An Experimental Analysis on Integrating Multi-Stream Spectro-Temporal, Cepstral and Pitch Information for Mandarin Speech Recognition

Yow-Bang Wang, Shang-Wen Li, and Lin-shan Lee, *Fellow, IEEE*

***Abstract*—Gabor features have been proposed for extracting spectro-temporal modulation information from speech signals, and have been shown to yield large improvements in recognition accuracy. We use a flexible Tandem system framework that integrates multi-stream information including Gabor, MFCC, and pitch features in various ways, by modeling either or both of the tone and phoneme variations in Mandarin speech recognition. We use either phonemes or tonal phonemes (tonemes) as either the target classes of MLP posterior estimation and/or the acoustic units of HMM recognition. The experiments yield a comprehensive analysis on the contributions to recognition accuracy made by either of the feature sets. We discuss their complementarities in tone, phoneme, and toneme classification. We show that Gabor features are better for recognition of vowels and unvoiced consonants, while MFCCs are better for voiced consonants. Also, Gabor features are capable of capturing changes in signals across time and frequency bands caused by Mandarin tone patterns, while pitch features further offer extra tonal information. This explains why the integration of Gabor, MFCC, and pitch features offers such significant improvements.**

***Index Terms*—Pitch, spectro-temporal features, tandem system, toneme.**

I. Introduction

OVER the past few decades, Mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) features have been commonly used as features for speech recognition. However, both MFCCs and PLPs consider only very local information due to the short window length (typically 25 ms) used when extracting these features. In an MFCC framework, the mel-filter bank is followed by a discrete cosine transform (DCT), which extracts spectral modulation information from the signal, thereby compressing the high-dimensional spectrogram to the relatively low-dimensional cepstral domain. Although useful, adding delta and acceleration terms still does not capture well the rich information present in successive observations.

In recent years, much work has been put into improving the performance of speech recognition by including longer context and temporal modulation information in features; this is different from conventional MFCC and PLP features. Beginning in the mid-1990s, Hermansky and Morgan proposed extracting temporal trajectory information with RelAtive SpecTrA (RASTA) features [1], which estimate the modulation of signals in a longer time interval over the critical band spectrogram using temporal filters, as well as multi-resolution RASTA [2], which further analyzed the temporal modulation using a bank of band-pass filters with varying resolutions. The power of these features came largely from the temporal modulation information. In addition, the MFCC Tandem system was also very successful in utilizing a much longer context window to capture the temporal modulation information of speech signals with the aid of artificial neural networks (ANN) [3]. Further modified ANN structures, such as hierarchical or parallel multilayer perceptrons (MLP), and MLPs with two or three hidden layers, were shown to yield even better performance [4]–[9].

It has also been found that spectro-temporal modulation plays an important role in speech signals. Intonation, coarticulation, and transitions 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 [10]. These findings led to substantial efforts in parameterizing those behaviors: autoregressive models and frequency-domain linear prediction have been used to extract spectro-temporal features [11], [12], and approaches using independent component analysis and non-negative sparse coding have also been proposed [13]. These approaches have all resulted in large improvements.

Parameterizing the log mel-spectrogram with 2-D Gabor filters is another way to extract the spectro-temporal behavior of signals [14]–[16]. Kleinschmidt and Gelbart employ a feature-finding neural network (FFNN) to iteratively adjust the parameters of Gabor filters for different recognition data [14]. In contrast, Zhao and Morgan select Gabor filter parameters by dividing 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 [15]; this has been found to most closely correspond to human speech perception [17]. They divide Gabor filters into several streams,

Manuscript received December 18, 2012; revised April 19, 2013; accepted May 03, 2013. Date of publication June 07, 2013; date of current version July 22, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steve Renals.

Y.-B. Wang is with the Graduate Institute of Electrical Engineering, National Taiwan University, Taipei City 10617, Taiwan (e-mail: d98921028@ntu.edu.tw).

S.-W. Li is with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139 USA (e-mail: swli@mit.edu).

L. Lee is with the Department of Electrical Engineering, National Taiwan University, Taipei, 10617 Taiwan (e-mail: lslee@gate.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2263803

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 [18]. Recently, this multi-stream approach has been extended to use even more powerful spectro-temporal features and further improve the performance with an increased number of feature dimensions [19]–[21].

Pitch is another type of information in the speech signal that MFCC/PLP features are not to capture. The frequency of the pitch of human speech roughly falls between 50 Hz to 500 Hz. The window shift between frames for extracting MFCC/PLP is not short enough to resolve pitch information temporally; also, because the bandwidth of the filters applied in frequency domain for extracting MFCC/PLP is generally wider than pitch, pitch information is also smoothed away spectrally. Pitch may not be a critical feature for languages which are not tonal (English, for instance). Mandarin Chinese, however, is a tonal language, in which every syllable is assigned a tone; the tone carries lexical meaning, which helps distinguish ambiguous words. For decades, pitch information has been known to be useful in Mandarin speech recognition [22]–[24]. The works that take into account pitch and tones can be roughly divided into explicit and embedded tone modeling [25], as summarized below. In explicit tone modeling, the prosodic and acoustic evidence is assumed independent given word sequences. The likelihood scores based on one set of these features are weighted-summed with the other, usually done either during decoding or by rescoring the word lattices or N-best lists obtained in first-pass ASR [26], [27]. However, the independence assumption of prosodic and acoustic evidences could be over-simplified, and rich information between them could be completely lost. In contrast, in embedded tone modeling we append tone-related features to spectral features, and model the tonal acoustic units within existing recognition frameworks. Embedded tone modeling as such may overlook long-term prosody information in the speech signal. To alleviate such oversight, some works have applied the Tandem system with a long context window [28]–[31]; others have incorporated embedded tone modeling with explicit tone modeling and showed that these two types of approaches are complementary [32]–[36].

Recently, we parameterized the spectro-temporal modulation information following Zhao and Morgan's 2-D Gabor filter approach [15], and integrated it with MFCCs at the phonetic posterior level using simple or hierarchical MLPs [37], [38]. We found that these two sets 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—and yield better discriminability over different types of phonemes. We further integrated them with pitch features in Mandarin speech recognition, using tonal phonemes (or tonemes) for MLP posterior estimation and tonal acoustic units for HMM recognition in Tandem system. This yielded improved performance [39].

In this paper, we use a flexible Tandem system framework to perform a more complete experimental analysis over different feature integration configurations for Mandarin speech recognition. We examine different combinations of MLPs based on different assumptions of dependency among feature sets and classification targets. We also analyze the complementarity among different sets of features for different classification targets. The rest of this paper is organized as follows. The feature extraction procedure and Tandem system framework are respectively explained in Sections II and III. Detailed experimental analysis is then presented in Section IV. We conclude in the last section.

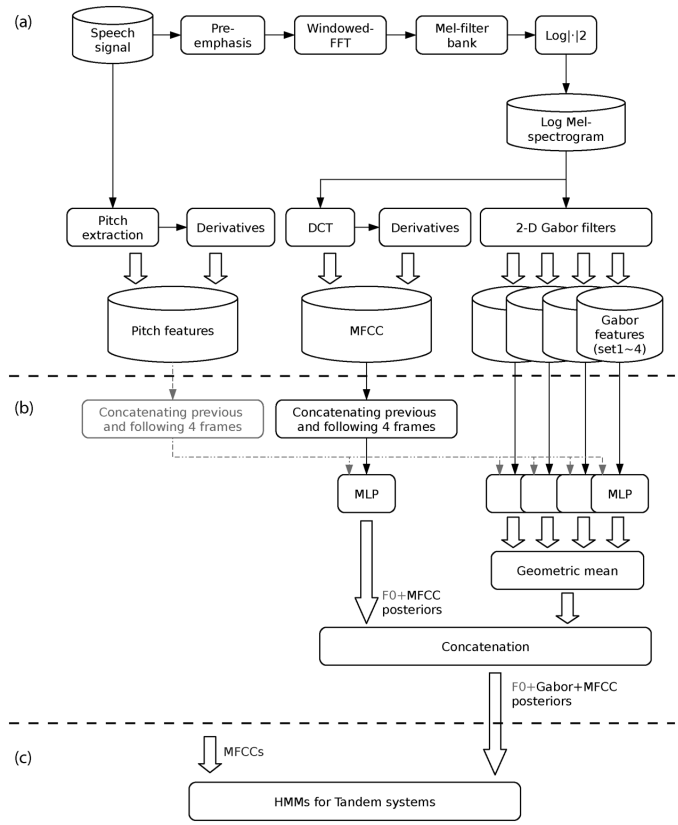


Fig. 1. Flow chart of (a) feature extraction, (b) posterior derivation and (c) the Tandem system. The gray dotted parts are optional in the experiments. Results with and without pitch information are both reported.

II. Feature Extraction

Fig. 1(a) illustrates how we extract the multi-stream spectro-temporal, cepstral, and pitch features from the speech signal. After pre-emphasizing the speech signal, taking windowed-FFT, passing through the mel-filter bank and taking the logarithm, the log Mel-spectrogram is obtained. The window length and shift are 25 ms and 10 ms respectively in windowed-FFT. Both multi-stream spectro-temporal and cepstral features are extracted from the log Mel-spectrogram; pitch features are extracted directly from time-domain signal. Below we briefly present the different sets of features used in this work and how they are extracted.

A. Cepstral Features

For cepstral features we use 39-dimensional MFCCs, including c0 to c12 plus derivatives and accelerations. This 39-dimensional MFCC feature set is used both for HMM acoustic modeling and also to derive the posteriorgrams (a

TABLE I
PARAMETER VALUES FOR FOUR GABOR FILTER SETS

Spectro-temporal filters		
	ω_t(Hz)	ω_f(rad/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	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(second)	ω_f(rad/mel-channel)
Set1	0.25, 0.13, 0.07, 0.05, 0.03	0.09
Set2		0.21
Set3		0.33
Set4		0.45

vector of posterior probabilities) with MLP; this is discussed further in Section III.

B. Spectro-Temporal Features

As shown in the right stream of Fig. 1(a), the log mel-spectrogram is convolved with 2-D Gabor filters to extract spectro-temporal modulation information, or Gabor features. The impulse response of a Gabor filter $G(t, f)$ is the product of a Gaussian envelope $g(t, f)$ and a modulation term $m(t, f)$:

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

where

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

$$m(t, f) = exp\left[i\omega_f(f - f_0) + i\omega_t(t - t_0)\right], \tag{2b}$$

and σ_t, σ_f, ω_t, and ω_f are the four parameters that determine the shape of each filter. The parameter values follow Zhao and Morgan [15] and are listed in Table I. In the upper half of Table I, σ_t and σ_f are set to π/ω_t and π/ω_f for constant modulation cycles in a Gaussian window, while ω_t (temporal modulation) and ω_f (spectral modulation) of the Gabor filters are selected based on human knowledge. These parameter sets 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; this was found to most closely correspond to human speech perception [17]. In the lower half of Table I we also set either ω_t or ω_f to zero to produce spectral-or temporal-only modulation filters. Because of the zero in the denominator, σ_t or σ_f is additionally chosen.

As also shown in Table I and Fig. 1(a), these Gabor filters are divided into four streams (Set1 to Set4), each corresponding to one Gabor filter bank, from low to high spectro-temporal modulation frequency bands [15].

C. Pitch Features

Pitch information is extracted using the Snack Sound Toolkit [40]. Since pitch is present only in voiced segments, the Snack toolkit assigns zero pitch to unvoiced frames. In unvoiced regions the pitch contour can be interpolated into non-zero values to make it smoother and to avoid the variance problem in recognition [33]. More precisely, the log-F0 contour is refined using cubic spline interpolation and then normalized by mean subtraction per utterance. For pitch features, we use pitch and its first and second derivatives. The frame length and shift of the Snack toolkit and of MFCC/Gabor feature extraction are synchronized so that these feature sets can be simultaneously generated for each frame. In the following sections, the pitch feature set is denoted as "F0."

III. TANDEM SYSTEMS

Below we use a flexible Tandem system framework to determine the recognition performance achievable by the multi-stream features mentioned above. In HMM training and recognition, we utilize a set of posteriors derived from different sets of features, different MLPs, and some post-processing.

A. System Architecture

In the Tandem system, we estimate the posteriors for Mandarin phonemes or tonemes (tonal phonemes) using the three sets of features illustrated in Section II and Fig. 1(a), with different configuration of MLPs. The MLP classification target can be either monophone (mono-phoneme) or mono-toneme. The output of each MLP is a vector of posteriors, with each element corresponding to the posterior probability of a specific monophone or mono-toneme given the input features of the present time frame.

As shown in the right stream of Fig. 1(b), each of the four streams of Gabor features obtained frame-by-frame is fed individually into an MLP, with pitch features (including those obtained from the previous and following four frames) augmented or not. We then merge the four streams of posteriors by taking the geometric mean over the four vectors frame-by-frame. We denote the resulting vectors as F0 + Gabor posteriors if pitch features are augmented. We similarly augment the MFCC features, including those of the previous and following four frames, with pitch features, and feed them into an MLP. The resulting posteriorgrams are denoted as F0 + MFCC posteriors. In the case of no pitch feature augmentation, the posteriors obtained are simply referred to as Gabor or MFCC posteriors. Note that for the MLP input, each MFCC and F0 feature vector of a frame is concatenated with its previous and following four frames; the Gabor feature vectors, however, are not.

To utilize the complementarity between Gabor and MFCC features investigated previously [38], [39], we further merge the F0 + Gabor (or Gabor alone) and F0 + MFCC (or MFCC alone) posteriors by concatenation, shown at the bottom of Fig. 1(b), and denote the results as F0 + Gabor + MFCC posteriors (or simply Gabor+MFCC posteriors if pitch features are not included).

In the Tandem systems, the posteriors are first transformed by a logarithm function which maps the range of the value between zero and one to a negative real number. We then use linear

discriminant analysis (LDA) transformation for dimension reduction and decorrelation. We retain 95% of the total variance. Both steps are employed for better Gaussian modeling in the following HMMs. We further perform mean and variance normalization (MVN) on the after-LDA log-posteriors just as done in previous studies [21], [38]. Although the noise level is relatively low in the corpus used here, the MVN step helps to reduce the mismatch between training and testing data due to speaker variation, and also results in distributions better modeled as Gaussians.

The normalized posteriors are then concatenated with MFCC frame-by-frame, and used in HMM training and recognition in the Tandem system as shown in Fig. 1(c). In this way, we have Tandem systems that differ in the posteriors obtained with different input feature sets. For analysis purposes, the target units for MLP posterior estimation and HMM recognition can be either tonal or not, as we will explain below.

B. Different MLP Combinations for Toneme Posterior Derivation

In prosodic modeling for Mandarin speech recognition, the following condition is often assumed for simplicity:

$$Pr(P,T|X,F) \approx Pr(P|X,F)Pr(T|X,F) \quad (3a)$$

$$\approx Pr(P|X)Pr(T|F), \quad (3b)$$

where P and T represents phoneme and tone sequences and X and F acoustic and prosodic feature sequences of an utterance, respectively. The first approximation in (3a) assumes the phoneme and tone sequences are independent given acoustic and prosodic feature sequences, and the second approximation in (3b) assumes phoneme sequences depend only on acoustic feature sequences, while tone sequences depend only on prosodic feature sequences. Similar assumptions may also apply to features of frames. If we let P and T represent the phoneme and tone classes and X for MFCC and Gabor features and F for pitch features, all for a frame, the terms in (3a) and (3b) $(Pr(P,T|X,F), Pr(P|X,F), Pr(T|X,F), Pr(P|X)$ and $Pr(T|F))$ are posteriors of phoneme and/or tone classes, given some feature sets of a frame as discussed above. We experimentally analyze results when (3a) and (3b) are applied to frames.

Fig. 2 illustrates three different approaches to implementing any of the MLPs in Fig. 1(b), either for MFCC or Gabor features, with or without pitch features, for deriving toneme posteriors. Fig. 2(a) shows a monolithic MLP used to calculate the posterior probability of all the tonemes, corresponding to finding $Pr(P,T|X,F)$ on the left hand side of (3a) by jointly modeling tone and phoneme. In contrast, Fig. 2(b) shows that we may use two MLPs, one for tone classification to obtain $Pr(T|X,F)$ and the other for phoneme classification to obtain $Pr(P|X,F)$, and combine the results by simply multiplying them together as $Pr(P,T|X,F)$, as in (3a). One may question that the toneme posteriors estimated as Fig. 2(a) could be dominated by the acoustic features due to their much higher dimensionality than pitch features, and makes the comparison between different assumptions of independence less convincing.

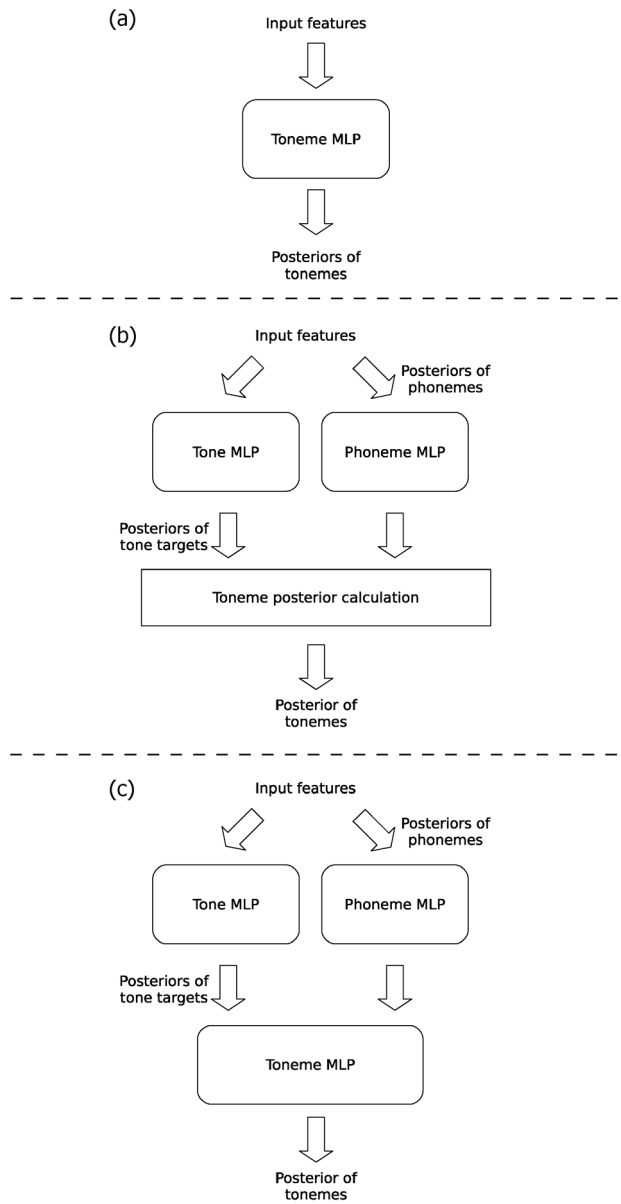


Fig. 2. Using different combinations of MLPs to derive toneme posteriors.

For more complete examination, we also tested using the hierarchical MLPs as shown in Fig. 2(c), where we concatenate the output of the phoneme MLP and tone MLP to train yet another MLP for toneme classification. In this way the acoustic and prosodic posteriorgram features have closer dimensionality before being concatenated and fed into the toneme MLP, as a comparison to Fig. 2(a).

Note that (3a) is only an assumption of approximation for the purpose of reducing the number of model parameters given limited training data. If sufficient training data is available, the posteriors of tonemes can certainly be better inferred using a monolithic MLP by considering the dependency between phonemes and tones given acoustic and pitch features. Yet given insufficient data, the resulting monolithic MLP could suffer from overfitting. In that case using two compact MLPs may be a good compromise, assuming some conditions of independence.

IV. Experiments

In this section, we describe the experimental setup, results, and analysis.

A. Experiment Setup

The experiments were conducted on the MATBN (Mandarin Across Taiwan-Broadcast News) corpus, available via the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) [41]. The training set included 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.

We started with 36 Mandarin phonemes (monophones), 23 consonants and 13 vowels, and expanded each vowel to its tonal variants while leaving the consonant part unchanged (we assumed the consonant parts do not carry tones). We thus obtained a set of 75 tonemes as the MLP training target. As will be reported below, the 36 phonemes were also used as the MLP training target for comparison. Note that in Mandarin there are four lexical tones plus a neutral tone, but here only the four lexical tones were included in the toneme set; the neutral tone was mapped to tone-3. Because there were a very small number of syllables produced as the neutral tone, adding it to the toneme set caused a data imbalance problem [30], [42], [43]. Therefore the "Tone MLP" in Fig. 2(b) was trained to classify feature vectors into five targets: consonant, tone-1, tone-2, tone-3, and tone-4; the "Phoneme MLP" was trained to classify 36 phonemes including 23 consonants and 13 vowels. In Fig. 2(a), the "Toneme MLP" was trained to classify feature vectors into 75 tonemes.

We essentially used the intra-syllable right-context-dependent Initial-Final (RCDIF) as the unit of acoustic models (HMMs) in the Tandem systems; we also tested triphones for some baseline systems for comparison. RCDIF has been widely used as the unit for Mandarin speech recognition, especially for tasks with limited quantities of training data. The Initial is the onset consonant of a syllable, and the Final is the vowel (or diphthong) part plus an optional medial or nasal ending. Because syllables in Mandarin Chinese have Initial-Final structure, only the Initials need be right-context-dependent to encode intra-syllable dependency; all Finals can be context-independent. This yields good recognition accuracy even with limited quantities of training data. Of course tri-phones (or even quin-phones) offer better performance, but this comes at the cost of much higher computation/corpus requirements. Since the purpose of this work is to analyze the contribution of different sets of features in Mandarin speeh recognition, rather than to achieve the highest possible accuracy, as done in most studies [30], RCDIFs instead of tri-phones or quin-phones were used for a relatively small corpus. It is believed that the analysis obtained here can be equally extended to acoustic models based on units that take into account more context dependency.

Each RCDIF unit (tonal or not) was modeled as an HMM of three states. 147 RCDIF units (112 right-context-dependent Initials plus 35 contect-independent Finals) were expanded to 257 tonal RCDIFs in the same way monophones were expanded to tonemes, including the original tone-independent Initials and tonal Finals, as the recognition target for HMM training. In the experiments below, the 147 RCDIF toneless units were also used as HMM units for comparison. The RCDIF-level training labels were based on forced alignment results.

We used a word-based trigram language model trained with the newswire text corpus, consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2001 and 2002 [44], and processed with Katz smoothing [45] using the SRI Language Modeling Toolkit (SRILM) [46]. Our 73K-word lexicon was also generated from this text corpus [47]. To further generate a tonal lexicon, we used the tones of each word provided in the Chinese Electronic Dictionary from the ACLCLP [48], and added the tone labels into our original lexicon.

B. Experimental Results (I): Recognition Error Rate

We list in Table II the character error rate (CER) for each recognition experiment for different configurations. Row (a) and (e) are the conventional HMM system using the triphones and RCDIFs as the HMM units respectively, with MFCC features and serves as the non-Tandem baseline. The other rows are all for Tandem systems. The first column lists the feature sets used for estimating the posteriors, and the second and third columns show respectively the MLP target (36 phonemes or 75 tonemes) and HMM units for recognition (147 RCDIFs or 257 tonal RCDIFs) in each Tandem system, as explained in detail in Section III.

In rows (b)–(d) and (f)–(h), the MLP targets are the 36 phonemes without respect to tone, and the HMM units are respectively the triphones and RCDIFs, both without tones. Row (b) and (f) are the conventional MFCC Tandem system and row (c) and (g) are the Gabor Tandem system [15]. The Gabor Tandem system in row (c) outperforms MFCC in row (b) with triphones as HMM units, while the two Tandem systems in rows (f)(g) with RCDIFs as HMM units achieved comparable results; and integrating MFCC and Gabor features by concatenation at the posterior level (row (d) and (h) as illustrated in the bottom of Fig. 1(f)) enhanced the performance for both triphone and RCDIF systems. This is clearly because of the complementarity of the cepstral and spectro-temporal features as noted in previous work [16], [38]; for both MLP classification and HMM recognition, although the performance of the individual MFCC and Gabor posteriors depends on the type of phone, their integration always outperforms either one individually. Since in these baseline systems using the RCDIFs as HMM units (rows (e)–(h)) outperforms triphones (rows (a)–(d)) by roughly 7% in general, we continue the experiments using RCDIFs as HMM units.

Replacing HMM units with tonal RCDIFs in rows (i)–(k) yielded consistent improvements among the three sets of posteriors; rows (i)(j) yielded statistically significant improvements over rows (f)(g); row (k) achieved statistically significant improvements over rows (f)–(h). Here the confidence level of the significance test was set to 95%. This shows that using tonal RCDIFs as HMM units allowed the HMMs to learn the variation of MFCCs and posteriors across different tones and improved the performance. Note this also implies Gabor and MFCC features may possess tonal information; otherwise better results would not have been achieved if the Tandem system could not distinguish vowels with different tones via Gabor and MFCC.

More evidence is presented of this inference in the following analysis.

In rows (l)–(n), we replaced the MLP targets with tonemes. Gabor toneme posteriors, row (m), achieved statistically significantly better performance than the phoneme counterpart, row (j); but the MFCC toneme posteriors, row (l), did not, with respect to row (i). A combination of the two posteriors, row (n), yielded statistically significant improvement over row (k). We analyze the results as follows:

First, because MLP training with phoneme targets as in rows (i)–(k) was mainly aimed at discriminating between phonemes, some tonal clues were inevitably lost in the posteriors. Changing MLP targets from phonemes to tonemes thus helped to learn more tonal information at the front-end (rows (m) and (n) vs. (j) and (k)). Also, comparing rows (m) and (n) to (j) and (k), we see more improvement (2.1% to 1.8%) than for rows (j) and (k) to (g) and (h) (1.0% to 0.8%). Because rows (j)(k) and rows (g)(h) used posteriors that do not take into account tones, the tonal HMM units are limited in their ability to discriminate tones.

Next, although the improvement from row (f) to (i) verified that MFCCs carry tonal information, from row (i) to (l) MFCC toneme posteriors offered little additional benefit. On the other hand, obviously Gabor toneme posteriors retained more information for modeling the tonal targets in HMMs and yielded improvements from row (j) to (m). Furthermore, the benefits from changing recognition targets into tonal ones for MLPs and HMMs are additive (from rows (g)(h) to (j)(k) to (m)(n)).

In rows (o)–(q), we report the results when the MLP input features are augmented with F0 as in the gray dotted parts in Fig. 1(f). We can see that adding pitch features boosted the performance consistently over rows (l)–(n). The improvements from rows (l) to (o), (m) to (p) and (n) to (q) were statistically significant. Because the pitch features were extracted directly from the signals, they clearly offered additional information to the MFCC and Gabor features used above, which were extracted from mel-spectrogram.

Finally in row (r) and (s), we changed all MLPs in Fig. 1 from monolithic MLP in Fig. 2(a) to the combination of two MLPs in Fig. 2(b) and the combination of three MLPs in Fig. 2(c) respectively, with all other settings unchanged as in row (q). The CER using two MLPs was further reduced to 16.2%, which was a statistically significant improvement. This was the best performance achieved with the present experimental setup, with 5% absolute or 23.6% relative CER improvement compared to the conventional Tandem system baseline in row (f). The improvement from row (q) to (r) was obviously due to the reduced number of parameters to be estimated in the expert MLPs as compared to the monolithic MLP with the relatively small corpus, and resulted in more reliable posteriors. On the other hand, the CER using three MLPs was only slightly reduced to 16.9%, which was not a statistically significant improvement over row (q). This further shows that estimating phoneme and tone posteriors separately could be better when limited training data is given.

C. Experimental Results (II): Frame-Level Classification by MLP

After showing the recognition results, we investigate the frame-level classification accuracies obtained with different MLP classifiers constructed with different input features and different classification targets. First, the frame accuracies of the four Mandarin tones and the total are illustrated in Fig. 3. To calculate these frame-level tone accuracies, for each frame we excluded the consonants and summed the posterior probabilities for all the rest of tonemes having the same tone but varied phoneme parts. Therefore for each frame only four probabilities were obtained for the four tones despite there being 75 tonemes. We then classified the tone for each frame following the maximum a posteriori (MAP) criterion on the tone posteriors.

The general relationships among bars (1)–(3) in Fig. 3 remained essentially unchanged in bars (4)–(6) for all four tones and the total. In bars (1)–(3) for each tone and the total, the Gabor features (bar(2)) always resulted in significantly better tone classification accuracy than MFCC (bar(1)). Because MFCC was much worse than Gabor, adding MFCC to Gabor (bar(3)) was worse than Gabor alone (bar(2)) in tones 1, 2, 3 and total. The only exception was for tone 4, in which the accuracy for both MFCC and Gabor were 70% or above. With additional pitch features (bars (4)–(6)), the tone accuracy was consistently enhanced in every tone and every feature set.

An interesting observation is that the classification accuracy using F0 + MFCC (bar(4)) is still inferior to using Gabor features only (bar(2)) for all the four tones and the total. Intuitively the Mandarin tones differ primarily in their pitch patterns, yet the Gabor features extracted as shown in Fig. 1 did not retain pitch information, because the window shift (10 ms) used was not short enough to resolve the pitch information temporally, while the bandwidth of the Mel-filters were wider than pitch so the pitch information was also smoothed away spectrally. This implies that different Mandarin tones result in not only different pitch patterns, but also different spectro-temporal patterns; i.e., the Mandarin tone is not just a purely prosodic phenomenon, but carries acoustic information as well. We may assume that when a speaker produces patterns with varying pitch, some parts of the vocal tract are actually influenced simultaneously, which means varying pitch patterns also results in varying acoustic feature patterns. Obviously the Gabor filters with different sloped patterns across the two-dimensional time and frequency bands extracted such acoustic feature patterns well. Therefore, approximating the frame-wise tone posterior $Pr(T|X,F)$ by $Pr(T|F)$ can be lossy, because it ignores the dependency from the acoustic features, as in the second term of (3b).

That also explains why in general Gabor features were better than MFCCs in classifying the tones. Because the two-dimensional Gabor filters with different sloped patterns over the spectrogram responded to different spectro-temporal modulation, the Gabor features were more sensitive than cepstral features to frequency component changes across the time and frequency band. Gabor features are thus capable of capturing more information of varying acoustic patterns needed for tone classification.

Of course, pitch features were also helpful for tone classification even with the presence of Gabor features which clearly parameterized considerable spectro-temporal modulation components. Pitch features always provide additional tonal information from the speech signal directly, and benefits the discrim-

TABLE II
CER FOR EACH RECOGNITION SYSTEM WITH VARIOUS FEATURES AND ACOUSTIC UNITS. 36 PHONEMES OR 75 TONEMES FOR MLP TARGET, AND 147 RCDIFS OR 257 TONAL RCDIFS AS HMM UNITS

Features for estimating the posteriors	MLP target	MLP layout	HMM unit	CER(%)
(a) MFCC(non-Tandem)	N/A	N/A	triphone	31.2
(b) MFCC	phoneme	monolithic	triphone	28.2
(c) Gabor	phoneme	monolithic	triphone	27.7
(d) Gabor+MFCC	phoneme	monolithic	triphone	27.2
(e) MFCC(non-Tandem)	N/A	N/A	RCDIF	24.6
(f) MFCC	phoneme	monolithic	RCDIF	21.2
(g) Gabor	phoneme	monolithic	RCDIF	21.3
(h) Gabor+MFCC	phoneme	monolithic	RCDIF	20.4
(i) MFCC	phoneme	monolithic	tonal RCDIF	20.3
(j) Gabor	phoneme	monolithic	tonal RCDIF	20.3
(k) Gabor+MFCC	phoneme	monolithic	tonal RCDIF	19.6
(l) MFCC	toneme	monolithic	tonal RCDIF	20.4
(m) Gabor	toneme	monolithic	tonal RCDIF	18.2
(n) Gabor+MFCC	toneme	monolithic	tonal RCDIF	17.8
(o) F_0+MFCC	toneme	monolithic	tonal RCDIF	17.8
(p) F_0+Gabor	toneme	monolithic	tonal RCDIF	17.5
(q) F_0+Gabor+MFCC	toneme	monolithic	tonal RCDIF	17.0
(r) F_0+Gabor+MFCC	toneme	2 MLPs	tonal RCDIF	16.2
(s) F_0+Gabor+MFCC	toneme	3 MLPs	tonal RCDIF	16.9

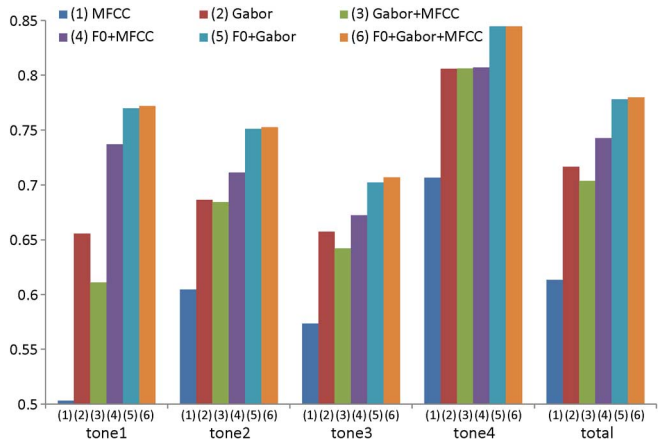


Fig. 3. The frame-level tone accuracy of different features used with MLP for tonemes.

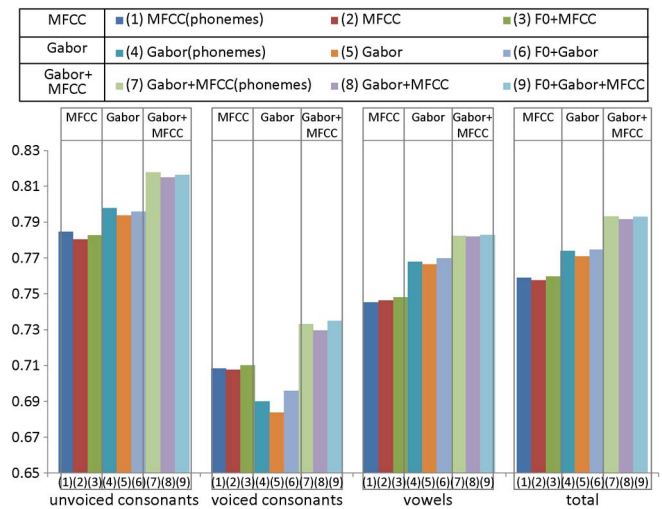


Fig. 4. Frame-level phoneme accuracy of different MLP targets and input features.

inability over the tones with MLPs. But since both Gabor and pitch features carry tonal information, their complementarity is not as extensive as MFCC and pitch features. In Fig. 3 we note that the improvement from bar (1) to (4) is more significant than from bar (2) to (5). In the recognition accuracy in Table II we note a similar situation: the improvement from row (l) to (o) is more obvious than that from row (m) to (p).

In Fig. 4 we show the frame-level accuracy for three different phoneme types (unvoiced consonants, voiced consonants, and vowels) and their total. To calculate the phoneme accuracy, similar to Fig. 3 but in a different way, for each frame we summed the posterior probabilities of all tonemes with the same phoneme but varied tones. Thus, 36 frame-wise phoneme posteriors were obtained for the 36 phonemes for MAP classification, although there were 75 tonemes. In each phoneme type, the first three bars (1)(2)(3) are for MFCC features, the next three (4)(5)(6) for Gabor, and the last three (7)(8)(9) for MFCC + Gabor. In each section of the three bars, the first bar is for MLP with 36 phonemes as targets directly, while the other two bars are for the 75 tonemes. The third bar in each section is with F0 features as input in addition.

We can see that for unvoiced consonants, vowels, and the total, Gabor posteriors (the middle three bars (4)–(6)) outperformed MFCC (the first three (1)–(3)), and their integration (the last three (7)–(9)) helped still more; for voiced consonants, MFCC (bars (1)–(3)) were superior to Gabor (bars (4)–(6)), and their integration (bars (7)–(9)) performed even better. Consistent with previous findings [38], [39], this shows the complementarity among Gabor and MFCC features in phoneme recognition (i.e., Gabor and MFCC features are respectively stronger for different types of phonemes), and explains why integrating them yields improvements. A similar situation can be found in the recognition results in Table II: for example, rows (b)(c) are comparable but row (d) is significantly better. However, different MLP targets (phoneme or toneme in bars (1)(2)) or pitch feature concatenation (bars (3)) resulted in very similar phoneme accuracies. This supports the assumption that the frame-wise posterior probability $Pr(P|X,F)$ may be approximated by $Pr(P|X)$ ignoring the pitch information in the first term of (3b). Although phoneme classification may

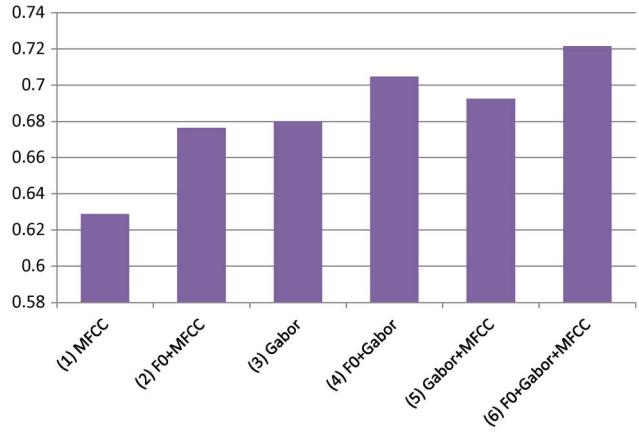


Fig. 5. Frame-level toneme accuracy of different features used with MLP for tonemes.

benefit less from the pitch features, using toneme posteriors as Tandem features still offered more information to the following HMM as discussed above.

Fig. 5 illustrates the frame-level toneme accuracy for different features, i.e., the accuracy of the combination of tone and phoneme was considered. Consistent with the trends we found in the above analysis, Gabor (bars (3)(4)) outperformed MFCC (bars (1)(2)), and their integration (bars (5)(6)) improved still more. Also, augmenting F0 features boosted the accuracy. Note that concatenating F0 features with Gabor (bars (4)) resulted in better toneme accuracy than concatenating MFCC with Gabor (bar (5)), although the dimensionality of the MFCC feature vector was much larger than that of the pitch feature vector (39*9 vs. 3*9). This again highlights the importance of pitch features in toneme recognition for Mandarin Chinese.

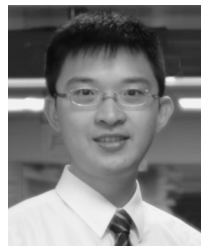
V. Conclusion

We offered a comprehensive analysis on spectro-temporal, cepstral, and pitch features, and their application in a Mandarin speech recognition task. We analyzed carefully the contributions made by each feature set to the recognition accuracy and the complementarity between different sets of features, based on the experimental results. We designed a Tandem system framework to flexibly integrate different sets of features extracted from speech signals, including cepstral, spectro-temporal, and pitch features, and also to model the tone and phoneme variation simultaneously (tonemes for MLP and tonal RCDIFs for HMM) in the experiments. We also investigated how tone and phoneme accuracies are influenced by these features. The results indicate that 2-D Gabor filters are capable of capturing vowel, unvoiced consonant, and tonal information in the spectral-temporal domain, and that both Gabor and pitch features are complementary to the conventional MFCC features in boosting the accuracy of Mandarin speech recognition.

References

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[2] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for tandem-based ASR," in *Proc. Interspeech*, 2005.
[3] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.
[4] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. ICASSP*, 2008, pp. 4165–4168.
[5] Q. Zhu, B. Chen, F. Grezl, and N. Morgan, "Improved MLP structures for data-driven feature extraction for ASR," in *Proc. Interspeech*, 2005.
[6] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, 2006, vol. 1, pp. I–I.
[7] H. Ketabdar and H. Bourlard, "Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation," in *Proc. ICASSP*, 2008, pp. 4065–4068.
[8] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. ICASSP*, 2008, pp. 4729–4732.
[9] S. Chang and L. Lee, "Data-driven clustered hierarchical tandem system for LVCSR," in *Proc. Interspeech*, 2008.
[10] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, no. 3, p. 1220, 2001.
[11] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Process. Lett.*, vol. 15, pp. 681–684, 2008.
[12] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of Hilbert envelopes," in *Proc. ICASSP*, 2010, pp. 4286–4289.
[13] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical spectro-temporal features for robust speech recognition," in *Proc. ICASSP*, 2008, pp. 4417–4420.
[14] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. ICSLP*, 2002, vol. 5, pp. 16–38.
[15] S. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Proc. Interspeech*, 2008.
[16] B. Meyer and B. Kollmeier, "Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities," in *Proc. Interspeech*, 2009.
[17] T. Chi, Y. Gao, M. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, pp. 2719–2732, 1999.
[18] D. Gelbart, "Ensemble feature selection for multi-stream automatic speech recognition," Ph.D. dissertation, Univ. of California, Berkeley, CA, USA, 2008.
[19] S. Zhao, S. Ravuri, and N. Morgan, "Multi-stream to many-stream: Using spectro-temporal features for ASR," in *Proc. Interspeech*, 2009.
[20] N. Mesgarani, S. Thomas, and H. Hermansky, "A multistream multiresolution framework for phoneme recognition," in *Proc. Interspeech*, 2010.
[21] S. Ravuri and N. Morgan, "Using spectro-temporal features to improve AFE feature extraction for ASR," in *Proc. Interspeech*, 2010.
[22] L. Lee, C. Tseng, H. Gu, F. Liu, C. Chang, Y. Lin, Y. Lee, S. Tu, S. Hsieh, and C. Chen, "Golden Mandarin (I)-a real-time Mandarin speech dictation machine for Chinese language with very large vocabulary," *IEEE Trans. Speech Audio Process.* , vol. 1, no. 2, pp. 158–179, Apr. 1993.
[23] H. Wang, T. Ho, R. Yang, J. Shen, B. Bai, J. Hong, W. Chen, T. Yu, and L. Lee, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 2, pp. 195–200, Mar. 1997.
[24] L. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal Process. Mag.*, vol. 14, no. 4, pp. 63–101, Jul. 1997.
[25] T. Lee, W. Lau, Y. Wong, and P. Ching, "Using tone information in Cantonese continuous speech recognition," *ACM Trans. Asian Lang. Inf. Process.*, vol. 1, no. 1, pp. 83–102, 2002.
[26] L. Cheng and L. Lee, "Improved large vocabulary Mandarin speech recognition by selectively using tone information with a two-stage prosodic model," in *Proc. Interspeech*, 2008.
[27] H. Wei, X. Wang, H. Wu, D. Luo, and X. Wu, "Exploiting prosodic and lexical features for tone modeling in a conditional random field framework," in *Proc. ICASSP*, 2008, pp. 4549–4552.
[28] X. Lei, M. Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *Proc. Interspeech*, 2005.
[29] M. Hwang, W. Wang, X. Lei, J. Zheng, O. Cetin, and G. Peng, "Advances in Mandarin broadcast speech recognition," in *Proc. Interspeech*, 2007.
[30] M. Hwang, G. Peng, W. Wang, A. Faria, A. Heidel, and M. Ostendorf, "Building a highly accurate Mandarin speech recognizer," in *Proc. ASRU*, 2007, pp. 490–495.

[31] F. Valente, M. Doss, C. Plahl, S. Ravuri, and W. Wang, "A comparative large scale study of MLP features for Mandarin ASR," in *Proc. Interspeech*, 2010.
[32] H. Wang, Y. Qian, F. Soong, J. Zhou, and J. Han, "A multi-space distribution (MSD) approach to speech recognition of tonal languages," in *Proc. Interspeech*, 2006.
[33] X. Lei, M. Siu, M. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. Interspeech*, 2006.
[34] H. Wang, Y. Qian, F. Soong, J. Zhou, and J. Han, "Improved Mandarin speech recognition by lattice rescoring with enhanced tone models," in *Proc. ISCSLP*, 2006, pp. 445–453.
[35] X. Lei and M. Ostendorf, "Word-level tone modeling for Mandarin speech recognition," in *Proc. ICASSP*, 2007, vol. 4, pp. IV-665–IV-668.
[36] X. Wang, Y. Yu, X. Wu, and H. Chi, "Maximum entropy based tone modeling for Mandarin speech recognition," in *Proc. ICASSP*, 2010.
[37] S. Li, L. Sun, and L. Lee, "Improved phoneme recognition by integrating evidence from spectro-temporal and cepstral features," in *Proc. Interspeech*, 2010.
[38] S. Li, L. Sun, and L. Lee, "Multi-stream spectro-temporal and cepstral features based on data-driven hierarchical phoneme clusters," in *Proc. ICASSP*, 2011, pp. 5196–5199.
[39] S. Li, Y. Wang, L. Sun, and L. Lee, "Improved tonal language speech recognition by integrating spectro-temporal evidence and pitch information with properly chosen tonal acoustic units," in *Proc. Interspeech*, 2011.
[40] The Snack Sound Toolkit. [Online]. Available: http://www.speech.kth.se/snack/
[41] MATBN (Mandarin Across Taiwan-Broadcast News) corpus. [Online]. Available: http://www.aclclp.org.tw/use_mat_c.php
[42] X. Liu, M. J. F. Gales, J. L. Hieronymus, and P. C. Woodland, "Investigation of acoustic units for LVCSR systems," in *Proc. IEEE ICASSP*, 2011, pp. 4872–4875.
[43] F. Diehl, M. Gales, X. Liu, M. Tomalin, and P. Woodland, "Word boundary modelling and full covariance Gaussians for Arabic speech-to-text systems," in *Proc. Interspeech*, 2011, pp. 777–780.
[44] Chinese Gigaword. [Online]. Available: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T09
[45] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 3, pp. 400–401, Mar. 1987.
[46] SRILM—the SRI Language Modeling Toolkit. [Online]. Available: http://www.speech.sri.com/projects/srilm/
[47] L. Chien, "PAT-tree-based keyword extraction for Chinese information retrieval," *ACM SIGIR Forum*, vol. 31, no. SI. ACM, pp. 50–58, 1997.
[48] Chinese Electronic Dictionary—The Association for Computational Linguistics and Chinese Language Processing. [Online]. Available: http://www.aclclp.org.tw/use_ced.php



Yow-Bang Wang was born in 1984. He received his B.S. degree from the Dept. of Electrical Engineering, National Taiwan University in 2006, and the M.S. degree from the Graduate Institute of Communication Engineering, National Taiwan University in 2008.

He is currently a Ph.D. student in the Speech Processing Laboratory, National Taiwan University. His research interests include Computer-Assisted Language Learning and Mandarin Chinese tone recognition.

Shang-Wen Li was born in 1987. He received his B.S. degree from the Dept. of Electrical Engineering, National Taiwan University in 2009, and the M.S. degree from the Graduate Institute of Communication Engineering, National Taiwan University in 2011. While at NTU, he worked on acoustic features and Tandem system for large vocabulary speech recognition in the Speech Processing Laboratory.

He is currently a graduate student in the Spoken Language Systems Group of the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT). He joined MIT in September of 2012. His research interests include speech and language processing.

Lin-shan Lee (F'93) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, Taiwan, since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world including text-to-speech systems, natural language analyzers, dictation systems, and voice information retrieval systems.

Dr. Lee was Vice President for International Affairs (1996–1997) and the Awards Committee chair (1998–1999) of the IEEE Communications Society. He was a member of the Board of International Speech Communication Association (ISCA 2002–2009), a Distinguished Lecture (2007–2008) and a member of the Overview Paper Editorial Board (since 2009) of the IEEE Signal Processing Society, and the general chair of ICASSP 2009 in Taipei. He is a fellow of ISCA since 2010, and received the Meritorious Service Award from IEEE Signal Processing Society in 2011.