Improved Phoneme Recognition by Integrating Evidence from Spectro-temporal and Cepstral Features

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

Gabor features have been proposed for extracting spectro-temporal modulation information, and yielding significant improvements in recognition performance. In this paper, we propose the integration of Gabor posteriors with MFCC posteriors, yielding a relative improvement of 14.3% over an MFCC Tandem system. We analyze for different types of acoustic units the complementarity between Gabor features with long-term spectro-temporal modulation information in the mel-spectrogram and MFCC features with short-term temporal information in the cepstral domain. It is found that Gabor features are better for vowel recognition while MFCCs are better for consonants. This explains why their integration offers improvements.

Index Terms: spectro-temporal modulation, Gabor features

1. Introduction

Over the past few decades, mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) have been commonly used as features for speech recognition. However, both MFCCs and PLP constitute only very local information due to the short window length (typically 25 ms) used when extracting these two features. Adding delta and acceleration terms is useful, but fails to capture the rich information present in successive observations. The successful MFCC Tandem systems [1] use artificial neural networks (ANN) to capture the temporal behavior of speech signals using a much longer context window, yielding significant performance improvements.

In the MFCC paradigm, the mel-filter bank is followed by a discrete cosine transform (DCT), which extracts spectral modulation information from the signal, thereby compressing the high-dimension spectrogram to the rather low dimension cepstral domain. Beginning in the mid-1990s, Hermansky and Morgan proposed extracting temporal trajectory information with RelAtive SpecTrA (RASTA) features [2], which involve the use of filtering to estimate modulation of signals in a longer time interval over the critical band spectrogram, and multi-resolution RASTA [3], which involves further analysis of the temporal modulation using a bank of band-pass filters with varying resolutions. The usefulness of these features derives in large part from the temporal modulation information.

However, it was also found that spectro-temporal modulation plays an important role in speech signals. Intonation, coarticulation, and transition across phonemes naturally produce sloped patterns on the two dimensional spectrogram. This is supported by recent findings in physiological experiments which show that a large percentage of neurons in the primary auditory cortex of mammal species respond to signals with different spectro-temporal modulations [4, 5]. These findings lead to substantial efforts to use 2-D Gabor filters to extract the spectro-temporal behavior of signals in the mel-spectrogram [6, 7, 8]. For instance, Kleinschmidt and Gelbart employed a feature-finding neural network (FFNN) to iteratively adjust the parameters of Gabor filters for different recognition data using a greedy algorithm [6]. In contrast, Zhao and Morgan selected Gabor filters 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 [7], which was found to closely correspond to human speech recognition [9]. 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 stream [10]. More powerful Gabor filters have also been proposed recently [11].

We follow Zhao and Morgan in extracting spectro-temporal information using several sets of 2-D Gabor filters (here denoted as Gabor features) [7], obtaining posteriors of Gabor features from ANN, and integrating them with MFCC posteriors. We have found that these two set of posteriors are complementary, one with longer spectro-temporal modulation information in the mel-spectrogram and the other with shorter temporal correlation information in the cepstral domain; thus they are able to better classify phonemes of different types. This explains why the integration of the two sets of posteriors, Gabor and MFCC, works well for speech recognition. While the results here were obtained for a Mandarin Chinese task, we believe the techniques described here are language-independent.

2. Proposed approach

Here we summarize how the two sets of posteriors from the Gabor and MFCC features are obtained (Fig. 1) and how they are integrated (Fig. 2).

2.1. MFCC posteriors

This set of posteriors is obtained in the conventional way, as shown on the right side of Fig. 1. MFCCs are first calculated using the HTK [12] supplied front-end with the typical parameter settings. The 39-dimension MFCC feature vector includes c0 to c12 as well as the delta and acceleration terms, and is calculated over a 25ms window with a 10ms window shift. Each vector is then concatenated with its previous and following four feature vectors and fed into a multi-layer tron (MLP). The MLP output is a vector of posterior probabilities, each of which corresponds to a specific phoneme. For better modeling of Gaussian mixtures in the Tandem system here, we took the logarithm of the probabilities as our MFCC posteriors.

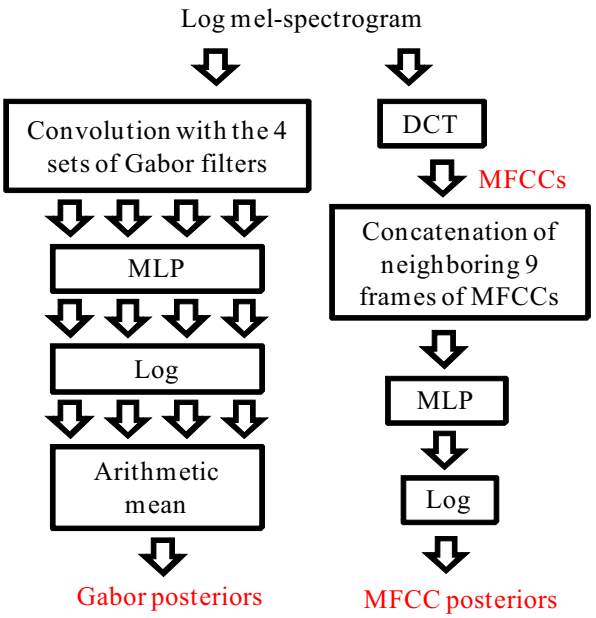

Figure 1: *The generation of the two sets of posteriors.*

2.2. Gabor posteriors

We use 2-D Gabor filters to extract the spectro-temporal modulation information. The impulse response of Gabor filter *G(t,f)* is the product of a Gaussian envelope *g(t,f)* and a modulation term *m(t,f)*

$$G(t,f) = g(t,f) \times m(t,f), \tag{1}$$

where
$$g(t,f) = \frac{1}{2\pi\sigma_f\sigma_t} \times \exp\left[\frac{-(f-f_0)^2}{2\sigma_f^2} + \frac{-(t-t_0)^2}{2\sigma_t^2}\right] \tag{2}$$

and
$$m(t,f) = \exp[i\omega_f(f-f_0) + i\omega_t(t-t_0)], \tag{3}$$

where σ_t, σ_f, ω_t, and ω_f are the four parameters to be chosen. σ_t and σ_f are set to π/ω_t and π/ω_f for constant modulation cycles in a Gaussian window. ω_t (temporal modulation) and ω_f (spectral modulation) of Gabor filters were selected using knowledge-based approaches [7]. These parameters are listed in the upper half section of Table 1. We also set either ω_t or ω_f to 0 to produce spectral- or temporal-only modulation filters. Because of the 0 in the denominator, σ_t or σ_f is additionally chosen, as shown in the lower half of Table 1. Table 1 also shows the division of these Gabor filters into four sets, from low to high spectro-temporal modulation bands.

As shown in the left stream of Fig. 1, after pre-emphasizing the speech signal, taking the FFT, passing through the 23 mel-filter banks and taking the logarithm, the resulting signal, the log mel-spectrogram, is convolved with each Gabor filter in the four sets. The output of these filters is collected into four streams of Gabor features for different bands. Each stream of these Gabor features is used to train an individual MLP respectively. Our final Gabor posteriors are then the arithmetic mean of the four sets of logarithmic posteriors from each stream at the frame level.

2.3. Tandem systems with different sets of posteriors and their integration

Fig. 2 shows all of the experiments performed in this work, including the proposed integration and baselines. In the lower long bar in Fig. 2, all of the posteriors, regardless of how they are obtained, are concatenated with the MFCCs and used in the HMM phoneme recognizer. Hence all of the results here are for Tandem systems, and differ only with respect to the input posteriors. While the rightmost experiment stream (b) uses only the MFCC posteriors, after principal component analysis (PCA) for decorrelation, as the input, the two leftmost experiment streams (c) and (g) use the Gabor posteriors, after PCA or linear discriminant analysis (LDA) for decorrelation. These three experiments serve as the baselines here.

In the three middle streams of Fig. 2, the two sets of posteriors are integrated using three different approaches followed by PCA in experiments (d) (e) (f) or by LDA in experiments (h) (i) (j). Here the MFCC and Gabor posteriors are integrated at the frame level in different ways and result in a stronger set of posteriors containing temporal, spectral, and spectro-temporal modulation information in the mel-spectrogram and cepstral domains. The resulting posteriors are also used in the same Tandem system. In experiments (d) and (h), we use the arithmetic mean (AM) for these two sets of posteriors. The inverse entropy (IE) [13] is used instead in

Spectro-temporal filters	ω_t (Hz)	ω_f (rads/mel-channel)
Set1	±2	3.14, 2.26, 1.51, 0.82, 0.25
	±4	0.25
Set2	±4	3.14, 2.26, 1.51, 0.82
	±7	0.82, 0.25
Set3	±7	3.14, 2.26, 1.51
	±11	1.51, 0.82, 0.25
Set4	±11	3.14, 2.26
	±16	3.14, 2.26, 1.51, 0.82, 0.25
Temporal filters	ω_t (Hz)	σ_f (mel-channel)
Set1	3.5	1, 1.39, 2.08, 3.85, 12.5
Set2	7.5	
Set3	11.5	
Set4	15	
Spectral filters	σ_t (seconds)	ω_f (rads/mel-channel)
Set1	0.25, 0.13, 0.07, 0.05, 0.03	0.09
Set2		0.21
Set3		0.33
Set4		0.45

Table 1: *Parameters for the four Gabor filter sets.*

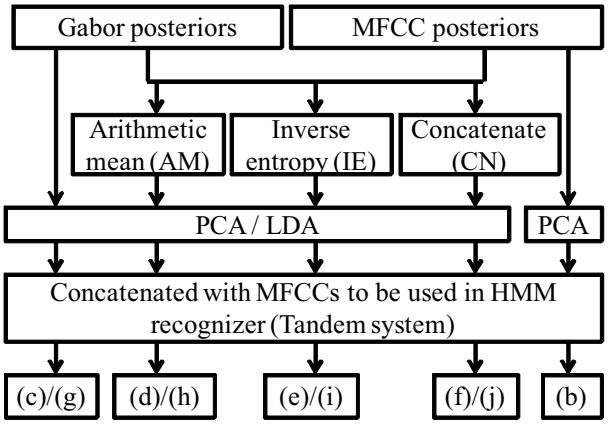

Figure 2: *Different ways of using posteriors in experiments (b) to (j).*

experiments (e) and (i). We also directly concatenate (CN) the two sets of posteriors in experiments (f) and (j). The integration methods AM, IE, and CN are respectively followed by PCA or LDA as mentioned above for decorrelation and dimension reduction, resulting in two experiments for each integration method.

3. Experiments

3.1. Experimental setup

The experiments were conducted on the Mandarin read speech corpus TCC300 [14]. It was recorded in an ordinary office environment via close-talking microphones at a 16 kHz sampling rate. Part of the corpus is phonetically balanced sentences produced by 50 males and 50 females; from this part we used 4596 utterances as the training set and 511 utterances for testing.

As our phone set we used right context dependent Initial-Finals, for a total of 148 intra-syllable right context dependent Initial-Final units as the recognition target [15]. The Initial is the initial consonant of a syllable, while the Final is the vowel (or diphthong) part and includes the optional medial and nasal ending. Each Initial-Final unit was modeled as a hidden Markov model (HMM) of three states. The phone-level training label was based on forced alignment. A lexicon and language model was not used.

3.2. Experimental results

Table 2 lists the results of phone recognition for the experiments shown in Fig. 2. The first row (a) of Table 2 is for the MFCC baseline, the only result not for a Tandem system. Rows (b) to (j) are for experiments (b) to (j) from Fig. 2 and described in Sec. 2.3, all for a Tandem system but with different posteriors used as input features. For the PCA and LDA shown in Fig. 2, we retained 95% of the total variance.

Rows (b), (c), and (g) were our baselines: (b) is the conventional MFCC Tandem system, and (c) and (g) are the PCA and LDA Gabor Tandem systems, respectively. Row (c) is the exact system proposed by Zhao and Morgan [7]. For PCA systems, the integration of the Gabor and MFCC posteriors yielded improvements in experiments (d) to (f) when arithmetic mean (AM), inverse entropy (IE), and concatenation (CN) were used, respectively, as compared to the individual posteriors in the baseline experiments (b) and (c). As shown in rows (h) to (j) of Table 2, integrating the two sets of posteriors with the same three integration methods followed by LDA yielded further improvements as compared to row (g) with only the Gabor posteriors, all with LDA. Clearly, integrating these two sets of posteriors considerably outperformed using any single set of posteriors. The best scheme proposed here is that for experiment (j), which yields a 14.3% relative improvement over the conventional MFCC Tandem system (b) and 9.5% over the Gabor Tandem system (c).

3.3. Complementarity analysis for MFCC and Gabor posteriors

We further investigated the complementarity of these two sets of posteriors to find out why the integration yielded better results. In this experiment, we investigated the frame accuracy achieved by each set of posteriors: MFCC, Gabor, and their integration by arithmetic mean (AM). The phone set was divided into three categories: unvoiced consonants, voiced consonants, and Finals. Unvoiced consonants included three types: plosive, fricative and affricate; voiced consonants three types: nasal, liquid and semivowel; Finals two types: vowel and diphthong; for a total of eight types. Frame by frame, we classified the speech signals to 148 Initial-Final units based on a maximum a posteriori (MAP) criterion. The frame accuracy for each type of unit as well as for the entire phone set is shown in Fig. 3. In each group in Fig. 3, the three bars correspond to MFCC and Gabor posteriors and their integration. Note that while the better of the MFCC and Gabor posteriors depends on the phone type, their integration always outperforms either type of posteriors individually. This is clear evidence of their complementarity.

ASR system	Phoneme acc. (%)	Relative error rate reduction (%) compared to	
		(b)	(c)
(a)MFCC	66.2	×	×
(b)MFCC, PCA	75.5	—	×
(c)Gabor, PCA	76.8	5.3	—
(d)Integration, AM, PCA	77.8	9.4	4.3
(e)Integration, IE, PCA	77.4	7.8	2.6
(f)Integration, CN, PCA	77.5	8.2	3.0
(g)Gabor, LDA	77.8	9.4	4.3
(h)Integration, AM, LDA	78.6	12.7	7.8
(i)Integration, IE, LDA	78.1	10.6	5.6
(j)Integration, CN, LDA	79.0	14.3	9.5

Table 2: *Phoneme accuracy for experiments shown in Fig. 2.*

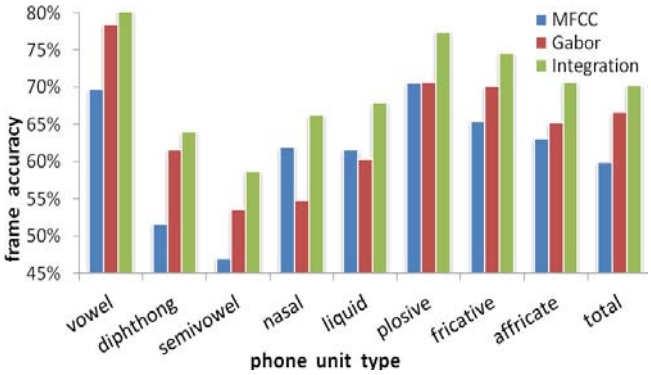

Figure 3: *Frame accuracy for different types of phone units. The classification targets were 148 Initial-Final units. For each type the three bars are for MFCC and Gabor posteriors and their integration.*

In the first three groups of Fig. 3 (vowel, diphthong, and semivowel), Gabor posteriors significantly outperformed MFCC posteriors; this is consistent with previous results [8]. For these vowel and vowel-like units, the spectrogram structure patterns are quite stable, and are therefore best represented as Gabor posteriors. Moreover, the spectrum transitions in diphthongs lead to clear spectro-temporal modulation components in the spectrogram which are also well reflected in Gabor posteriors. These two facts result in the superiority of Gabor posteriors over the MFCC alternative.

For the next two groups in Fig. 3 (nasal and liquid), while MFCC posteriors outperformed Gabor posteriors, the integration of the two was still better than either taken individually. This shows that even if Gabor posteriors are inferior to MFCC posteriors for some unit types, they do include complementary information due to difference in nature.

For unvoiced consonants (plosive, fricatives, and affricates), Gabor posteriors were on par with or slightly outperformed MFCC posteriors. This is somewhat contrary to earlier

findings [8], probably due to the different dimensions of Gabor features in the two studies. Frequency components of unvoiced consonants are spread out in the spectrogram and sometimes change with time. Perhaps more Gabor filters covering different spectral and temporal modulation frequencies are needed to extract enough information for unvoiced consonant classification. We performed an additional experiment to investigate the frame accuracy by classifying these unvoiced consonants using the four individual streams of posteriors shown in Table 1 extracted from each Gabor filter set (before the arithmetic mean-based merging to Gabor posteriors in Fig. 1). We found that the performance of all of the individual Gabor posterior streams was consistently and significantly lower (the frame accuracy of the four streams respectively was 45.6%, 62.3%, 62.6%, and 56.4% for plosive, 53.7%, 63.2%, 59.4%, and 50.3% for fricative, and 48.1%, 58.3%, 56.3%, and 47.1% for affricate) than the merging of the four (i.e., Gabor posteriors: the red bar in Fig. 3, the frame accuracy was 70.7% for plosive, 70.1% for fricative, and 65.2% for affricate). This indicates that the information for unvoiced consonants is spread out in every frequency band: therefore some information may be lost if fewer Gabor filters are used. Here, MFCC posteriors yielded performance comparable to that for Gabor posteriors, presumably because the DCT thereof retains information from every frequency band. This is also supported by the fact that MFCC posteriors (the blue bar in Fig 3, the frame accuracy was 70.6% for plosive, 65.4% for fricative, and 63.0% for affricate) actually outperformed all of the individual Gabor streams (before the merging to Gabor posteriors), each of which contains information in only part of the frequency range.

We find in Fig. 3 that the integration of MFCC and Gabor posteriors improved the accuracy consistently for all phone types and overall (overall frame accuracy was 60% and 67% for MFCC and Gabor posteriors, and 70% for their integration). In other words, regardless of the performance differences between these two types of posteriors, they complement each other; properly exploiting this complementarity by integrating the two yields clear performance benefits.

We also investigated a rougher classification task for different sets of posteriors. In Fig. 4, the classification target was limited to three categories (i.e., voiced and unvoiced consonants and Finals). Thus only the correct category out of the three – as opposed to the correct phone – was needed for correct classification. Here MFCC posteriors outperformed Gabor posteriors for voiced and unvoiced consonants but were tied for Finals. This seems to imply that MFCC posteriors are more useful for category classification, while Gabor posteriors discriminate well different phone units belonging to the same category: this is another aspect of the complementarity of the two types of posteriors.

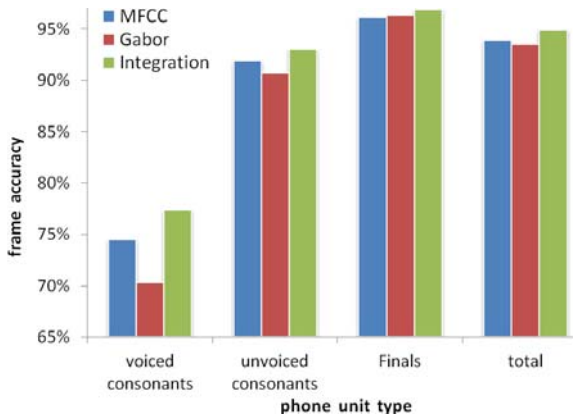

Figure 4: *Frame accuracy when the classification targets were only the three categories: voiced, unvoiced consonants, and Finals.*

4. Conclusions

We integrated Gabor and MFCC posteriors in Tandem systems, yielding significant performance improvements: 14.3% over a MFCC Tandem and 9.5% over a Gabor Tandem system [7]. The complementarity between the two sets of posteriors was analyzed using frame accuracy results. We discovered that the respective contributions of MFCC and Gabor posteriors for recognition depend on the phone unit type. This complementarity explains why the integration of the two posteriors yields significant performance improvements. This is clearly because the two sets of posteriors are quite different in nature.

5. References

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

[2] Hermansky, H. and Morgan, N., "RASTA Processing of Speech", *IEEE Trans. Speech and Audio Proc.*, 2(4):578–589, 1994.

[3] Hermansky, H. and Fousek, P., "Multi-resolution RASTA filtering for TANDEM-based ASR", in *Proc. Interspeech*, 2005.

[4] 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.

[5] Klein, D.J., Depireux, D.A., Simon, J.Z. and Shamma, S.A., "Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design", *J. Comp. Neuroscience*, 9:85–111, 2000.

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

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

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

[9] 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.

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

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

[12] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ol-lason, D., Povey, D., Valtchev, V. and Woodland, P., *The HTK book (for HTK version 3.2)*, Cambridge University, Eng. Dept., 2002, Technical report.

[13] Misra, H., Bourlard, H. and Tyagi, V., "New entropy based combination rules in HMM/ANN multi-stream ASR", in *Proc. ICASSP*, 2003.

[14] The Association for Computational Linguistics and Chinese Language Processing, `http://www.aclclp.org.tw/` .

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