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

Research Report RR-06-185

Investigations Into Tandem Features

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

Contents

Abstract

This report proposes and evaluates a number of tandem feature extraction schemes. The proposed schemes use confidence measures estimated from the MLP outputs to derive tandem-like features. The analysis of variance shows that the proposed features discriminate better between phone classes than conventional tandem features. But they become less discriminant as the HMM model become more complex in term of number of gaussians. This report investigates also the use of contextual knowledge and its benefit to tandem based HMM system. We evaluate the use of context-dependent modeling techniques and the use of language model. Experimental results on TIMIT database show that, while tandem features, compared to standard MFCCs improve significantly the performance with context-independent models, these improvements did not generalized to context-dependent models. The same conclusion, with less effect, could be drawn for the language model. When both context-dependent and the language model are used, all features perform almost equally. This report investigates also the capacity of tandem features to handle intrinsic variabilities. Experiments are carried out using OLLO corpus.

1 Introduction

The role of speech recognizer consists of analysing the speech signal and detecting or generating the sequence of words representing the pronounced text. A main obstacle towards good performances is the high variability in the speech signal characteristics. This variability can be attributed to several sources such as the speaker himself (e.g, age, mood, dialect), the environment (e.g, background noise) and the coarticulation effect. In speech recognizer, the purpose of feature extraction component is, then, to extract from the speech signal, the relevant information that is useful to discriminate between a set of different subword units (such as phonemes) and, discard all other harmful information that depends on the speaker or the environment. In other words, the relevant information should be dependent on the discriminant characteristics of the subword units. Therefore, a good feature extraction could be a good classifier [1]. If we define the *posteriori probability* as the probability of being correct, then an accurate estimation of the posteriors will result in good discriminative features. A good candidate to this task could be a Multi-Layer Perceptron (MLP).

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 regions 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 posteriori probabilities of phone classes conditioned on the input feature vector. In hybrid Hidden Markov Models/MLP (HMM/MLP) systems [2], these phone *posterior* probabilities 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, MLP outputs are used to derive acoustic features in conventional HMM/GMM systems [3]. 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. Tandem features achieved significant improvements in the accuracy with context-independent acoustic models [3, 4, 5, 6].

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 [7]. They are derived using the *a posteriori* probability variable γ as defined in the HMM formalism. These *phone gamma posteriors* are successfully used to derive acoustic features in HMM/GMM systems [7] and, as local state score [8].

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 [9, 10, 11]. 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.

It is well known that using context-dependent modeling techniques can improve significantly the performance of the speech recognizer, as they reduce the coarticulation effect. However, an issue with the tandem features is that, while they improve significantly -compared to standard MFCC or PLP features- the accuracy with context-independent models (CI), these improvements did not carry over to context-dependent (CD) models [4, 5]. In this report, we will discuss and analyse further this issue. Based on our finding, we thought that it is worth to investigate the use of language model (bigram model) during the decoding. We would like to check if the use of linguistic context with tandem features will result in a significant or partial improvements compared to MFCCs, and whether the use of language model together with context-dependent model will improve further the performance or not.

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

It is well known that intrinsic variabilities (speaker, gender, speech rate, accents, speaking style...) affect the acoustic-phonetic properties of the speech signal in different ways, both in temporal and spectral domains. Current feature extraction and speech modeling techniques found to be sensitive to such kind of variations and performances might degrade dramatically. This report, investigates also the capacity of tandem features to handle intrinsic variabilities.

The rest of the report is organized as follows: Section 2 describes the conventional tandem features extraction approaches 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 reports and analysis the evaluation results. Experiments with intrinsic speech variabilities are described in Section 6. 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 approaches that will be discussed in this report, 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.

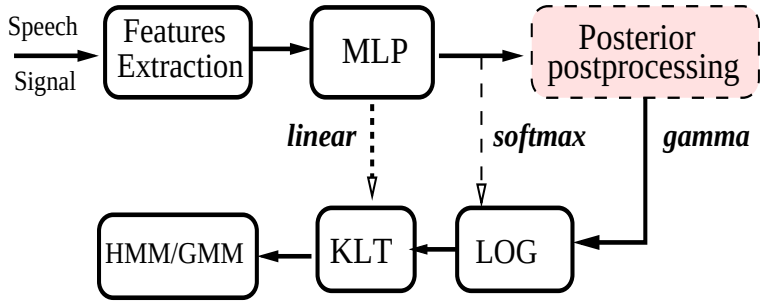

Figure 1: *Block-diagram of the tandem feature extraction*

2.1 Conventional tandem features

In the conventional tandem approach [3], 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 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 [7], 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 [12]. 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 [7], 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 [13, 14, 15].

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 this work, 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 to 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:

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

where K is the number of phones. This is equivalent to the Higgins criterion [14] 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 (after *softmax*) 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 its cohort, the obtained features are referred to as *modified relative posterior* derived features.

4 Database and Experimental set-up

To evaluate the performance of different features described above, we have used TIMIT database. The training set contains all the *si* and *sx* sentences, making a total of 3696 utterances. 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 complemented with their first and second order derivatives. These coefficients are calculated every 10 ms over 30 ms window, resulting in 39 coefficients. 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 [16]. The word insertion penalty parameter was optimized on the test set. A test of the significance of the improvements using the matched pairs sentence-segment word error is performed [17]. The experiments are conducted using HTK Toolkit [18].

5 Evaluation results & discussion

5.1 ANOVA analysis

Analysis of variance (ANOVA) is a statistical tool that is useful to identify sources of variability from several potential sources, by splitting the overall observed variance into different source variances. The method analysis the difference between two or more class means to decide if differences exist between these classes. Therefor, it can be used as a mean to evaluate the discriminant

[2]This confidence measure was presented first in [11] and used as emission probabilities in HMM/ANN system for utterance verification task.

capabilities of different acoustic features. It assumes that the samples or observations are normally distributed. Speech signal conveys several types of information such as linguistic information (due to phone characteristics), speaker information (due to speaker characteristics), speaking style information (due to speaker moods), dialect information (due to speaker regions),...etc. Each of these types of information can be considered as a source of variability of the speech signal. The goal of feature extraction component is to extract only the (desired) linguistic information and discard the other types of (undesired) information usually considered as error. So, the total variance can be decomposed into between phone variance due to differences in phone classes and within phone variance due to errors [19]:

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

where

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

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

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

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

The results indicate that tandem features are more discriminant than MFCC features. They show also, that the discriminant capabilities of the proposed *modified relative gamma* and *modified relative posterior* derived features are better than the other tandem features. We can expect that the use of these two features might yield to better performance.

5.2 Conventional tandem feature evaluation

The first set of experiments compare the performance of the approaches described in Section (2.1) and Section (2.2). Results are reported in Table (2). The first line corresponds to the baseline HMM/GMM system trained with MFCC. Results show that tandem features perform significantly better than standard MFCC features. They show also that *gamma* derived features are slightly better than *softmax* derived features. This confirm what was reported in [7]. But the best

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 the HMM baseline system using MFCC and HMM systems trained with conventional tandem-like features.*

results are obtained by the *linear* derived features. Although, there is no significance differences in the performance between all tandem feature extraction approaches.

5.3 The use of confidence measures

In Figure (2), we plot the variations of the accuracy -using *relative gamma* derived features- as a function of the size of the phone cohort. The best performance is achieved with $N = 1$, corresponding to the use of equation (4). The same conclusion was kept for *relative posteriors* derived features without conducting any further comparison experiments.

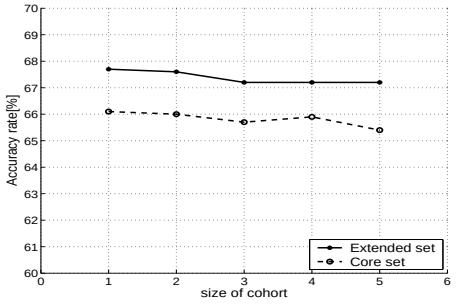

Figure 2: *The variation of the accuracy as a function of the size of the phone cohort on both the core set and the extended set using the relative gamma derived features.*

Table (3) reports the performance of these two tandem-like features compared with their modified schemes. That is at each time t, we use equation (4) for all features, except for the best one, we use the second best feature[3]. There are several observations that can be made from these results.

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 HMM systems trained with confidence measures derived features.*

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

[3]By the best feature, we mean the one with the highest scaled likelihood or the highest posterior probability.

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.

Second, 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 different feature sets with different number of gaussians/state. Curves are plotted in Figure (3).

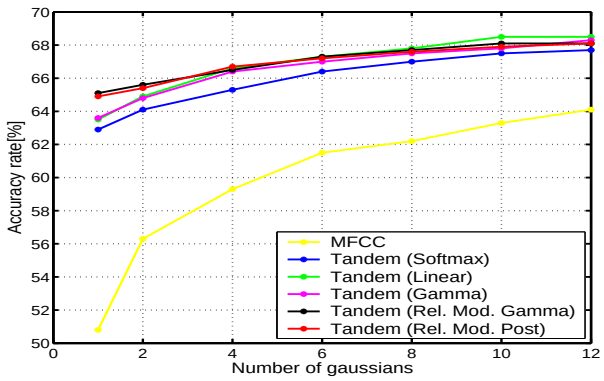

Figure 3: *Accuracy variation as a function of the complexity of the HMM system.*

It can be seen that, with one gaussian/state, HMMs trained with the proposed tandem features perform better than the conventional tandem features (Accuracies are: 62.9% for softmax, 63.5% for linear, 63.6% for gamma, 64.9% for modified relative posteriors and 65.1% for modified relative gamma). But differences in performances tend to be negligible with an increasing number of gaussians/state.

It can be observed also that the performance of the HMM trained with conventional tandem features increases gently compared to HMM's performances trained with MFCCs. When we increase the number of gaussians from 1 to 12, the absolute improvements in the performances using conventional tandem features are 4.8%, 5.0% and 4.7% for *softamx*, *linear* and *gamma* derived features, respectively, whereas this improvement is 13.4% with MFCCs. When we know that the main purpose of increasing the number of gaussians, is to better model the variability within each phone class, these observations indicate that tandem features have the advantages to reduce this variability. Moreover, the absolute improvement are only 3.2% and 3.0%, using, respectively, *modified relative gamma* and *modified relative posteriors* derived features, and the performances get saturated around 10 gaussians/state. This indicates that the within phone class variability can be further reduced using confidence measures. In [5], it has been found that HMM trained with tandem features need less data than HMM trained with standard features. Our analysis indicate that HMMs trained with the proposed confidence measures derived features need less data than conventional tandem features. In practice this is an important issue.

5.4 The use of context-dependent models

It is well known that part of the within phone variation (or the intra-phone variability) is due to the phone pronunciation which is highly dependent on the phonetic context (co-articulation effect). That is, for a given phone, the acoustic characteristics of the transition parts, at the

beginning and the end of the phone are affected differently depending on the preceding or the following phone. A good feature extraction should discard or at least reduce this co-articulation effect. In practice, a solution to this issue is to extend the phone context-independent modeling to phone context-dependent modeling.

It has been found that 27.9% of the total variation in TIMIT can be attributed to the phonetic context variation [20], indicating a great potential in modeling context-dependent speech units. In our experiments, the context-dependent triphone model uses 2015 tied states with 6 mixtures/state. The obtained results with different feature sets are reported in Table (4). It can be observed that:

	Extended set		Core set	
	CI	CD	CI	CD
MFCC	64.1%	68.1%	63.0%	67.1%
Tandem (softmax)	67.7%	68.2%	66.1%	67.7%
Tandem (Linear)	68.5%	68.9%	67.0%	67.2%
Tandem (Gamma)	68.2%	68.3%	66.8%	67.6%
Tandem (Mod. Rel. Gamma)	68.1%	68.8%	66.6%	67.7%
Tandem (Mod. Rel. Post)	68.1%	68.4%	66.5%	67.4%

Table 4: *Accuracy of HMM baseline system and HMMs trained with tandem-like features. No grammar model is used.*

- The use of context-dependent models is more beneficial to MFCC features than tandem-like features. Indeed, with MFCC features the relative improvements are 6.24%‡ and 6.5%‡, respectively[4], for extended and core sets, whereas with tandem features, the best relative improvements are only 1.0% and 2.4%, respectively, for extended and core sets. These results indicate that the context-dependent modeling is not appropriate for tandem features. The same conclusion was reported in [4, 5].

- Moreover, the difference in the performance between CD-HMM trained with MFCC and CD-HMM trained with tandem features is not significant (with a significance level 5%).

A possible reason is that the input vector to the MLP consisted of 9 consecutive frames. So, it incorporates information about context. During the MLP training, the target output (i.e., the target phone) is associated with the frame in the center of this input vector. Therefor, two frames which have the same phone label but with different left or right context, will be mapped to the same phone space represented by output vectors of the MLP. This mapping reduces the variation due to the context. As a result, the output vectors of the MLP associated with these two frames will be less dependent on the context, making the tandem-like features less effective with CD modeling than MFCCs. To consolidate this explanation and evaluate the importance of the context at the input of the MLP, we have trained an MLP with one frame as input vector (i.e, without any context). We generated tandem features using linear outputs. The performance of the context-independent HMM trained with such features is equal to 64.2%, which is similar to the HMM baseline system trained with MFCCs.

5.5 The use of language model

In the previous experiments, we found that training subword units that are dependent on the context with tandem-like features resulted in negligible improvements. Another contextual information that can be used to greatly improve the recognition accuracy is the linguistic knowledge usually embedded in the "language model". The role of the language model consists of reducing

[4]The ‡ means that this improvement is significant at the significance level 5%.

the search space to accept the possible word sequences. In the following set of experiments, we would like to evaluate the benefits of integrating such contextual knowledge in HMM based speech recognizers trained with tandem-like features. For this purpose we have trained a bigram language model. The obtained results are reported in Table (5). Several observations can be made from

	Extended set		Core set	
	CI+LM	CD+LM	CI+LM	CD+LM
MFCC	68.0%	71.4%	67.1%	71.1%
Tandem (softmax)	69.9%	71.8%	68.0%	70.6%
Tandem (Linear)	70.1%	72.1%	68.6%	71.0%
Tandem (Gamma)	70.3%	71.8%	68.3%	71.4%
Tandem (Mod. Rel. Gamma)	70.9%	72.3%	68.9%	71.6%
Tandem (Mod. Rel. Post)	70.8%	72.4%	68.9%	71.5%

Table 5: *Accuracy of HMM systems trained with MFCCs and tandem-like features. A bigram language model is used.*

these results:

- The use of language model with tandem based HMM gives better improvements compared to context-dependent models. These improvements are significant (with a significance level 5%) in case of *modified relative gamma* and *modified relative posterior* derived features.
- Performances of the tandem based HMM system are significantly better than MFCC based HMM system. This is the case for HMMs trained with *gamma*, *modified relative gamma* and *modified relative posterior* derived features.
- The contribution of the language model is much higher with MFCC based HMM than with tandem based HMM. In the former system, the relative improvements are, 6.1%‡ and 6.5%‡, respectively, on the extended and core sets, whereas, in the later system, the best relative improvements are 4.1%‡ and 3.5%‡, respectively. A possible reason is that, with a context of 9 frames at its input layer, the MLP did not only suppress some information about the phonetic context, but also it learned some linguistic knowledge. It means that some *a priori* knowledge embedded in the language model are convoyed in the tandem features. This makes the contribution of the language model in HMM system train with tandem systems less important than that in HMM system trained with MFCCs.
- As it could be expected, when the language model is used with context-dependent models, all systems perform almost equally.

6 Tandem features & intrinsic variabilities

To evaluate the capacity of tandem features in handling the intrinsic variability in speech, a set of experiments are conducted on the OLLO read speech corpus [**?**]. The OLLO corpus consists of 150 different CVC (consonant vowel consonant) and VCV (vowel consonant vowel) logatomes from 50 speakers (40 German and 10 French speakers) and with different variabilities. These include speaking rate (fast, normal, slow), speaking effort (soft, normal, loud), and speaking style (statement, question). In these experiments, only German speakers are used. The training set consists of 11927 logatoms from the normal speech. For testing, Table (6) summarizes the number of test logatoms for each variability.

The baseline HMM context-independent model trained with MFCC features consists of 26 left-to-right HMM phone models. Each phone model has 3 states with 40 mixtures/state. Unlike

	Number of logatoms
Normal	5864
Fast	17623
Slow	17888
Loud	17913
Soft	17393
Question	17862

Table 6: *Number of test logatoms for each variability.*

TIMIT database, the OLLO corpus is not manually segmented. Because the MLP training requires such segmentation, a forced Viterbi alignment was performed on the training data using the baseline MFCC system. We could also used embedded training of the MLP to enhance this segmentation and, hence the tandem features. But we have not done that. The MLP consists of 351 input units, 900 hidden units and 26 output units. Each output is associated with a specific phoneme. Once this MLP is trained, it is used to generate tandem features to train HMM models that have the same configuration as the baseline system.

Table (7) reports phone accuracies of HMM systems trained with different features. For each system, the word insertion penalty was optimized on the normal test logatoms and kept unchanged for the other variabilities.

	Normal	Fast	Slow	Loud	Soft	Question
MFCC	90.8%	85.4%	**86.7%**	81.8%	84.4%	85.0%
Tandem (softmax)	92.1%	85.7%	84.1%	83.9%	86.6%	88.0%
Tandem (Linear)	92.2%	86.0%	84.4%	**84.6%**	86.6%	**88.3%**
Tandem (Gamma)	92.3%	86.0%	84.5%	84.3%	86.7%	88.1%
Tandem (Mod. Rel. Gamma)	**92.7%**	**86.7%**	84.0%	**84.6%**	**87.0%**	87.8%
Tandem (Mod. Rel. Post)	92.5%	86.5%	83.7%	84.5%	86.7%	88.1%

Table 7: *Accuracy of HMM systems trained with normal speech and tested with different speech intrinsic variabilities.*

The main conclusion from these results is that HMM systems trained with tandem features on normal speech underwent a degradation in performances when they are tested on other speech variabilities. Although these performances are still better compared to the baseline system, in almost all variabilities, except for slow speech. It is worth mention here that these performances could be further improved if an embedded training of the MLP was performed.

In previous experiments (Section 5.3) we found that making the HMM more complex had different impacts on the HMM performance depending on the features. To check the generalization capacity of this finding, we have conducted the same analysis as in Section 5.3. In Figure (4) we plot the variation of the accuracy as a function of the complexity of the HMM. This is done for each variability and for each feature set.

It can be observed that curves have almost the same behavior (except for slow speech) as in Figure (3). This confirm that with the proposed *modified relative gamma* and *modified relative posteriors*, HMM parameters require less data to be trained than other tandem features.

7 Conclusion and Future work

This report has investigated simple techniques to enhance the discriminant capabilities of tandem features. It has then evaluated different HMM systems trained with different tandem-like

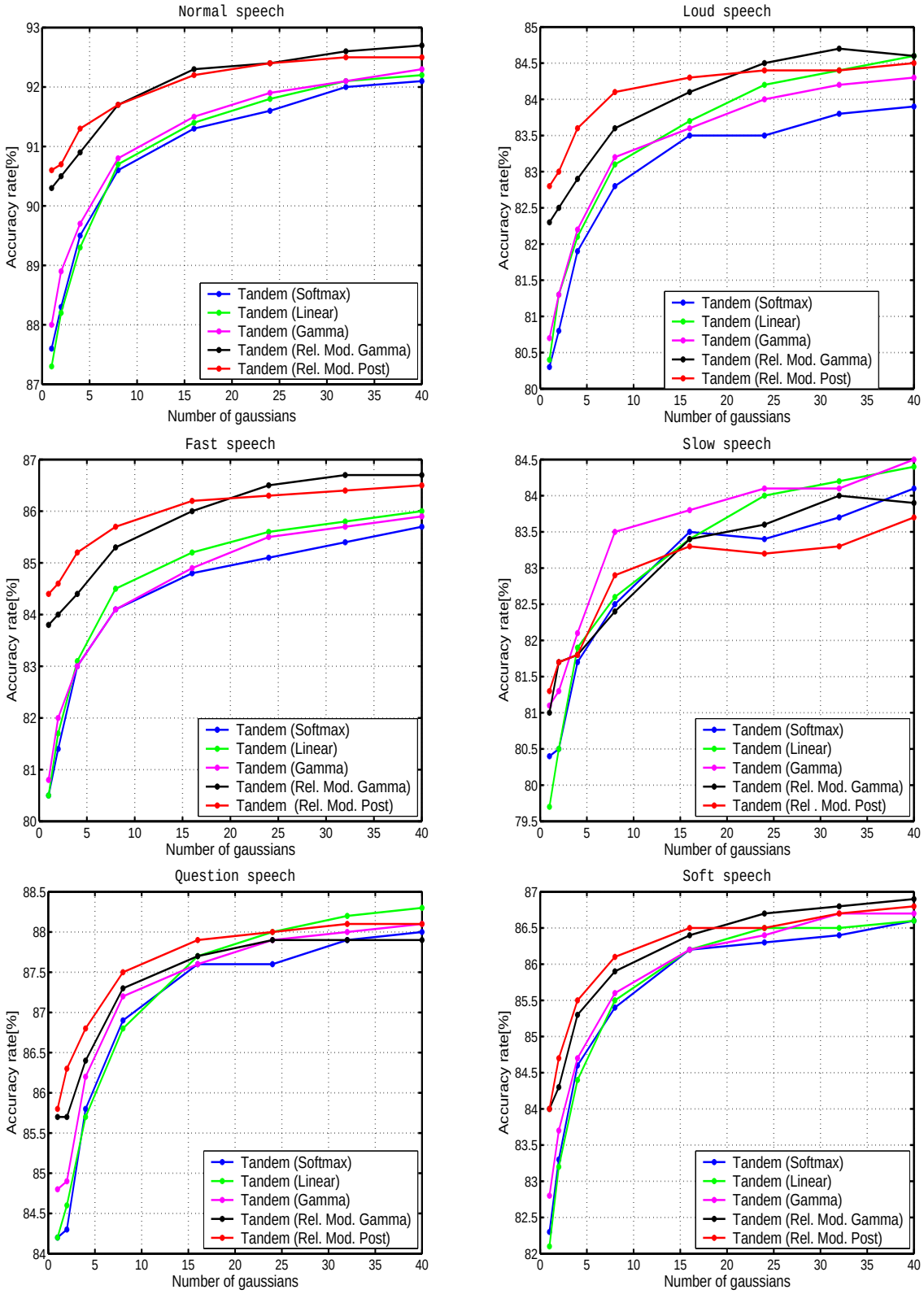

Figure 4: Accuracy variation as a function of the HMM complexity for all variabilities and different tandem feature set.

feature sets. Results are compared to a baseline HMM system trained with MFCCs. Evaluations were conducted using different acoustic modeling techniques and under different speech intrinsic variabilities. The main conclusions of this investigation are:

1. The use of confidence measures derived tandem features enhanced the discriminant capabilities of conventional tandem features. But these capabilities get reduced with more complex phone HMM model.

2. HMM trained with confidence measures derived features requires less amount of training data compared to the other features.

3. Context-dependent modeling techniques are not appropriate to tandem features. Indeed, the MLP reduces significantly the phone variability due to the context.

4. With tandem features, it is more beneficial to incorporate the language model with context-dependent models than using context-dependent models.

5. With a number of consecutive frames at its inputs, the MLP did not only reduce the coarticulation effect but also it captured some linguistic knowledge, normally embedded in the language model.

6. As the standard MFCCs features, tandem features are sensitive to the intrinsic variabilities.

As an extension to this investigation, it is worth to compare context-dependent feature extraction techniques such as the one proposed in [22] with tandem features, to really understand the role of the MLP.

8 Acknowledgments

This work is supported by the EC 6^{th} Framework project DIVINES under the contract number FP6-002034. The views expressed here are those of the authors only. The community is not liable for any use that may be made of the information contained Therine.

References

[1] R. O. Duda and P. E. Hart "Pattern Classification and Scene Analysis" *John Wiley & Sons*, 1973.

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

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

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

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

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

[7] 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

[8] H. Ketabdar, J. Vepa, S. Bengio and H. Bourlard "Developing and Enhancing Posterior Based Speech Recognition Systems" *Proceedings of Eurospeech'05*

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

[10] 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.

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

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

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

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

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

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

[17] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms" *Proceedings of ICASSP 89*, pp. 532-535.

[18] 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

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

[20] D. X. Sun and L. Deng "Analysis of Acoustic-phonetic Variation in fluent speech using TIMIT" *Proceeding of ICASSP'95*, pages 201-204, Detroit.

[21] T. Wesker, B. Meyer, K. Wagener, J. Anemller, A. Mertins, B. Kollmeier "Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines" *Interspeech 2005*, Lisbon, Portugal.

[22] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M. A. Picheny "Robust Methods for Using Context-Dependent Features and Models in a Continuous Speech Recognizer." *Proceedings of ICASSP 94*, Vol. 1, pp. 533-536