MULTI-STREAM SPECTRO-TEMPORAL AND CEPSTRAL FEATURES BASED ON DATA-DRIVEN HIERARCHICAL PHONEME CLUSTERS

Shang-wen Li, Liang-che Sun and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
b94901123@ntu.edu.tw, lgsun@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

We propose a method to enhance multi-stream Gabor and MFCC features using data-driven hierarchical phoneme clusters to yield more discriminating posteriors. We take into account different hierarchy structures, and in addition perform mean and variance normalization. A relative improvement of 11.5% over the conventional MFCC Tandem system was achieved in experiments conducted on Mandarin broadcast news. We analyze the complementarity between Gabor and MFCC features for different types of phonemes, and investigate the benefits that come from using hierarchical phoneme clusters.

Index Terms— spectro-temporal features, LVCSR, clustered hierarchical MLP

1. INTRODUCTION

In recent years, much work has been put into improving the performance of automatic speech recognition (ASR) by including longer context and temporal modulation information in features as compared to the conventional Mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) features. The MFCC Tandem system was successful in its use of a much longer context window to capture the temporal modulation information of speech signals with the aid of artificial neural networks (ANN) [1]. Modified ANN structures, such as hierarchical or parallel multilayer perceptrons (MLP) and MLPs with two or three hidden layers, were shown to provide even better performance [2, 3, 4, 5, 6, 7].

However, spectro-temporal modulation also plays an important role in speech signals. Intonation, coarticulation, and transition across phonemes naturally produce sloped patterns on the 2-D spectrogram. This is supported by recent findings in physiological experiments which showed that a large percentage of neurons in the primary auditory cortex of mammal species respond to signals with different spectro-temporal modulations [8]. These findings lead to substantial efforts in parameterizing those behaviors: autoregressive models and frequency domain linear prediction were used for extracting spectro-temporal features [9, 10], and approaches using independent component analysis and non-negative sparse coding were also proposed [11]. These approaches all resulted in significant improvements.

Filtering the log mel-spectrogram with 2-D Gabor filters is another way to extract the spectro-temporal behavior of signals [12, 13, 14]. Kleinschmidt and Gelbart employed a feature-finding neural network (FFNN) to iteratively adjust the parameters of Gabor filters for different recognition data [12]. In contrast, Zhao and Morgan selected Gabor filter parameters to divide the temporal modulation frequency from 1 to 16 Hz and the spectral modulation frequency from zero to two cycles per octave equally on a logarithmic scale [13], which was found to most closely correspond to human speech recognition [15]. They further divided Gabor filters into several streams, each of which covers a subset of Gabor filters within a specific spectro-temporal modulation frequency band. This multi-stream approach can be viewed as an ensemble of several recognition systems that yields better performance than each individual system [16]. Recently, this multi-stream approach was extended to obtain even more powerful spectro-temporal features and further improvement with increased number of feature dimensions [17, 18, 19].

Gabor features extracted following Zhao and Morgan's approach [13] have also been integrated with MFCCs at the phonetic posterior level, with the posteriors obtained using ANNs [20]. These combinatory posteriors were shown to yield improvements over the individual feature sets in phoneme accuracy; the authors also analyzed the complementarity between the two posteriors. A data-driven clustered hierarchical MLP (CHMLP) was also proposed to obtain more discriminating posteriors [7]. Here, we incorporate the concept of CHMLP into the multi-stream, or integrated, spectro-temporal and cepstral features and show improved performance in large vocabulary continuous speech recognition (LVCSR). We also investigate the benefits of hierarchical phoneme clusters, or CHMLP, when recognizing different types of phonemes. We further perform mean and variance normalization (MVN) on log-posteriors after dimension reduction, and demonstrate even larger performance gains.

2. PROPOSED APPROACH

Here we summarize MFCC and Gabor feature extraction, the clustered hierarchical posteriors from CHMLP, and how to put them together, in Sec. 2.1, 2.2, and 2.3, respectively.

2.1. Input for MLP

MFCC is one of the input feature sets for MLP. MFCCs are obtained with the typical parameter settings, 39-dimension MFCC features from c0 to c12 plus derivatives obtained with a 25ms window and a 10ms shift. Each vector is then concatenated with its previous and following four feature vectors as the MFCC input for MLP. This is denoted as **M** in the following.

We use 2-D Gabor filters to extract the spectro-temporal modulation information. The impulse response of Gabor filter *G(t,f)* is

$$G(t,f)=\frac{1}{2\pi\sigma_f\sigma_t}\times\exp\left[\frac{-(f-f_0)^2}{2\sigma_f^2}+\frac{-(t-t_0)^2}{2\sigma_t^2}\right] \times\exp[i\omega_f(f-f_0)+i\omega_t(t-t_0)], \quad (1)$$

where σ_t, σ_f, ω_t, and ω_f are the four parameters that must be chosen. Per previous work [13], we selected four sets of parameters of

Gabor filters, each corresponding to one Gabor filter bank, to cover from low to high spectro-temporal modulation bands. We convolve the log mel-spectrogram with the ith Gabor filter bank (i = 1 to 4 is the index for Gabor filter set). Used to train an individual MLP, the respective output is denoted as $\mathbf{G}_i$.

2.2. Data-driven clustered hierarchical MLP

The purpose of clustered hierarchical MLP (CHMLP) is to obtain more discriminating posteriors for recognition [7]. We first construct a data-driven phonetic hierarchy using the hierarchical agglomerative clustering (HAC) algorithm to cluster easily-confused phonemes, based on the probability of misclassification for every phoneme pair as evaluated using a monolithic MLP (the targets of which are the entire phoneme set). In each leaf node of the hierarchy, an MLP is specially trained for a cluster of easily-confused phonemes; in each higher-layer node, an MLP is used for classifying the input feature vector among its children nodes.

Shown in Fig. 1.1d is a simplified example of CHMLP with only two layers, one root MLP, and several specialized MLPs respectively for HAC-induced subsets of phonemes C_1, C_2, …, C_n (or C_k with k as the phoneme subset index). In each CHMLP stream with input feature stream **X**, each specialized MLP is fed the same feature set **X**; root MLP, though, is fed set **X'**, which may be different from **X**. Both **X** and **X'** can be either feature set **M** (MFCC) or one of $\mathbf{G}_i$ (Gabor). In Fig. 1, each specialized MLP is denoted as MLP(**X**,C_k), where the first parameter **X** indicates the input feature and the second parameter C_k the output target; the root MLP is denoted as root MLP(**X'**), where its parameter indicates the input feature. We multiply the output posterior vectors of the specialized MLPs with the posterior probability for the corresponding clusters obtained from the root MLP, and concatenate the resulting vectors as the hierarchical version posteriors for use in the Tandem system.

A hierarchy with more than two layers can be imagined as a soft decision tree with the root MLP as a parent node in layer m and specialized MLPs as its children in layer $m + 1$. The output posteriors of the parent node make the soft decision among its children, while a child MLP decides among its grandchildren (layer $m + 2$). Only leaf nodes decide among a phoneme subset. While different number of layers and hierarchies can be trained for different tasks, in our work and in a previous work [7], two layers were found to be optimal.

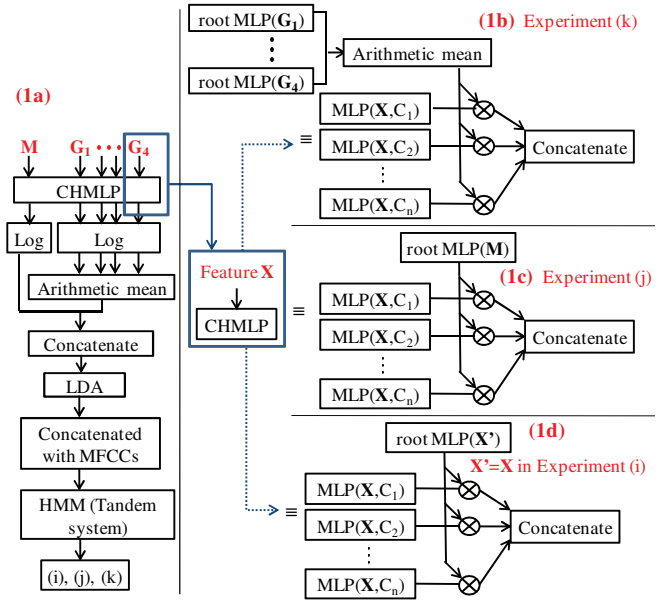

Fig. 1. (1a) Clustered hierarchical Tandem system with multi-stream posteriors in experiments (i), (j), and (k), which are the three different implementations of each stream of the top CHMLP block in Fig 1.1a.

2.3. Tandem system with clustered hierarchical combinatory posteriors

In Fig. 1.1a, we show the proposed Tandem system. Experiments (i), (j), and (k) represent three different ways of obtaining multi-stream, or integrated, clustered hierarchical posteriors from the features **M** and $\mathbf{G}_i$ and utilizing them in the Tandem system. The only difference between the three experiments is in the computation of the root MLP's output posterior (i.e. the estimation of the leaf cluster probability) in the top CHMLP block. For experiment (i), any one of the five streams (i.e., **M,** and $\mathbf{G}_1$ to $\mathbf{G}_4$) of the top CHMLP block are the same as in Fig. 1.1d, and the input of root MLP **X'** is the same as the input of the other specialized MLPs. As shown in Fig. 1.1c, for experiment (j), we use **M** as **X'** for all five streams. In Fig. 1.1b, we show that for experiment (k), in each stream, we first train four root MLPs with $\mathbf{G}_1$ to $\mathbf{G}_4$, respectively, and then multiply the arithmetic mean of their outputs with the output of the corresponding specialized MLPs. The process following the CHMLP block for experiments (i), (j), and (k) is illustrated in Fig. 1.1a: the log-arithmetic mean of the four CHMLP outputs with $\mathbf{G}_i$ inputs is concatenated with the logarithm of the outputs of the CHMLP with **M** input, linear discriminant analysis (LDA) is performed for dimension reduction and decorrelation, and the output is then concatenated with the MFCCs for the Tandem system.

We also designed several baselines, all of them Tandem systems, each with a different process for obtaining the posteriors. In experiments (a) and (b), the posteriors are obtained from a monolithic MLP individually with either **M** or $\mathbf{G}_i$ as input. Therefore, they are simply Tandem systems baselines with either MFCC or Gabor posteriors [13]. In experiment (d), posteriors from monolithic MLPs with input **M** and $\mathbf{G}_i$ are integrated to yield better posteriors for the Tandem system [20]. In experiment (c), our CHMLP baseline [7], we use CHMLP posteriors obtained from **M** only. In the above four experiments, we used principal component analysis (PCA) for dimension reduction and decorrelation. Replacing PCA with linear discriminant analysis (LDA) results in the four experiments (e) to (h).

We took into account different hierarchical structures for CHMLP, concluding that for this work's task the highest performance was achieved with three phoneme subsets for experiments (c) and (g) and two phoneme subsets for experiments (i), (j), and (k), all with two layers. We report the results for these settings below.

In experiment (l), the posteriors in experiment (k) are enhanced by MVN after performing LDA. The normalized output is then concatenated with the MFCCs and used in the Tandem system.

3. EXPERIMENTS

3.1. Experiment setup

The experiments were conducted on the MATBN (Mandarin Across Taiwan-Broadcast News) corpus. The training set includes

Table 1. CER for experiments described in Sec. 2.3

TANDEM system	CER (%)	Relative error rate reduction (%) compared to	
		(a)	(b)
(a) MFCC, PCA	22.6	—	×
(b) Gabor, PCA	22.3	1.3	—
(c) MFCC, CHMLP, PCA	22.1	2.2	0.9
(d) Integration, PCA	21.6	4.4	3.1
(e) MFCC, LDA	22.2	1.8	0.4
(f) Gabor, LDA	22.2	1.8	0.4
(g) MFCC, CHMLP, LDA	21.7	4.0	2.7
(h) Integration, LDA	21.1	6.6	5.4
Integration, CHMLP, LDA			
(i) Root MLP(**X**)	21.1	6.6	5.4
(j) Root MLP(**M**)	20.7	8.4	7.2
(k) Arithmetic mean	20.6	8.8	7.6
(l) Experiment (k) + MVN	20.0	11.5	10.3

13 hours of gender-balanced broadcast news collected in Taiwan from November 2001 to December 2002. A one-hour set of broadcast news collected in 2003 was used for testing. Two sets of acoustic units were used: a 36 monophone set as the MLP training target, and a total of 147 intra-syllable right-context-dependent Initial-Final (RCDIF) units as the recognition target for HMM training. The Initial is the initial consonant of a syllable, while the Final is the vowel (or diphthong) part which includes an optional medial or nasal ending. RCDIF is widely used for Mandarin speech recognition. There is a mapping table between the two sets of acoustic units; the phoneme-level training labels were based on forced alignment results.

Every MLP was trained with feature-dependent input nodes, 1000 hidden nodes, and label-dependent (either the entire monophone set or a cluster-induced subset) output nodes. Each RCDIF unit was modeled as an HMM of three states. A trigram language model was used in decoding.

3.2. LVCSR result

Table 1 lists the character error rate (CER) for experiments (a) to (l), described in Sec. 2.3, all for a Tandem system but with different processes for obtaining the posteriors. For PCA and LDA, we retained 95% of the total variance. The CER is 24.6% for MFCC system (i.e. no Tandem).

Rows (a) (b), (e), and (f) are the baselines using a single posterior type: (a) is the conventional MFCC Tandem system and (b) the Gabor Tandem system [13]. Replacing PCA by LDA yields baselines (e) and (f). The integration of the Gabor and MFCC posteriors, the subject of previous work [20], yielded statistically significant improvements at the 95% confidence level in experiments (d) and (h) as compared to the corresponding baselines for individual feature sets: (d) vs (a) (b) and (h) vs (e) (f). This shows that integration yields improvements not only for the phoneme recognition task [20] but also for LVCSR. CHMLP baselines (c) and (g) used hierarchical MFCC posteriors as features and outperformed (a) and (e). As presented above, we incorporated the integrated Gabor and MFCC posteriors with CHMLP in experiments (i) to (k), yielded additional improvements over the multi-stream Gabor and MFCC posteriors in experiment (h), and clustered hierarchical MFCC posteriors in experiment (g). We report only the results using LDA, as LDA consistently outperformed PCA in experiments (e) to (h). The CHMLP experiments (j) and (k) yielded statistically significant improvements at the 95% confidence level over experiment (h); experiment (i) did not yield statistically significant improvements (The 95% confidence interval is 21.45% to 20.75%). This is because the individual stream of Gabor filters used contained only part of the spectro-temporal modulation band and not enough for estimating posterior probabilities of clusters. Thus, we obtain better cluster estimates by taking into account the entire frequency band when averaging the output posteriors of the root MLPs over the four different inputs $\mathbf{G}_i$ as in experiment (k), or by feeding the root MLP with the MFCCs as in experiment (j).

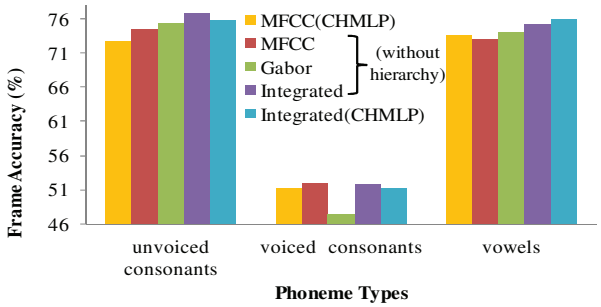

Fig. 2. Frame accuracy for different types of phonemes, with 36 monophones as classification targets.

The MVN added in experiment (l) yielded the best result, an 11.5% relative improvement over the conventional MFCC Tandem system (a) and a 10.3% improvement over the Gabor Tandem system (b).

3.3. Improvement analysis

3.3.1. Comparison between MFCC, Gabor, integrated, and hierarchical posteriors

We first investigated the performance of different posteriors including three sets obtained from MLP (MFCC, Gabor, and their integration) and two sets from CHMLP (MFCC and the integration of MFCC and Gabor), on different types of phonemes in terms of the frame accuracy achieved. The phoneme set was divided into three categories: unvoiced consonants, voiced consonants, and vowels, the frame proportions of which are 31%, 6%, and 63%, respectively. We classified the speech signals frame-by-frame to 36 monophone units based on a maximum a posteriori (MAP) criterion. The frame accuracy was averaged over each type of unit and is shown in Fig. 2.

In each category in Fig. 2, the three middle bars respectively correspond to MFCC and Gabor posteriors, and their integration, all without hierarchy. We see that this integration (4th bar, purple) was always superior or comparable to the other two results (2nd and 3rd, red and green); this is consistent with the results in Table 1. Comparing the 2nd and 3rd bars, we see that Gabor outperforms MFCC in vowels and unvoiced consonants while MFCC outperforms it for voiced consonants. The former is presumably due to the greater number of Gabor filter streams which retain more frequency component information. In addition, it may be that the spectro-temporal information extracted by Gabor filters more accurately reflects the formant transitions between vowels for diphthongs, which are abundant in Mandarin, as well as the transitions from silence to voiced parts in plosives. On the other hand, voiced consonants may be easily influenced by neighboring phonemes because they frequently follow vowels as the short nasal ending of Finals. The longer time span of Gabor filters may also include too much information from neighboring phonemes and thus overlook voiced consonants.

The results for hierarchical posteriors are the first (orange) and the last (blue) bars in each category in Fig. 2. Thus the first two bars are for MFCC posteriors alone, with and without the hierarchical structure, and the last two bars are for the integrated Gabor and MFCC posteriors, similarly with and without the hierarchical structure. In both cases, CHMLP (hierarchical posteriors) yielded higher accuracy for vowels, with degraded accuracy for voiced and unvoiced consonants. In the hierarchies constructed for this task, unvoiced consonants are clustered into one or two subsets in two- or three-leaf CHMLP, with the remaining group for vowels and voiced consonants. However, the vowel part is usually much longer than the consonant part; this frame sample imbalance between different leaves works against the unvoiced-consonant leaves for the root MLP and leads to degradation for unvoiced consonants. Correspondingly, this favors the leaf for vowels and voiced consonants, and thus leads to improved vowel recognition. The more specialization on vowel recognition of leaf MLP is another reason for the improvement. The data imbalance problem leads also to degraded recognition for voiced consonants, which are overlooked in leaf MLP.

Although CHMLP did not consistently outperform in all categories, the improved vowel classification is more important in decoding and thus contributes more to character accuracy. This is partially because of the longer time span for vowels, which greatly affects the frame-by-frame accumulation of acoustic model scores during Viterbi decoding. Similar discovery was found previously [21]. This explains the improved LVCSR performance with CHMLP.

In summary, the integration of MFCC and Gabor posteriors yields consistent improvements, and the clustered hierarchical approach for integrated posteriors yields even better performance for vowels, which dominate character accuracy, at the expense of accuracy for other categories. The best character accuracy was achieved using hierarchical integrated posteriors.

3.3.2. Mismatch between training and testing data

Although there is no noise in the corpus used, speaker variation still results in a mismatch between training and testing data. The histograms of the values of a sample component (the second) of the after-LDA log-posteriors obtained in experiment (l) before and after MVN are shown in the right and left panels of Fig 3, with the blue curve for training data and the red for testing. We see that the training and testing data distributions in the left panel are more consistent and symmetric around zero, and are thus better modeled by Gaussians. This observation holds for all components. We also computed the Bhattacharyya distance between the probability distributions for the posterior vectors for the training and testing data. We found that the distance was reduced by 27% after performing MVN. Thus we see why MVN yielded further improvements in experiment (l).

4. CONCLUSION

We showed that the integration of Gabor and MFCC posteriors in Tandem systems yielded relative improvements of 6.6% over an MFCC Tandem and 5.4% over a Gabor Tandem [13] for LVCSR, and proposed a hierarchical structure for this integration to obtain more discriminative posteriors and a further improvement of 8.8% over a MFCC Tandem. Adding an MVN step then yielded the best result, an 11.5% improvement. The hierarchical structure puts more emphasis on vowel recognition, which dominates acoustic model scores during decoding at the expense of other phoneme types. MVN compensates for the mismatch between training and testing data and results in distributions that are better modeled as Gaussians.

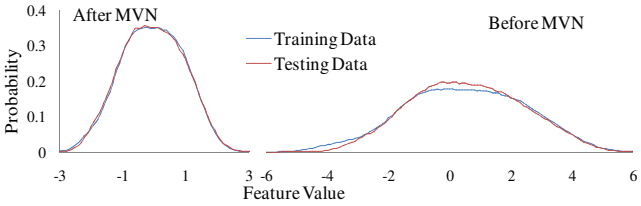

Fig. 3. The histograms of the values of a sample component of the after-LDA log-posteriors from training and testing data before and after MVN.

5. REFERENCES

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

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

[3] Zhu, Q., Chen, B., Grezl, F. and Morgan, N., "Improved MLP structures for data-driven feature extraction for ASR", in *Proc. Interspeech*, 2005

[4] Schwarz, P., Matejka, P. and Cernock, J., "Hierarchical structures of neural networks for phoneme recognition", in *Proc. ICASSP*, 2006

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

[6] Grezl, F. and Fousek, P., "Optimizing Bottle-Neck features for LVCSR" in *Proc. ICASSP*, 2008.

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

[8] Depireux, D.A., Simon, J.Z., Klein, D.J. and Shamma, S.A., "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex", *J. Neurophysiology*, vol. 85, pp. 1220–1234, 2001.

[9] Thomas, S., Ganapathy, S. and Hermansky, H., "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction", IEEE Sig. Proc. Let., Vol. 15, pp. 681-684

[10] Ganapathy, S., Thomas, S. and Hermansky, H., "Robust spectro-temporal features based on autoregressive models of Hilbert envelopes", in *Proc. ICASSP*, 2010.

[11] Domont, X., Heckmann, M., Joublin, F. and Goerick, C., "Hierarchical spectro-temporal features for robust speech recognition", in *Proc. ICASSP*, 2008.

[12] Kleinschmidt, M. and Gelbart, D., "Improving word accuracy with Gabor feature extraction", in *Proc. ICSLP*, 2002.

[13] Zhao, S. and Morgan, N., "Multi-stream spectro-temporal features for robust speech recognition", in *Proc. Interspeech*, 2008.

[14] Meyer, B. and Kollmeier, B., "Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities", in *Proc. Interspeech*, 2009.

[15] Chi, T., Gao, Y., Guyton, M.C., Ru, P. and Shamma, S.A., "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, 106:2719–2732, 1999.

[16] Gelbart, D., *Ensemble feature selection for multi-stream automatic speech recognition*, Ph.D. dissertation, University of California, Berkeley, http://www.icsi.berkeley.edu, 2008.

[17] Zhao, S., Ravuri, S. and Morgan, N., "Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR", in *Proc. ICASSP*, 2009.

[18] Thomas, S., Mesgarani, N. and Hermansky, H., "A Multistream Multiresolution Framework for Phoneme Recognition", in *Proc. Interspeech*, 2010.

[19] Ravuri, S. and Morgan, N., "Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR", in *Proc. Interspeech*, 2010.

[20] S.-W. Li, L.-C. Sun and L.-S. Lee., "Improved phoneme recognition by integrating evidence from spectro-temporal and cepstral features", in *Proc. Interspeech*, 2010

[21] Stilp, C. and Kluender, K., "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility", in *Proc. National Academy of Sciences of the USA*, 2010