : (+33) 4 93 00 81 88
Fax : (+33) 4 93 00 82 00
Email : {mohamed.benzeghiba, christian.wellekens}@eurecom.fr

[1]Institut Eurécom's research is partially supported by its industrial members: Bouygues Télécom, France Télécom, Hitachi Europe, SFR, Sharp, ST Microelectronics, Swisscom, Texas Instruments, Thales.

Abstract

This report proposes and compares a number of tandem-like feature extraction schemes. The proposed schemes use relative phone posteriors as confidence measures estimated from the MLP outputs directly or using Gamma function. The analysis of variances shows that the proposed tandem-like features discriminate better between phone classes than the conventional tandem features. But these capabilities are lost when the complexity of the model (number of gaussians) increases. Evaluations are conducted on TIMIT database.

1 Introduction

There are two advantages -among others- that make Multi-Layer Perceptrons (MLP), particularly very useful for speech recognition tasks. First, MLPs are trained to discriminate between phone classes, focusing on the critical region to learn or to model the boundaries between phonemes. Second, under certain conditions and using the one-hot encoding paradigm, the outputs of the MLP can be interpreted as a posteriori probabilities of phone classes conditioned on the input feature vector. In hybrid Hidden Markov Models/MLP (HMM/MLP) systems [1], the phone *posterior* probabilities -estimated by the MLP outputs- are divided by the phone *prior* probabilities to obtain *scaled likelihoods*. These scaled likelihoods are then used as state or phone emission probabilities in HMM Viterbi decoding, instead of likelihood. These systems showed comparable performances with HMM/GMM systems.

Recently, MLPs are used as acoustic features extractor instead of state emission probabilities estimator [2]. In this approach, the MLP outputs are used to derive acoustic feature vectors in conventional HMM/GMM system. These features are referred to as *tandem features*. Here, the MLP is considered as a non-linear transformation technique that map the input feature vector to another but more discriminative feature vector based on phone posteriors estimate. If we define the posteriori probability as the probability of being correct, then an accurate estimation of the posteriors will results in a good discriminative features, which is the aim of the feature extraction components in speech recognition systems. Tandem features achieved significant improvements in the accuracy with context-independent acoustic models [2, 3, 4, 5].

More recently, to enhance the estimation accuracy of phone posteriors and hence the discriminant capabilities of tandem features, *phone gamma posteriors*[1] derived features are proposed [6]. They are derived using the *a posteriori* probability variable γ as defined in the HMM formalism. Compared to the previous features, the phone gamma posterior derived features showed better performances [6].

Towards better enhancement of the discriminant capabilities of tandem features, we investigate the use of local phone confidence measures (CM). These phone confidence measures are successfully used in several speech recognition applications, particularly in utterance verification [7, 8, 9]. In such applications, confidence measures are used to quantify how well the acoustic model matches the acoustic data. In this work, these confidence measures are used to enhance the discriminant capabilities of the tandem features.

The rest of the report is organized as follows: Section 2 describes the conventional tandem features extraction approach as well as the phone gamma derived features. Section 3 describes the use of confidence measures to enhance tandem and phone gamma features. Speech databases and the experimental set-up are described in Section 4. Section 5 discussed the evaluation results. Finally we give some conclusions and guidelines for future work.

2 Tandem feature extraction

Figure 1 illustrates the general Block-diagram of tandem features extraction procedure. The only difference between the approaches that will be discussed in this paper, is that the phone posteriors are postprocessed differently. So, the goal of this report is to investigate a number of posterior postprocessing schemes in order to enhance the tandem features. In this section, we will describe briefly the conventional tandem features, and the phone gamma derived features. In the next section we will discuss the use of confidence measures.

2.1 Conventional tandem features

In the conventional tandem approach [2], first an MLP is trained to estimate phone posteriori probabilities $P(q_t^k|x_t)$ of each phone q^k conditioned on the input feature vector x_t at time t. This

[1] We suppose that one state corresponds to one phone

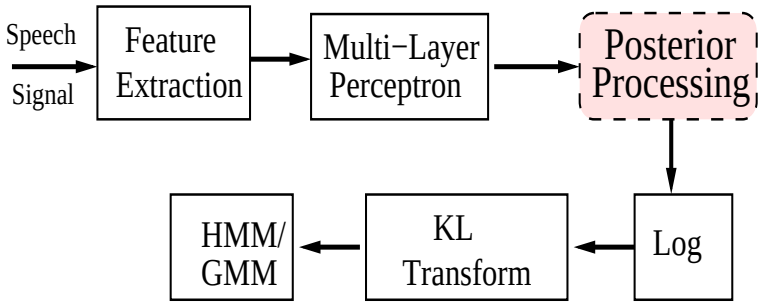

Figure 1: *Block-diagram of the connectionist feature extraction*

is done by using *softmax* activation function in the output layer of the MLP. Because the shape of the posterior distribution is sharp, they have to be gaussianized, so they can be modeled by an HMM/GMM model. The gaussianization is performed by taking the logarithm of these posteriors. The transformed phone posteriors are then decorrelated using the Karhunen-Loeve Transform (KLT). Another technique to derive the tandem features is to use the estimate of the MLP outputs before *softmax* (i.e., *linear* outputs). These estimates have a gaussian-like distribution. So, they do not need to be gaussianized, but still they need to be decorrelated using KLT. In both cases, the obtained features (after KLT) are then used in HMM/GMM system.

2.2 Phone gamma posteriors derived features

In this approach [6], the local phone posteriors $P(q_t^k|x_t)$ are postprocessed to generate *phone gamma posteriors* using gamma recursion as defined in the HMM formalism [10]. The *phone gamma posterior* $\gamma(i,t)$ is defined as the posterior probability of being in phone q^i at time t, given the observation sequence x_t^T and the HMM model M. The variable $\gamma(i,t)$ can be expressed as follows:

$$\gamma(i,t) = P(q_t^i|x_1^T, M) \tag{1}$$

As it can be observed, the estimation of gamma takes into account all the observation sequence and a priori information (if there is any) about the model M. It has been shown in [6], that when the model M is ergodic with uniform transition probabilities (i.e., no use of any a priori information), the *phone gamma posteriors*, $\gamma(i,t)$ is simply the **normalized scaled likelihood** defined as follows:

$$\gamma(i,t) = \frac{\frac{P(q_t^i|x_t)}{P(qi)}}{\sum_{k=1}^{K} \frac{P(q_t^k|x_t)}{P(q^k)}} \tag{2}$$

where K is the number of phones and $P(q^k)$ is the *a priori* probability of the phone q^k. These *phone gamma posteriors* (as defined in (2)) will be gaussianized and decorrelated using logarithm and KL transforms, respectively. The obtained features are then used as input feature vectors to the HMM/GMM system. These features will be referred to as *gamma* derived features.

3 The use of confidence measures

To enhance the discriminant capabilities of a feature vector, the feature associated with the best phone should contribute more to the discriminant information convoyed by the feature vector. In other words, the relative "goodness" of the best phone compared to the other phones has to be enhanced. To achieve this goal, we have used some *online normalization* techniques. These normalization techniques can be considered as a confidence measure that tell us how good a feature vector discriminate between the best phone and the other phones. Similar techniques have been used in speaker and utterance verification tasks [11, 12, 13].

3.1 Relative phone gamma posterior derived features

The normalization factor $\left(\sum_{k=1}^{K} \frac{P(q_t^k|x_t)}{P(q^k)}\right)$ in (2) is the sum of scaled likelihoods over all phones. However, this factor is dominated by the few highest values, corresponding to phones that are not easily separable from the best phone (under the actual MLP architecture and training procedure). In this work, we propose to increase the contribution of the normalization factor using *online phone cohort normalization*. There are several criteria for phone cohort selection. In our case, the *gamma function* in (2) is approximated as follows:

$$\gamma(i,t) \approx \frac{\frac{P(q_t^i|x_t)}{P(qi)}}{\left(\sum_{j \in C_j} \frac{P(q_t^j|x_t)}{P(q^j)}\right)^{\frac{1}{N}}} \tag{3}$$

where N is the size of the phone cohort C_j. It is worth mention here, that the size N of the cohort will depend on the performance of the MLP (i.e.; how good the posterior estimates are). If N equals 1, then the *gamma function* in (2) will be approximated as follows:

$$\gamma(i,t) \approx \frac{\frac{P(q_t^i|x_t)}{P(qi)}}{\max_{1 \leq k \leq K} \frac{P(q_t^k|x_t)}{P(q^k)}} \tag{4}$$

where K is the number of phones. This is equivalent to the Higgins criterion [12] used in speaker verification. This will enhance the relative goodness of the feature associated with the best phone. After normalization, the obtained feature vector is then used to derive tandem features using logarithm and KL transforms, respectively. These features will be referred to as *relative gamma* derived features. A modified version of the schemes defined in (3) and (4) is to deprive the cohort of the best phone from including the best phone itself. In case of (4), for example, this means that the normalization factor for the best phone will be the scaled likelihood of the second best phone. The obtained features will be referred to as *modified relative gamma* derived features.

3.2 Relative phone posterior derived features

In this scheme, the phone posteriors estimated by the MLP are normalized before using the logarithm. We performed the same *online-normalization* techniques described above. That is, the relative posterior $RP(i,t)$ of phone q^i at time t is expressed as follows:

$$RP(i,t) = \frac{P(q_t^i|x_t)}{\left(\sum_{j \in C_j} P(q_t^j|x_t)\right)^{\frac{1}{N}}} \tag{5}$$

If N equal 1, then each phone posteriori probability $P(q_t^k|x_t)$ will be divided by the best posteriori probability at time t[2]. These $RP(i,t)$ values are used as an input feature vectors to the HMM/GMM system after gaussianization and decorrelation using logarithm and KL transforms, respectively. These features will be referred to as *relative posterior* derived features. When the best phone is not included in the cohort, the obtained features are referred to as *modified relative posterior* derived features.

4 Database and Experimental set-up

The performances evaluation of different features described above are conducted using TIMIT database. The training set contains all the *si* and *sx* sentences, making a total of 3696 utterances.

[2]This confidence measure was presented first in [9] and used as emission probabilities in HMM/ANN system for utterance verification task.

For testing, we have used both the standard core test set which contains 192 and the extended test set which contains 1680 utterances. The acoustic feature vectors used to train the MLP consist of 13 MFCC with their first and second order derivatives resulting in 39 coefficients. These coefficients are calculated every 10 ms over 30 ms window, The MLP which is used to estimate the posteriors consists of 351 input units with 9 consecutive frames, 500 hidden units and 48 output units, such that each output is associated with a specific phoneme. During the MLP training, 30% of the training data is used as cross-validation set. To have the same complexity for all the HMM/GMM models and make the results comparable, the dimension of the tandem feature vector is reduced to 39 features. The HMM/GMM context-independent model consists of 48 left-to-right HMM phone models. Each HMM phone model has 3 states with 12 mixtures/state. The test is performed on the reduced phone set containing 39 phones [14]. The experiments are conducted using HTK Toolkit [15].

4.1 ANOVA analysis

To evaluate the discriminant capabilities of different tandem-like features as well as MFCC features, we used the analysis of variance (ANOVA) technique. Because we are interested only in the phone information (variation), the total variation can be then decomposed as follows [16]:

$$\sum_{total} = \sum_{between_phone} + \sum_{within_phone} \tag{6}$$

where

$$\sum_{total} = \frac{1}{N} \sum_{p} \sum_{i} (X_{pi} - \overline{X}_{..})^t (X_{pi} - \overline{X}_{..}) \tag{7}$$

$$\sum_{between_phone} = \sum_{p} \frac{N_p}{N} (X_{p.} - \overline{X}_{..})^t (X_{p.} - \overline{X}_{..}) \tag{8}$$

$$\sum_{within_phone} = \frac{1}{N} \sum_{p} \sum_{i} (X_{pi} - \overline{X}_{p.})^t (X_{pi} - \overline{X}_{p.}) \tag{9}$$

where N_p is the number of frames associated with the phone p and N is the total number of frames. $\overline{X}_{..}$ and $\overline{X}_{p.}$ are the total and phone mean vectors, respectively. The contribution of the phone variation is then computed as the $trace(\sum_{between_phone})/trace(\sum_{total})$. The higher is the contribution, the better are the features. All the features are normalized to have zero mean and unit variance. Table (1) reports the ANOVA results.

Features	Phone Contribution
MFCC	14.5%
Tandem (softmax)	17.3%
Tandem (Linear)	17.2%
Tandem (Gamma)	17.3%
Tandem (Rel. Gamma (N=1))	17.3%
Tandem (Mod. Rel. Gamma)	**21.0%**
Tandem (Rel. Post (N=1))	17.3%
Tandem (Mod. Rel. Post)	**21.4%**

Table 1: *The contribution percentage of the phone variation to the total variation for MFCC and different tandem-like features. Rel. and Mod. mean relative and modified.*

The results confirm that tandem features are more discriminant than MFCCs. The results indicate also, that the discriminant capabilities of the *modified relative gamma* and the *modified relative posterior* derived features are better than those of conventional tandem extraction schemes. So, the use of these two features might yield to better performance.

5 Evaluation results & discussion

5.1 Conventional tandem features

The first set of experiments compare the performance of the approaches described in Section (2). Results are reported in Table (2). The first line corresponds to the standard HMM/GMM system trained with MFCC features. The results show that *phone gamma posteriors* derived features are

Features	Extended set	Core set
MFCC	64.1%	63.0%
Tandem (softmax)	67.7%	66.1%
Tandem (Linear)	68.5%	67.0%
Tandem (Gamma)	68.2%	66.8%

Table 2: *Accuracy of standard HMM/GMM systems using MFCC and conventional tandem-like features.*

slightly better than *softmax* derived features. This confirm what was reported in [6]. But the best results are obtained by the *linear* derived features. Although, there is no significance differences between all tandem features. The results show also that tandem features perform significantly better than standard MFCC features.

5.2 The use of confidence measures

In Figure (2), we plot the variations of the accuracy as a function of the size of the phone cohort. We have found for $N = 1$, that the *relative gamma posteriors* derived features perform worse than their modified schemes. So, only features derived using the *modified relative gamma posteriors*. It can be seen that features with $N = 1$ perform marginally better. Because of this finding, only *relative posteriors* and their *modified* scheme derived features with $N = 1$ are evaluated (see equ: 5). Table (3) reports the accuracy of the proposed schemes described in Section (3).

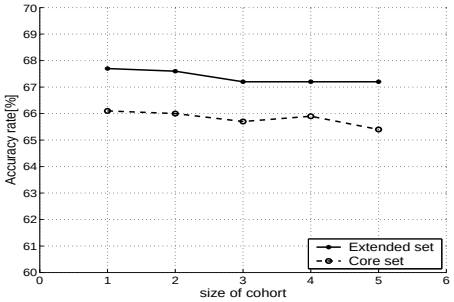

Figure 2: *The variation of the accuracy as a function of the size of the phone cohort on both the core set and the extended set using the modified relative gamma posteriors derived features.*

There are several observations that can be made from these results. First, the use of the relative gamma and relative posteriors derived features did not improve the performance, even marginally, contrary to their modified schemes. A possible reason is that by using the estimate of the best phone (scaled likelihood or posterior probability) as normalization factor, we are giving the same importance to all features, making the discriminant capabilities of the feature vector unchanged. In the modified schemes, the normalization is done in such a way that it gives more importance to the best phone, which increases the discriminant capabilities of the feature vector.

Features	Extended set	Core set
Tandem (Relative Gamma)	67.7%	66.2%
Tandem (Mod. Rel. Gamma)	68.1%	66.6%
Tandem (Relative Post)	67.8%	66.3%
Tandem (Mod. Rel. Post)	68.1%	66.5%

Table 3: *Accuracy of standard HMM/GMM systems using confidence measure derived features.*

However, the improvement obtained by the modified relative gamma and modified relative posterior derived features is far from what we would expected. As there are no significant differences in the performance compared to the *linear* derived features. A possible reason is that, in ANOVA, the distribution of phone samples is assumed to be mono-gaussian, while in HMM/GMM model, the same distribution is modeled by (3×12) gaussians. Such modeling seems to be more effective for the other tandem-like features than for the proposed ones. To check further this explanation, we have evaluated the best three tandem-like features with different number of gaussians/state. Results are plotted in Figure (3). It can be seen that the difference in the performance between the proposed features and

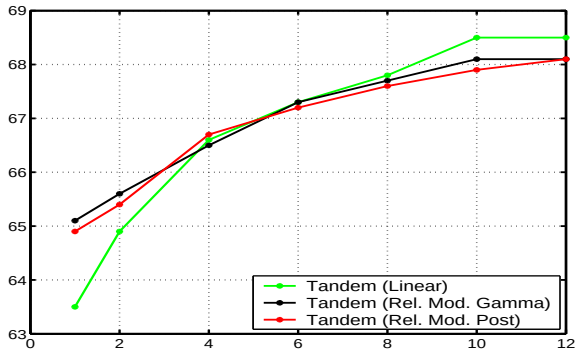

Figure 3: *Accuracy variation as a function of the complexity of the HMM/GMM systems. The evaluated features are linear, modified relative gamma and modified relative posterior derived features.*

the linear derived features is significantly large with 1 gaussian/state. But this difference tend to be negligible with an increasing number of gaussians/state.

Finally, it has been found in [5, 6], that by concatenating tandem and gamma based features with the baseline features (e.g.,MFCC) the performance can be improved. We have conducted an experiment where we have concatenated the baseline MFCC features with the different features presented in this paper. The resulting feature vectors (39+39 coefficients) are then used in the HMM/GMM system. The results are reported in Table (4). The results show no difference in the performance between different features on the extended set but on the core set the *phone gamma* derived features perform the worse.

6 Conclusion and Future work

Confidence Measures derived from the MLP outputs are successfully used in many speech recognition tasks. In this paper, we have investigated the use of relative posterior confidence measure (derived from the MLP outputs directly or using gamma function) to enhance the discriminant capabilities of the conventional tandem features. Analysis of variances has shown that the proposed variants of the tandem features discriminate better between phone classes. But these capabilities are lost with

	Extended set	Core set
MFCC+T(softmax)	69.1%	67.7%
MFCC+T(Linear)	69.2%	67.5%
MFCC+T(Gamma)	69.2%	66.6%
MFCC+T(Mod. Rel. Gamma)	69.2%	67.2%
MFCC+T(Mod. Rel. Post)	69.3%	67.4%

Table 4: *Accuracy of standard HMM/GMM system using different tandem-like features concatenated with the baseline MFCC features.*

an increasing number of gaussians/state, making them a good tradeoff between the accuracy and the complexity of the model.

7 Acknowledgments

The authors would like to thank Vivek Tyagi for fruitful discussion. This work is supported by the EC 6^{th} Framework project DIVINES under the contract number FP6-002034.

References

[1] H. Bourlard, N. Morgan "Connectionist Speech Recognition: A Hybrid Approach" *Kluwer Acadamic Publishers*, Boston, 1994.

[2] H. Hermansky, D. P. W. Ellis and S. Sharma "Connectionist Feature Extraction for Conventional HMM systems" *proceedings of ICASSP'00*, Istanbul, turkey.

[3] D. W. P. Ellis, R. Singh and S. Sivadas, "Tandem Acoustic Modeling In Large Vocabulary Recognition" *proceeding of ICASSP'01* Salt Lake City, 2001.

[4] S. Sivadas and H. Hermansky "Hierarchical Tandem Feature Extraction" *proceedings of ICASSP'02*, Orlando, Florida, USA.

[5] Q. Zhu, B. Chen, N. Morgan and A. Stolcke "On Using MLP Features In LVCSR" *ICSLP'04*,

[6] H. Bourlard, S. Bengio, M. M. Doss Q. Zhu, B. Mesot and N. Morgan " Towards Using Hierarchical Posteriors For Flexible Automatic Speech Recognition Systems" *DARPA RT-04 Workshop*, November 2004

[7] G. Williams and S. Renals, "Confidence measures for hybrid HMM/ANN speech recognition," *Proceedings of Eurospeech'97*, pp. 1955-1958, 1997.

[8] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN speech recognition systems", *Proc. of Intl. Conf. on Spoken Language Processing* (Sydney), pp. 775-779, 1998.

[9] E. Mengusoglu and C. Ris " Use of Acoustic Priori Information for Confidence Measure in ASR Applications", *EUROSPEECH'01*, pages 2557-2560

[10] L. Rabiner and B. H. Juang " Fundamentals of Speech Recognition". Prentice Hall

[11] A. E. Rosenberg and S. Parthasarathy "Speaker Background Models for Connected Digit Password Speaker Verification" *Proceeding of ICASSP'96*, pages 81-84.

[12] A. Higgins, L. Bahler and J. Porter "Speaker Verification using Randomized Phrase Prompting". *Digital Signal Processing*, vol. 1, pages 89-106.

[13] A. M. Ariyaeeinia and P. Sivakuaran "Analysis and Comparison of Score Normalization Methods For Text-Dependent Speaker Verification" *EUROSPEECH'97*, pages 1379-1382.

[14] K. F. Lee and H. W. Hon, "Speaker Independent Phone Recognition using Hidden markov Models" *IEEE Trans. Acous. Speech and Signal*, 37(11).

[15] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland " The HTK Book (For HTK Version 3.1)", Cambridge university, 2001

[16] S. Kajarekar and H. Hermansky "Analysis of Information in Speech based on MANOVA" *NIPS'02*, pp. 1189-1196