IMPROVED CLUSTERED HIERARCHICAL TANDEM SYSTEM WITH BOTTOM-UP PROCESSING

Shuo-Yiin Chang and *Lin-Shan Lee*

Graduate Institute of Communication Engineering, National Taiwan University
Taiwan, Republic of China
shuoyiin@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

The outputs of multi-layer perceptron (MLP) classifiers have been successfully used in tandem systems as features for HMM-based automatic speech recognition. In a previous paper, we proposed Data-driven Clustered Hierarchical MLP (CHMLP) tandem system yielding improved performance by dividing the complicated global phone classification problem into simpler hierarchical tasks, in which specialized MLPs are trained to classify small clusters of confusing phones in a hierarchical structure. In this paper a bottom-up processing is further proposed to enhance the classification in the above CHMLP and offer even better performance. MLP rescoring for the tandem system is also investigated. The best result achieved 19.1% relative error reduction over the MFCC baseline.

Index Terms— Neural Network, Tandem system, LVCSR

1. INTRODUCTION

In recent years a great effort has been made to try to improve the performance of existing ASR system based on MFCC and HMMs. Utilizing the discriminative capabilities of artificial neural network (ANN) to help HMM has become an important direction. ANN can be trained to estimate the posterior probabilities for phonemes, which are useful information for ASR.

There have been many approaches of integrating ANN with HMM. Tandem systems [1], hybrid ANN/HMM [2], crandem systems [3] and lattice rescoring [4] are good examples. Many approaches for improving tandem systems in feature extraction or ANN structures have been proposed. Different features such as HATS [6], TRAPS [5] and MRASTA [7] carrying temporal information were shown to be complementary to short term features. The hierarchical or parallel structure MLPs [8, 10, 11, 12, 13, 14] and MLPs with two or three hidden layers [9, 15] were also shown to achieve better performance.

In a previous paper, a data-driven clustered hierarchical MLP(CHMLP) tandem system was proposed, in which the phone set is decomposed into hierarchical clusters, each consisting of confusing phones to be classified by a specialized MLP. Improved performance as compared to conventional monolithic MLP tandem system has been verified [14]. In this paper we further propose an improved CHMLP tandem system with bottom-up processing approach (CHMLP (BU)) to achieve even better performance. We also show integrating the proposed approach with word graph rescoring can offer better performance.

2. PROPOSED APPROACH

Using a monolithic MLP to optimize phone classification over the entire phone set, as was done conventionally, inevitably resulted in phone confusion, and thus limited the achievable performance. This is why in a previous paper [14] we proposed to cluster the phone set based on the confusion among the phones, and classify the phones in clusters with cluster-specific MLPs in a hierarchical structure. To create such a hierarchy of clusters based on the confusion among phones, a phonetic distance based on the confusion caused in a monolithic MLP was defined and a clustering algorithm was used [14]. Here we further propose a bottom-up approach to improve the performance.

2.1. Phonetic Distance and Hierarchical Agglomerative Clustering (HAC)

We first define a phonetic distance characterizing the confusion between each pair of phones in a monolithic MLP for clustering purpose. The distance $d(i, j)$ between phones i and j is thus defined as

$$d(i,j) = -\{w_i \log P(c_j \mid c_i) + w_j \log P(c_i \mid c_j)\} \quad (1)$$

$$P(c_j \mid c_i) = \frac{1}{n}\sum_{t=1}^{n_i} P(c_j \mid \overline{o_{it}}) \quad (2)$$

$$w_i = \frac{n_i}{n_i + n_j}, w_j = \frac{n_j}{n_i + n_j} \quad (3)$$

where c_i is phone class i , $\overline{o_{it}}$ is the t-th observation vector of phone i in the training set, n_i is the total number of such observation vectors, and $P(c_j \mid \overline{o_{it}})$ is the posterior probability for phone j obtained from a monolithic MLP given the observation vector $\overline{o_{it}}$ of phone i. It is clear that $P(c_j \mid c_i)$ represents the averaged posterior probability that the observation vector of phone i is confused as phone j, and vice versa, and w_i and w_i are used to give higher weights to more reliable posterior probabilities obtained with more data. This distance $d(i,j)$ is symmetric.

To construct the clustered phone hierarchy using the phonetic distance defined above, we exploit the hierarchical agglomerative clustering (HAC) algorithm to tie the closest phones together. The distance between two clusters C_i and C_j is defined as the average distance between all phone pairs respectively belonging to the two clusters, as in (4) below, based on the average-linkage

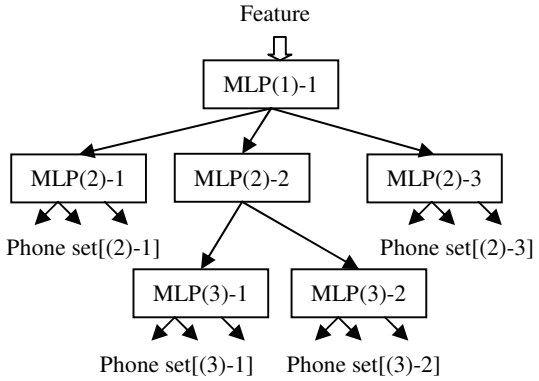

Fig.1 Clustered hierarchical MLP (CHMLP) for phone classification

agglomerative algorithm,

$$D(C_i, C_j) = \frac{1}{n_{c_i} n_{c_j}} \sum_{a \in c_i} \sum_{b \in c_j} d(a,b) \tag{4}$$

where n_C is the number of different phones in the cluster C , and a and b are two phones respectively in clusters C_i and C_j. The resulting HAC algorithm is straightforward. We first regard each phone c_i as a cluster C_i, and then find a pair of closest clusters C_i and C_j with minimum $D(C_i,C_j)$ and merge them into a new cluster. This process continues until the stop criterion is satisfied.

2.2. Clustered Hierarchical MLP (CHMLP)

The result of the above algorithm is a clustered hierarchical MLP (CHMLP) as shown in Fig. 1, where MLP(l)-k is the k-th MLP in level l of the hierarchy, and phone set [(l)-k] is classified by MLP(l)-k [14].

2.2.1. Higher level MLPs and Leaf MLPs

A clustered hierarchical classifier consists of higher-level MLPs and leaf MLPs as shown in Fig 1. The higher-level MLP, MLP(l)-k , separates a given cluster on level l into its child clusters on level l +1. Thus, given an observation vector $\overline{o_t}$, the higher-level MLP is to estimate the posterior probability $P(C_j \mid \overline{o_t})$ for $\overline{o_t}$ belonging to the child cluster C_j. On the other hand, different phones in a leaf cluster at the lower end of the hierarchy are easily confused. We thus train a specific leaf MLP for each cluster to distinguish between these competing phones [14].

2.2.2. Integration of higher-level and leaf MLPs

The process for integrating the clustered hierarchy MLP (CHMLP) structure is shown in Fig. 2. The posterior probability that each observation vector $\overline{o_t}$ belongs to each phone class c_i can be obtained by multiplying the outputs of the leaf cluster with the output of the higher level MLP immediately above the leaf cluster including c_i as in (5).

$$P(c_i \mid \overline{o_t}) = P(c_i, C_j \mid \overline{o_t}) = P(C_j \mid \overline{o_t}) P(c_i \mid \overline{o_t}, C_j), \tag{5}$$

where C_j is the leaf cluster including the phone class c_i, $P(c_i \mid \overline{o_t}, C_j)$ is estimated by the MLP for the leaf cluster C_j, and $P(C_j \mid \overline{o_t})$ is obtained from high-level MLPs;

$$P(C_j \mid \overline{o_t}) = P(C_j, C_k \mid \overline{o_t}) = P(C_k \mid \overline{o_t}) P(C_j \mid \overline{o_t}, C_k), \tag{6}$$

where C_k is the cluster immediately above the cluster C_j, etc. All these posterior probabilities $P(C_i \mid \overline{o_t})$ in (5) are then used as input to HMM.

2.2.3. CHMLP with Bottom-up Processing CHMLP (BU)

In the above structure of CHMLP, the higher-level MLPs are to perform cluster-level discrimination. However, when the number of leaf clusters becomes large, the phonetic distances between some phones belonging to different clusters may not be large enough any longer, and it may becomes difficult for the higher-level MLPs to provide good discrimination for clusters including such similar phones. This leads to the concept of bottom-up processing approach proposed here as shown in Fig. 3. It is based on exactly the same CHMLP structure as discussed above, except processed in a bottom-up manner.

Complete MFCC features for each speech frame at time t, $\overline{o_t}$ is used as the input of each leaf MLP at level l, MLP(l)-n for a phone cluster C_j, to classify all phones c_i in the phone set [(l)-n]. The output is then the posterior probabilities $P(c_i \mid \overline{o_t}, C_j)$. The parent MLP immediately above, MLP (l-1)-m for the parent phone cluster C_k, then takes all these output posterior probabilities $P(c_i \mid \overline{o_t}, C_j)$ from all its child MLPs (child clusters) as inputs, giving outputs $P(c_i \mid \overline{o_t}^{(l-1)-m}, C_k)$ for all phones c_i belonging to the cluster C_k, where $\overline{o_t}^{(l-1)-m}$ is the set of all posterior probabilities used as the input of MLP(l-1)-m, i.e. $P(c_i \mid \overline{o_t}, C_j), \forall c_i \in C_j$ and $\forall C_j \subset C_k$ for the case in Fig. 3. The outputs $P(c_i \mid \overline{o_t}^{(l-1)-m}, C_k)$ are then used as the inputs of the next higher level MLP, etc. This bottom-up process continues to the top MLP, MLP (1)-1 for cluster C_1, which includes all phones c_i, whose output $P(c_i \mid \overline{o_t}^{(1)-1}, C_1)$ is then used as the features of

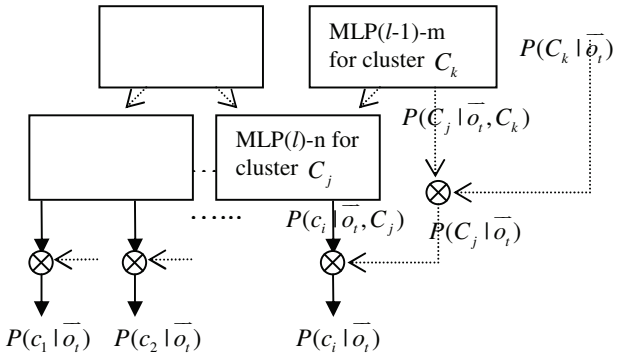

Fig. 2 Integrating hierarchical clustered MLPs (CHMLP)

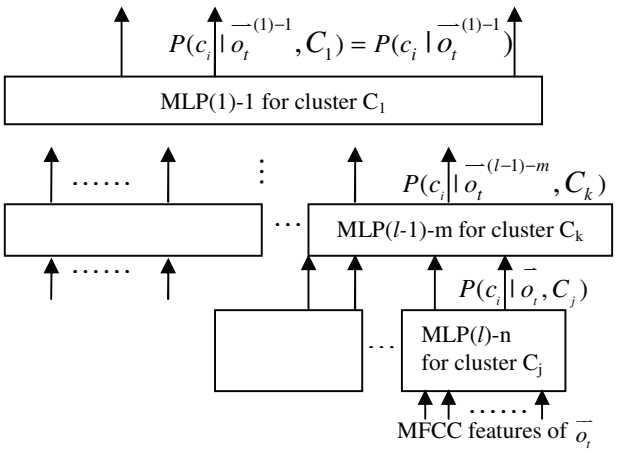

Fig. 3 CHMLP with bottom-up processing (CHMLP (BU))

HMM in the tandem system. This bottom-up process is used in both training and testing. Note here the CHMLP structure is first trained with approaches presented in Section 2.1, 2.2, in which each MLP is then re-trained using approaches in Section 2.3. In this way, the confusing phones in each leaf cluster in C_j are first discriminated by the leaf MLPs, and then the higher level MLPs further re-estimate all phone posterior probabilities from its child MLPs in a bottom-up way level by level, and the similar phones can therefore be better classified. This is referred as to as CHMLP with bottom-up processing (CHMLP (BU)). Its outputs are used as the features of HMMs after post processing (log-transformation and PCA decorrelation).

2.4. Word graph rescoring

We further improve the above tandem system by using outputs of MLPs in word graph rescoring as well. So a word arc *W* is rescored as below.

$$S(W) = \log P(o_{t_0}^{t_n} \mid W) + \lambda_l \log P(W) + \lambda_M \log P(W \mid o_{t_0}^{t_n}), \quad (7)$$

where the first two terms are scores from (tandem) acoustic and language models while the last term is from MLP, λ_l and λ_M are weight parameters and *W* is a word arc consisting of the phone sequence $c_1 c_2 .....c_n$ with boundaries $t_0, t_1, ...t_{n-1}, t_n$, obtained from the first pass HMM output,

$$P(W \mid o_{t_0}^{t_n}) \approx P(c_1 \mid o_{t_0}^{t_1}) P(c_2 \mid o_{t_1}^{t_2}) ....P(c_n \mid o_{t_{n-1}}^{tn}), \quad (8)$$

$$P(c_i \mid o_{t_{i-1}}^{t_i}) \approx \prod_{t=t_{i-1}}^{t_i} P(c_i \mid \overline{o_t}), \quad (9)$$

where the probability on the right hand side of (9) is the output of an MLP. Here we assume all posteriors of the frames are independent in (9).

3. EXPERIMENT

3.1 Experiment Setup

All experiments reported here were performed on the MATBN corpus (Mandarin Across Taiwan-Broadcast News). The training set includes 25 hours of gender-balanced broadcast news collected in Taiwan in 2001-2002. A 1.5-hour set of broadcast news collected in 2003 was used as the testing set. The baseline system used MFCC features with derivatives and accelerations (39 dims). We used two phone sets, one for MLP feature extraction and the other for HMM recognition and MLP rescoring, and defined a mapping table between the two. The former has 36 Mandarin phones. The latter included 112 right-context-dependent Initial models expanded from 22 Initials with different right contexts, 38 context independent Final models, plus a silence model. Just as the conventional tandem architecture the output posteriors of MLP were used as extra HMM features. After a log transform, PCA was further performed to reduce the dimensionality from 36 to 25, preserving 95% of the variance [3, 4]. MFCC features and MLP features were then concatenated, resulting in 64-dimension feature vectors. These concatenated feature vectors were used to train the Initial/Final HMM models. As features with more dimensions changed the range of the Gaussian mixture likelihood, a proper weight was adopted to make the range more reasonable [3].

All of MLP experiments reported used a window of 9 successive frames and 1000 hidden nodes. It was found that adding hidden nodes yielded a negligible impact on recognition results. In the rescoring case, we used bigram language model to generate word graph in the first pass, and then the posteriors from MLP and trigram language model were used in rescoring.

3.2 Baseline Results

For fair comparison, we performed a series of five baseline experiments with results listed in Table 1. The MFCC baseline is in column (1) of Table 1. We also constructed tandem system with a conventional monolithic MLP (MLP (mono)) as well as with conventional cascade two-stage monolithic MLPs (2-stage MLP(mono)), with results listed in columns(2)(3) of Table 1 respectively. In the latter case, the posteriors estimated from the first MLP were used as the input for the second stage MLP. We also implemented the rescoring process on top of the baseline tandem system in column (3), with the last term of rescoring in the right hand side of (7) obtained from either a conventional monolithic MLP (RSC by MLP (mono)) or a conventional two-stage monolithic MLP (RSC by 2-stage MLP (mono)), with results respectively listed in columns (4) and (5) of Table 1. From the character error rate (CER) listed in Table 1, incremental improvements step by step by the approaches can be easily observed.

	(1)	(2)	(3)	(4)	(5)
	MFCC baseline	(1)+ MLP(mono)	(1) + 2-stage MLP(mono)	(3) +RSC by MLP(mono)	(3) +RSC by 2-stage MLP(mono)
CER	25.33%	23.35%	22.27%	21.90%	21.30%

Table 1. Character error rate (CER) for the five baseline systems. (2)(3) tandem with conventional monolithic MLPs, (4)(5) with rescoring.

3.3 Recognition Results

In Table 2, the results for CHMLP(BU) proposed in this paper and those with rescoring are demonstrated. Since in the previous work it was found CHMLP with 2-level structures performed the best, here only three structures of CHMLP with 2 levels were compared, where CHMLP(m/n) means hierarchy of n levels and m leaves. So listed in rows (a)(b)(c) of Table 2 are CHMLP of 2 levels with 3,4

and 6 leaves. In columns labeled (2) and (3) of Table 2 are the results for Tandem systems with the previously proposed CHMLP [14] and the new approach of CHMLP(BU) proposed here, respectively to be compared with columns (2) and (3) in Table 1. For columns labeled (2) the comparison is between using conventional monolithic MLP and using CHMLP, while for columns labeled (3) the comparison is between using conventional 2-stage monolithic MLPs and the proposed 2-level hierarchical CHMLP(BU). In both cases it is clear CHMLP and CHMLP (BU) perform better. It is also clear by comparing columns labeled (3) with (2) in Table 2 that CHMLP (BU) proposed here is always better than CHMLP proposed previously. Also listed in columns labeled (4) and (5) of Table 2 are the results for rescored CHMLP(BU) Tandem systems, respectively using a conventional monolithic MLP or a conventional 2-stage monolithic MLP for rescoring, to be compared to columns (4) and (5) in Table 1.Again the new approach proposed here performed better. It is also clear that in each row of Table 2 the performance was improved step by step from left to right (columns labeled (2) to (5))

The above results are also shown in Fig 4, in which each set of 4 bars are the 4 results for columns labeled (2)(3)(4)(5) in Tables 1 and 2. The first set is for the baselines in Table 1, while the next three sets for the three rows in table 2 for three CHMLP structures. An important observation here is that for the previously proposed CHMLP (labeled (2)), the best results was obtained with 3 leaves (CHMLP (3/2)), and more leaf MLPs wasn't able to offer better accuracy. As mentioned previously, too many leaves may cause confusion in higher level MLPs. However, with the bottom-up processing proposed here (next 3 bars in each set labeled (3) (4) (5)), the accuracy were continuously improved as more leaf MLPs were used, so the above problem of CHMLP has been solved with bottom-up processing proposed here. The frame error rates (FER) of CHMLP (BU) (labeled (3)) are shown in Table 3. It can be found the decrease of FER with increase of leaf MLPs is consistent with the decrease in CER in Tables 1 and 2.

CER		(2) CHMLP	(3) CHMLP(BU)	(4) (3)+RSC by MLP(mono)	(5) (3)+RSC by 2-stage MLP(mono)
(a)	+ CHMLP(3/2)	22.30%	21.86%	21.50%	20.88%
(b)	+ CHMLP(4/2)	22.47%	21.44%	21.24%	20.72%
(c)	+ CHMLP(6/2)	22.94%	21.37%	21.01%	20.49%

Table 2. Character error rate (CER) for approaches proposed here to be compared with those in Table 1 respectively. (2) with previously proposed CHMLP(3) with bottom-up processing, (4)(5) with rescoring.

tandem system	MLP(mono)	CHMLP(3/2)	CHMLP(4/2)	CHMLP(6/2)
FER	17.52%	16.20%	15.66%	15.44%

Table 3. Frame error rate (FER) of CHMLP(BU) approaches in columns labeled(3) in Table (1) and (2).

To investigate the classification capability against confusing phones in the same leaf clusters, we evaluated the frame error rate (FER) for each leaf cluster or the percentage of frames for which the following is incorrect:

$$p(c_i \mid \overline{o_{it}}) > P(c_j \mid \overline{o_{it}}),\ \ c_i, c_j \in C_k \text{ and } i \neq j \qquad (10)$$

where C_k is the k-th leaf cluster. Only the phones belonging to the same cluster are considered. Table 4 shows the results for conventional monolithic MLP,CHMLP 4/2 and CHMLP 4/2 (BU) respectively in columns (1)(2)(3). Clearly CHMLP (BU) can bring some extra benefit for classification of the confusing phones within each cluster.

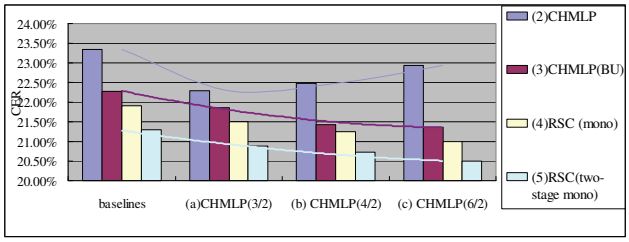

Fig. 4 Comparison of CER in Table 1 and Table 2 for columns labeled (2)(3)(4)(5)

FER	(1) Mono. MLP	(2) CHMLP 4/2	(3) CHMLP(4/2)(BU)
phone set[(2)-1]	15.42%	11.40%	9.57%
phone set[(2)-2]	11.32%	7.74%	5.55%
phone set[(2)-3]	14.36%	10.85%	8.90%
phone set[(2)-4]	17.46%	14.14%	11.41%

Table 4. Frame error rate (FER) within leaf cluster for monolithic MLP,CHMLP and CHMLP(BU)

4. CONCLUSION

This paper presents a bottom-up (BU) processing approach to improve the previously proposed clustered hierarchical MLP (CHMLP) tandem system. Improved accuracy can be obtained with more leaf MLPs with careful rescoring. The best result is obtained by CHMLP (BU) (6/2) rescored by two-stage MLP, yielded up to a 19.1% relative error reduction compared to the MFCC baseline.

5. REFERENCES

[1] Hermansky, H., Ellis, D.P.W. and Sharma, S., "Tandem connectionist feature extraction for conventional HMM systems", ICASSP 2000

[2] Bourlard, H. and Morgan, N., "Connectionist speech recognition: A hybrid approach", Kluwer Academic Publishers,Boston, USA, 1994

[3] Fosler-Lussier,E. and Morris,J. "Crandem systems: Conditional random field acoustic models for hidden macrov models", ICASSP 2008

[4] Siniscalchi,S., Schwarz,P and Chin-Hui,L. "High-accuracy phone recognition by combining high-performance lattice generation and knowledge-based rescoring", ICASSP 2007

[5] Hermansky, H., Ellis, "TRAP-TANDEM: Data-driven extraction of temporal features from speech", ASRU 2003

[6] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Incorporating Tandem/HATs MLP features into SRI's conversational speech recognition system," EARS RT-04F Workshop, 2004

[7] Hermansky H. and Fousek P., "Multi-resolution rasta filtering for tandem-based ASR.," Interspeech 2005.

[8] Valente,F. and Hermansky,H. "Hierarchical and parallel processing of modulation spectrum for ASR application", ICASSP 2008

[9] Zhu, Q., Stolcke, A., Chen, B., and Morgan, N, "Improved MLP structures for data-driven feature extraction for ASR", Interspeech 2005

[10] Schwarz, P., Matejka, P., Cernock J., "Hierarchical structures of neural networks for phoneme recognition", ICASSP 2006

[11] Antoniou, C., "Modular neural networks exploit large acoustic context through broad-class posteriors for continuous speech recognition", ICASSP 2001

[12] Sivadas, S. and Hermansky, H., "Hierarchical tandem feature extraction", ICASSP 2002

[13] Ketabdar, H and Bourlard, H. "Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation" ICASSP 2008

[14] S-Y. Chang and L-S Lee. "Data-driven clustered hierarchical tandem system for LVCSR", Interspeech 2008

[15] Frantisek,G and Fousek, P. "Optimizing Bottle-Neck features for LVCSR" ICASSP 2008