IDIAP RESEARCH REPORT

HIERARCHICAL TANDEM FEATURES FOR ASR IN MANDARIN

Joel Praveen Pinto   Mathew Magimai.-Doss
Hervé Bourlard

Idiap-RR-39-2010

NOVEMBER 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

HIERARCHICAL TANDEM FEATURES FOR ASR IN MANDARIN

Joel Pinto [1], *Mathew Magimai.-Doss* [1], *and Hervé Bourlard* [1,2]

[1] Idiap Research Institute, Martigny, Switzerland.
[2] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{jpinto, mathew, bourlard}@idiap.ch

ABSTRACT

We apply multilayer perceptron (MLP) based hierarchical Tandem features to large vocabulary continuous speech recognition in Mandarin. Hierarchical Tandem features are estimated using a cascade of two MLP classifiers which are trained independently. The first classifier is trained on perceptual linear predictive coefficients with a 90 ms temporal context. The second classifier is trained using the phonetic class conditional probabilities estimated by the first MLP, but with a relatively longer temporal context of about 150 ms. Experiments on the Mandarin DARPA GALE *eval06* dataset show significant reduction (about 7.6% relative) in character error rates by using hierarchical Tandem features over conventional Tandem features.

***Index Terms*—** Automatic speech recognition, multilayer perceptrons, Tandem features, hierarchical systems.

1. INTRODUCTION

Multilayer perceptron (MLP) classifier based acoustic modeling is being extensively used in state-of-the-art automatic speech recognition (ASR) systems [1][2][3][4]. The MLP is typically trained using standard acoustic features such as mel frequency cepstral coefficients (MFCC) or perceptual linear predictive coefficients (PLP) with a certain temporal context. A well trained MLP estimates the posterior probabilities of the output classes such as phonemes, conditioned on the input features [5]. The phonetic class conditional probabilities estimated by the MLP are typically used in hidden Markov model (HMM) based speech recognition as local emission scores in the HMM/MLP hybrid system [5] or as Tandem features to the standard HMM/GMM system [6].

More recently, we proposed an MLP based hierarchical system for estimating the phonetic class conditional probabilities [7][8][9][10]. In this hierarchical system, a second MLP classifier is trained on the phonetic class conditional probabilities (or posterior features) estimated by the MLP with a long temporal context of 150-230 ms. This hierarchical architecture is motivated towards exploiting the useful contextual information in the sequence of posterior features estimated by the MLP. A detailed analysis of the second MLP classifier using Volterra series showed that it learns the phonetic-temporal confusion patterns in the posterior features and to a certain extent the phonotactics of the language as observed in the training data. The motivation for the hierarchical architecture, its similarities/differences with previous works in the literature, and a detailed analysis of the system is presented in [7][8].

The effectiveness of the hierarchical system has been previously evaluated in recognition of phonemes in read speech recorded in clean conditions (TIMIT) as well as conversational speech recorded over a telephone channel (CTS) [8], where significant reduction in error rates have been observed. The reduction in word error rates in small vocabulary isolated word recognition was reported in [11]. In both these works the HMM/MLP hybrid system was used, where the phonetic class conditional probabilities estimated by the MLP were used as local scores in the states of the HMM. The objective of this work is to confirm if Tandem features estimated using the MLP based hierarchical architecture is useful in large vocabulary continuous speech recognition in challenging real-world scenarios.

We use the Mandarin database developed under the Global Autonomous Language Exploitation (GALE) project. On the state-of-the-art SRI-ICSI-UW ASR system [12], the hierarchical Tandem features yield an absolute (relative) reduction of 2.8% (8.7%) in character errors on broadcast conversations and a reduction of 1.1% (6.1%) in character errors on broadcast news when compared to the baseline Tandem features. The hierarchical Tandem features also outperform the standard MFCC plus pitch features. Furthermore reduction in errors is observed when hierarchical Tandem features are augmented with MFCC features.

This paper is organized as follows: In Section 2, we introduce MLP based hierarchical Tandem features. Section 3 describes the experimental setup and Section 4 presents the results. Section 5 is a discussion on some of the differences in the hierarchical system for recognition of phonemes and recognition of words. Section 6 concludes the paper.

2. HIERARCHICAL TANDEM SYSTEM

The HMM/MLP hybrid system is simple yet effective in recognition of phonemes and small vocabulary isolated word recognition. For large vocabulary continuous speech recognition, it is advantageous to use state-of-the-art modeling techniques such as context dependent modeling, state tying, speaker adaptation, etc. To this end, the Tandem approach provides an effective solution as the estimated posterior probabilities are used as features to the HMM/GMM system in the same way as conventional features. Standard Tandem features are obtained by applying Karhunen Loeve transformation (KLT) to $\log$ posterior probabilities estimated by the MLP. Fig. 1 (a) is a block schematic of the standard Tandem feature extraction used in the experiments.

The phonetic class conditional probabilities (a vector) estimated at a particular time instant represents the instantaneous soft-decision on the underlying phoneme and carries useful information such as the probability mass assigned to the competing phonemes. A temporal context on the phonetic class conditional probabilities (a sequence of vectors) carries additional contextual information such as the transition of the estimated probabilities within a phoneme and across neighboring phonemes. The second MLP in the hierarchical system is trained to learn the contextual information as shown in Fig 1 (b). Hierarchical Tandem features are obtained by applying KLT on the $\log$ posterior probabilities estimated by the second MLP.

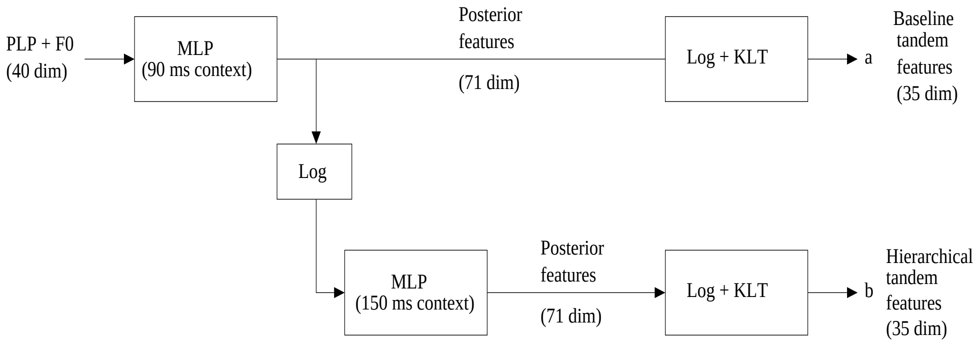

Fig. 1. (a) Standard Tandem feature extraction (b) Hierarchical Tandem feature extraction.

3. EXPERIMENTAL SETUP

The DARPA GALE task involves recognition of speech in audio segments acquired from various television programs broadcast in Mandarin. The broadcast segments include two types of genres, namely broadcast news (BN) and broadcast conversations (BC). The broadcast programs span a wide range of domains which include informal and colloquial language. In this section, we describe the experimental setup for the Mandarin ASR system.

Training and Test Data Definition

The training corpus consists of 95 hours of speech, which includes 50 hours of BN and 45 hours BC data. It is a subset of the training set of the 2008 SRI Mandarin speech-to-text system [13]. The snippet level genre classification on the training set was provided by SRI using a technique described in [14]. We use the GALE *eval06* data as the test set. The genre labels on the test set are provided by the Linguistic Data Consortium.

Hierarchical System

Fig. 1 shows the details of baseline and hierarchical Tandem feature extraction used in the experiments. The input features to the first MLP consists of the first 13 PLP cepstral coefficients appended to their delta and delta-delta parameters. As Mandarin is a tonal language, a smoothed estimate of the log-pitch value is appended to the cepstral features [15]. The 40 dimensional combined feature vector is applied at the input of the MLP with a temporal context of 90 ms. The size of the output layer of the MLP is 71 which corresponds to the number of phonemes. Only the first 35 components of the Tandem features are retained to capture at least 95% of the total variance in the data.

In the hierarchical Tandem system, a second MLP classifier is trained on the log posterior features estimated by the first MLP with a temporal context of 150 ms. This temporal context is based on the findings from task adaptation studies reported in [11], where it was observed that the word error rates begin to saturate at a context of around 130 ms - 150 ms. The output of the second MLP is transformed in the same way as the baseline system to obtain the hierarchical Tandem features.

Three layered MLP classifiers are used with sigmoid nonlinearity at the hidden layer and softmax nonlinearity at the output layer. The size of the hidden layer of the MLPs was chosen such that the total number of parameters is roughly equal to 5% of the total number of training samples.

Mandarin ASR System

We use the SRI-ICSI-UW Mandarin ASR system [15][12][13] developed for the DARPA GALE program. More specifically, we use the system setup described in [16] and this is briefly discussed here.

Speech-silence segmentation and automatic speaker clustering is first performed using Gaussian mixture modeling technique to derive “auto speakers”. The vocal tract length normalization factors are estimated for each auto speaker and are used in the estimation of MFCC features [17].

The acoustic modeling is based on the standard HMM/GMM technique. In the training phase, context independent models are first trained for each of the 71 phonemes. Context dependent models are subsequently trained and clustered down to 2000 shared Markov states, which are also known as senones. Each senone is modeled using a mixture of 32 Gaussians using phonetic decision tree based clustering. The acoustic model parameters are trained using the simple maximum likelihood criterion. Cross-word triphone modeling and speaker adaptive training is not performed in this study.

A trigram language model, which was estimated using an assortment of text corpora totalling over a billion words [17] was used for this study. The pronunciation dictionary consists of 60K characters, and is transcribed using 70 phonemes. A silence class was added, resulting in a total of 71 output classes. The decoding/testing phase involves two passes:

1. First pass search: A maximum likelihood decoding is performed using a trigram language model and the trained acoustic model to obtain one best hypothesis for each utterance. The system is referred to as the speaker independent system.

2. By using one best recognition hypothesis, the silence, vowel, and consonant regions are first identified. Constrained maximum likelihood linear regression transformation matrices are then estimated for each auto speaker. The features are subsequently transformed using these matrices.

3. Second pass search: A maximum likelihood decoding is again performed using the transformed features and the same acoustic model that was used in the first pass decoding. As feature transforms are estimated on a per speaker basis, the two pass system is referred to as the speaker adapted system.

The tunable parameters of the system, namely the language model scaling factor and the Gaussian scaling factor were fixed based on previous study [16].

Features	BC genre		BN genre		Both genres	
	SI (%)	SA (%)	SI (%)	SA (%)	SI (%)	SA (%)
mfcc-f0-42	33.4	31.0	20.9	19.3	27.0	25.0
baseline tandem-35	34.4	32.7	19.5	17.9	26.8	25.1
hierarchical tandem-35	**31.3**	**29.9**	**18.0**	**16.8**	**24.5**	**23.2**

Table 1. CERs obtained using mfcc-f0-42, baseline tandem-35, and hierarchical tandem-35 features. Boldface indicates the lowest CER.

Features	BC genre		BN genre		Both genres	
	SI (%)	SA (%)	SI (%)	SA (%)	SI (%)	SA (%)
baseline mfcc-f0-tandem-77	29.2	28.0	17.6	16.6	23.3	22.2
hierarchical mfcc-f0-tandem-77	**28.4**	**27.3**	**17.0**	**16.3**	**22.5**	**21.7**

Table 2. CERs obtained using baseline mfcc-f0-tandem-77 and hierarchical mfcc-f0-tandem-77 features. Boldface indicates the lowest CER.

Methodology

The HMM/GMM system is trained using three sets of features:

- mfcc-f0-42: The static feature vector consists of first 13 MFCC coefficients along with an estimate of the $\log$ pitch value (f0). The static features are appended to their first and second order temporal derivatives to obtain a 42 dimensional feature vector.
- tandem-35: The phoneme posterior probabilities estimated by the MLP classifier are transformed using logarithm and KLT, followed by dimensionality reduction to obtain a 35 dimensional feature vector. The tandem features are estimated in the conventional way using a single MLP classifier or the hierarchical approach as discussed in Fig. 1.
- mfcc-f0-tandem-77: Studies have shown that the best performance using Tandem features have been obtained when they are concatenated with the standard acoustic features [1]. To this end, we investigate an augmented feature vector formed by the concatenation of mfcc-f0-42 and tandem-35 features.

To distinguish between Tandem features estimated by standard single MLP approach and hierarchical approach, we prefix the features with qualifiers “baseline” and “hierarchical” respectively. For instance, baseline tandem-35 refers to Tandem feature estimated by the standard single MLP approach. The MLP classifiers were trained at Idiap Research Institute using the Quicknet toolkit. [1] Training ASR models and recognition experiments were performed at ICSI, Berkeley using the SRI Decipher system.

4. EXPERIMENTAL RESULTS

In this section, we discuss the results obtained on the GALE *eval06* test set. The speaker independent (SI) and speaker dependent (SA) systems are evaluated in terms of the character error rate (CER). The results are reported for the individual genres as well as on the entire test set.

Table 1 shows the character error rates on the *eval06* dataset obtained using mfcc-f0-42 features, baseline tandem-35 features, and hierarchical tandem-35 features. It can be seen that on broadcast conversations, the mfcc-f0-42 features yield a lower CER when compared to the baseline tandem-35 features. On broadcast news, the opposite trend is observed. Hence, on the entire test set, the mfcc-f0-42 and tandem-35 features yield similar performance. The hierarchical tandem-35 features yield the lowest CER on both broadcast news as well as broadcast conversations. These results clearly demonstrate the effectiveness of the MLP based hierarchical acoustic modeling in large vocabulary continuous speech recognition. The other main observations from this study are the following:

- The error rates on the BC genre are significantly higher when compared the BN genre as observed in some of the previous works in the literature [14]. Recognition on the BC genre is significantly harder when compared to BN because of two main reasons. Firstly, the conversational speech is spontaneous in nature and characterized by variable speaking rate, spectral reduction, [2] mispronunciations, false starts, repeated words, filled pauses, hesitations, and disfluencies. Secondly, the BC programs span a wide range of domains, which include political, economical, and cultural topics in China and around the world. In addition, the language model, which is estimated from text is more closer to the broadcast news than conversations.
- On the BC genre, the hierarchical yields an absolute decrease of 3.1% on the speaker independent system, whereas on the BN genre, the decrease in CER is about 1.5%. A similar trend was also observed in the recognition of phonemes in [8]. On the CTS task (conversations), the hierarchical approach resulted in an absolute increase of 9.0% in the phoneme accuracy over the baseline single MLP based system. On TIMIT (read speech), the improvement in the recognition accuracy was only about 3.5%.
- The decrease in CER obtained by using the hierarchical tandem-35 features over the baseline tandem-35 features is slightly higher in the case of speaker independent decoding when compared to the speaker adaptive decoding. It can be seen that on the BC (BN) genre, the decrease in CER is about 3.1% (1.5%) on the speaker independent system, whereas on speaker adapted system, the decrease is about 2.8% (1.1%).

Table 2 shows the CERs obtained using the baseline mfcc-f0-tandem-77 and hierarchical mfcc-f0-tandem-77 features. The important observations from the table in conjunction with Table 1 are as follows:

[1] http://www.icsi.berkeley.edu/Speech/qn.html

[2] When compared to read speech, the mean cepstral feature vector of a phoneme in conversational speech is closer to the global mean [18]. In addition, the variance of the cepstral coefficients is higher in spontaneous speech.

- Significant reduction in character error rates is observed when mfcc-f0-42 features are augmented with baseline tandem-35 features (a relative decrease of 9.7% on BC and 7.3% on BN compared to the best individual stream). This shows that mfcc-f0-42 and baseline tandem-35 features bear complimentary information. Further decrease in CER is observed by using hierarchical tandem-35 features. This shows that the hierarchical tandem-35 and mfcc-f0-42 features bear complimentary information in the same way baseline tandem-35 and mfcc-f0-42 features.
- The improvement in performance obtained by using hierarchical Tandem features over the baseline Tandem features is reduced when these features are augmented with mfcc-f0-42 features. This suggests that the improvement in recognition accuracies obtained by feature concatenation and hierarchical processing is not exactly additive. Nonetheless, the hierarchical mfcc-f0-tandem-77 features yield the lowest error rates in both the BC and BN genres.

To summarize, in this experimental setup, the lowest CER of 21.7% on the combined test set (BN and BC) is obtained using the hierarchical mfcc-f0-tandem-77, which is an absolute (relative) decrease of 3.3% (13.2%) over the conventional mfcc-f0-42 features.

5. DISCUSSION

Experimental results have shown that the hierarchical Tandem features outperform the conventional Tandem features in large vocabulary ASR. This is consistent with our previous study on recognition of phonemes. In this section, we discuss two differences in the hierarchical system for phoneme recognition and word recognition mainly based on empirical studies.

In recognition of phonemes, a temporal context of 210-230 ms on the posterior features yielded the lowest phoneme error rates [8]. In recognition of isolated words [11] as well as continuous words in this work, a temporal context of 150 ms yields the best performance. Beyond this point, the error rates begin to saturate. This suggests that learning the phonotactics implicitly by taking a longer temporal context on the posterior features is more useful in the recognition of phonemes than in the recognition of words. This aspect needs to be investigated further.

It was observed that slightly lower character error rates are obtained by training the second MLP using $\log$ posterior features rather than directly using raw posterior features. A similar observation was also made in recognition of isolated words. Taking a logarithm of the output of the MLP with a softmax output nonlinearity is equivalent to taking its linear activation values, except for a constant additive factor. On recognition of phonemes, however, this did not make any difference.

A possible extension of this work is to genre adaptation in similar lines as task adaptation [11]. Here, the first MLP in the hierarchical system is trained using data from different genres but the second MLP is trained on the genre-specific data. For example, by training the first MLP using data from both BN and BC genres and training the second MLP using only BC genre, we have observed further reduction in character error rates on the BC genre [7].

6. CONCLUSIONS

Experimental studies on the DARPA GALE Mandarin task show that the hierarchical Tandem features yield significantly lower character error rates when compared to the standard Tandem features as well as the traditional MFCC features. This reduction in error rates is observed in both broadcast conversations and broadcast news genres. Further reduction in error rates is obtained when the hierarchical Tandem features are augmented with standard MFCC features.

7. ACKNOWLEDGEMENTS

This work was supported in parts by the Swiss National Science Foundation through the Indo-Swiss joint research program on keyword spotting, the Swiss National Center for Competence in Research through the Interactive Multimodal Information Management (IM2) project, and by DARPA through the GALE program. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of funding agencies. We acknowledge SRI for the permission to use the Decipher ASR system and ICSI for the computational infrastructure. We also thank Wen Wang from SRI and Suman Ravuri from ICSI for their help in setting up the experiments.

8. REFERENCES

[1] N. Morgan et al., "Pushing the Envelope - Aside," *IEEE Signal Process. Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[2] A. Stolcke et al., "Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW," *IEEE Trans. Audio. Speech. Language. Process.*, vol. 14, no. 5, pp. 1729–1744, 2006.

[3] P. Fousek, L. Lamel, and J-L Gauvain, "Transcribing Broadcast Data using MLP Features," *Proc. of Interspeech*, pp. 1433–1436, 2008.

[4] J. Park, F. Diehl, M.F.J. Gales, M. Tomalin, and P.C. Woodland, "Training and Adapting MLP Features for Arabic Speech Recognition," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, pp. 4461–4464, 2009.

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

[6] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, pp. 1635–1638, 2000.

[7] J. Pinto, *Multilayer Perceptron based Hierarchical Acoustic Mmodeling for Automatic Speech Recognition*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, July 2010.

[8] J. Pinto, G. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP based Hierarchical Phoneme Posterior Probability Estimator," *IEEE Trans. Audio. Speech. Language. Process. (In press)*, 2011.

[9] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai.-Doss, "Exploiting Contextual Information for Improved Phoneme Recognition," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, pp. 4449–4452, 2008.

[10] H. Ketabdar and H. Bourlard, "Hierarchical Integration of Phonetic and Lexical Knowledge in Phone Posterior Estimation," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, pp. 4065–4068, 2008.

[11] J. Pinto, M. Magimai.-Doss, H., and H. Bourlard, "MLP Based Hierarchical System for Task Adaptation in ASR," *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[12] M.-Y. Hwang, G. Peng, M. Ostendorf, W. Wang, A. Faria, and A. Heidel, "Building a Highly Accurate Mandarin Speech Recognizer with Language-independent Technologies and Language-Dependent Modules," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 17, no. 7, pp. 1253–1262, 2009.

[13] X. Lei, W. Wu, W. Wang, A. Mandal, and A. Stolcke, "Development of the 2008 SRI Mandarin Speech-to-text System for Broadcast News and Conversation," in *Proc. of Interspeech*, 2009, pp. 2099–2103.

[14] W. Wang, A. Mandal, X. Lei, A. Stolcke, and J. Zheng, "Multifactor Adaptation for Mandarin Broadcast News and Conversation Speech Recognition," in *Proc. of Interspeech*, 2009, pp. 2103–2102.

[15] X. Lei, M. Siu, M. Ostendorf, and T. Lee, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," *Proc. of Interspeech*, pp. 1237–1240, 2006.

[16] F. Valente, M. Magimai.-Doss, C. Plahl, and S. Ravuri, "Hierarchical Processing of the Modulation Spectrum for GALE Mandarin LVCSR System," in *Proc. of Interspeech*, 2009, pp. 2963–2966.

[17] M.Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin Broadcast News Speech recognition," *Proc. of International Conference on Spoken Language Processing*, 2006.

[18] M. Nakamura, K. Iwano, and S. Furui, "Differences Between Acoustic Characteristics of Spontaneous and Read Speech and their Effects on Speech Recognition Performance," *Computer Speech and Language*, vol. 22, pp. 171–184, 2007.