INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

F_0 Contour Analysis Based on Empirical Mode Decomposition for DNN Acoustic Modeling in Mandarin Speech Recognition

Xiaoyun Wang[1,2], *Xugang Lu*[1], *Hisashi Kawai*[1], *Seiichi Yamamoto*[2]

[1] National Institute of Information and Communications Technology, Kyoto, Japan
[2] Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan

euo1101@mail4.doshisha.ac.jp,
{xugang.lu, hisashi.kawai}@nict.go.jp, seyamamo@mail.doshisha.ac.jp

Abstract

Tone information provides a strong distinction for many ambiguous characters in Mandarin Chinese. The use of tonal acoustic units and F_0 related tonal features have been shown to be effective at improving the accuracy of Mandarin automatic speech recognition (ASR) systems, as F_0 contains the most prominent tonal information for distinguishing words that are phonemically identical. Both long-term temporal intonations and short-term quick variations coexist in F_0. Using untreated F_0 as an acoustic feature renders the F_0 contour patterns differently from their citation form and downplays the significance of tonal information in ASR. In this paper, we explore the empirical mode decomposition (EMD) on F_0 contours to reconstruct F_0 related tonal features with a view to removing the components that are irrelevant for Mandarin ASR. We investigate both GMM-HMM and DNN-HMM based acoustic modeling with the reconstructed tonal features. In comparison with the baseline systems using typical tonal features, our best system using reconstructed tonal features leads to a 4.5% relative word error rate reduction for the GMM-HMM system and a 3.5% relative word error rate reduction for the DNN-HMM system.

Index Terms: tonal feature, F_0 contour, empirical mode decomposition (EMD), deep neural network (DNN), Mandarin speech recognition

1. Introduction

Mandarin Chinese is a tonal language that uses tone as the discriminative information for ambiguous characters. Many previous studies incorporating tonal information in continuous Mandarin speech recognition have showed great progress [1]–[7]. Two major methodologies have been used to integrate tonal information into Mandarin speech recognition – explicit and embedded tone modeling. In explicit tone modeling, tonal information and acoustic evidence are independently modeled and recognized [5]–[7], whereas in embedded tone modeling, tonal acoustic units are used and fundamental frequency (F_0) features are directly augmented to spectral features, and are recognized as part of the existing system [1, 2].

In Mandarin, F_0 is affected by many factors, such as long-term variations over the duration of prosodic units and short-term quick variations in the accented syllables [8, 9]. The F_0 contours usually characterize four tonal patterns of Mandarin Chinese in both methodologies mentioned above. However, surface F_0 contours of the speech signal show extensive variations, e.g., F_0 patterns of lexical tones influenced by contextual characters or tonal coarticulation, adjacent tones and phrase intonation [8]. Although this tone modeling obtains good performance in most state-of-the-art Mandarin automatic speech recognition (ASR) systems, the variations could render the models over-simplified tonal information and acoustic evidence, losting richer information between them [10].

To estimate the F_0 contours for ASR purposes, it is usually preferable to use a pitch track from a speech signal. Because of the specific nature of F_0, such as the fact that it is not defined for unvoiced regions, special treatment is required if it is to be used as an acoustic feature. There have been many effective techniques proposed for estimating F_0. RAPT [11] is a time domain algorithm for pitch tracking that makes a binary voicing classification regarding voiced or unvoiced frames in the speech signal. SAcC [12] is a new time domain algorithm that focuses on classifying the autocorrelations of a set of subbands using an MLP neural network. It provides robustness to noisy conditions in particular. The Kaldi pitch tracker [14] is a highly modified version of RAPT, and has been confirmed to achieve better accuracy than other canonical pitch trackers [11]–[13] for ASR systems. Conventionally, extracted F_0 related tonal features directly serve as acoustic features with tone augmentation. Although this approach alleviates the variance problems in acoustic modeling and improves ASR performance, untreated F_0 contours are still superimposed together with long-term temporal variations (intonation), short-term quick variations (coarticulations), and other factors that are irrelevant acoustic cue of tonal language for ASR purpose.

In this paper, we explore the empirical mode decomposition (EMD) [15, 16] method on an F_0 contour, which aims to reconstruct F_0 related tonal features with a view to removing the components of tonal information that are irrelevant for Mandarin speech recognition, by recombining a collection of intrinsic mode functions (IMF). The incentive for exploring EMD on F_0 is to adaptively obtain a series of particular pitch oscillations characterized by the physical time scales. Derived IMFs can obtain a full energy-frequency-time distribution of the tonal features, which include local energy, instantaneous frequency, and temporal intonation. We evaluated the recognition performance of both GMM-HMM and DNN-HMM based approaches to acoustic modeling along with the reconstructed F_0 related tonal features. The experimental results demonstrate that the reconstructed tonal features, as complementary tonal information, achieved a better performance than the other methods.

This paper is structured as follows. In Section 2, we illustrate F_0 extraction and the EMD analysis of F_0 contours. The baseline Mandarin systems and experimental results are presented in Section 3. We discuss the results and conclude with a brief summary of our study in Section 4.

http://dx.doi.org/10.21437/Interspeech.2016-653

2. F_0 extraction and EMD analysis of the F_0 contour

2.1. F_0 related tonal features in ASR for Mandarin Chinese

The use of F_0 as an acoustic feature greatly improves the performance of the ASR systems of tonal languages. However, the F_0 variations of tonal languages actually are constrained by coupling both individual lexical tones and various other factors, together with the temporal structures of the speech signal. In Mandarin Chinese, the F_0 contours of speech generally simultaneously manifest lexical tones at a local level and long-term temporal intonation at a global level. Previous studies have demonstrated that intonation perturbs lexical tones in terms of the change of the F_0 values of a speech signal when the sentence is uttered as an interrogatory narrative or with emphasized emotion [8, 9]. A few studies have found that the lexical tones are the most important factors for determining the local F_0 contours [18].

2.2. Effective tonal features extraction

To incorporate tonal information into the modeling features, we exploited the Kaldi pitch tracker [14] to extract the F_0 related tonal features and append them to the acoustic features. The pitch tracking algorithm used in the Kaldi ASR toolkit [17] is based on the famous *get_f0* [11] algorithm for F_0 estimation, but is a highly modified version of it. The main difference is that the Kaldi pitch tracker does not make a hard decision whether any given frame is voiced or unvoiced; instead, it assigns a pitch, even to unvoiced frames while constraining the pitch trajectory to be continuous [14]. After tracking, post-processing is used to interpolate the unvoiced region to avoid variance problems in acoustic modeling, and a short-time smoothing is used to reduce the noise in F_0.

The tonal features that serve as the baseline system include 3-element vectors, consisting of the probability of voicing (POV), F_0 and delta F_0. The POV feature is estimated from another pitch tracker called SAcC [12], using $log((p + 0.0001)/(1.0001 - p))$ as the resulting value, where $0 \leq p \leq 1$ is the voicing probability from SAcC. This aims to reduce the influence of pitch extraction errors. The details of F_0 smoothing and normalization are discussed in the next section.

2.3. EMD analysis of F_0 contour

In contrast to tones, intonation refers to the structured variation in F_0 that is not determined by lexical distinctions. The acoustic correlates of lexical tones and intonation inevitably interact with each other [19] because of the use of the same acoustic parameter F_0. Considering the purpose of ASR, intonation should be removed or normalized out from F_0 because it does not carry any lexical meaning for Mandarin Chinese. We hence investigate the structure of F_0 related tonal features using an EMD analysis of the F_0 contour.

EMD [15] is used as a temporal signal decomposition method to analyze data from nonstationary and nonlinear processes, and usually aims to filter out additive noise from the speech signal. Signals analyzed by EMD are decomposed into a series of oscillatory IMFs and a residual. The main purpose of EMD in this paper is not for signal denoising or data compression; instead, it is for the analysis of the F_0 related tonal feature, which aims to separate redundant components from relevant tonal information for effective tone modeling.

Combined with the description in Section 2.1, there are two main advantages to adopting EMD to extract relevant tonal information from the utterance F_0 contour. First, EMD is able to identify F_0 between two consecutive extreme, and decomposed F_0 contours can obtain the upper and lower envelope curves by interpolating local values. In contrast to lexical tones at a local level, long-term temporal intonation is mainly associated with tones at a global level. EMD should decompose the local tones and global intonation into individual components. Second, EMD is able to analyze F_0 in an entirely adaptive way, which is completely based on the local properties of F_0. It could assure the completeness of the F_0 contour reconstruction using IMFs.

2.3.1. EMD algorithm

The EMD decomposes an input signal into a series of IMFs through an iterative process called *sifting*. Each IMF needs to satisfy two conditions: (i) over the whole data, the numbers of extrema and zero crossings must either equal or differ at most by one and (ii) at any point, the mean value of the envelope defined by the local maxima and minima is a constant zero. The sifting process extracts a series of IMFs, which serves two main purposes: one is to eliminate the mutual overlapping waveform, and the other is to make a more symmetrical waveform. The main decomposition process can be summarized by the following steps:

(1) Calculate local mean value $m_l(t)$ of input signal $x(t)$ from extreme upper $e_u(t)$ and lower $e_l(t)$ envelopes as follows: $m_l(t) = (e_u(t) + e_l(t))/2$.

(2) Compute the difference based on the sifting process using $h_l(t) = x(t) - m_l(t)$.

(3) Iterate on the residual local trend until the corresponding difference satisfies the IMF properties, i.e., $h_{l^k}(t) = h_{l^{(k-1)}}(t) - m_{l^k}(t)$ at the $(k-1)^{th}$ sifting.

(4) Denote the corresponding IMF component by $c_n(t)$ after taking the $(k-1)^{th}$ sifting. Here $c_n(t) = h_{lk}(t)$.

The sifting process repeats several times until there are less than two extrema in the final computed residual. Input signal $x(t)$ is hence decomposed into a finite number of N IMFs and a final residual $r(t)$, as follows:

$$x(t) = \sum_{n=1}^{N} \text{IMF}_n(t) + r(t)$$

2.3.2. Reconstruction of F_0 contour

In the EMD algorithm, we note that the IMFs may have frequency overlaps, except at an instant time, and the instantaneous frequencies represented by each IMF are different. An example of the F_0 of an utterance from a data corpus and its IMF components are shown in Figure 1. The figure shows that :

(1) The upper order IMFs (e.g., C1 and C2) represent the high-frequency content of F_0, and are composed of faster oscillations than the middle (C3–C5) and lower order IMFs (C6 and C7), which in turn have faster fluctuations step-by-step.

(2) The EMD separates high-frequency versus low-frequency among the IMFs at each time interval.

(3) The different levels of IMF components mostly represent the F_0 variations between the IMFs at each time interval.

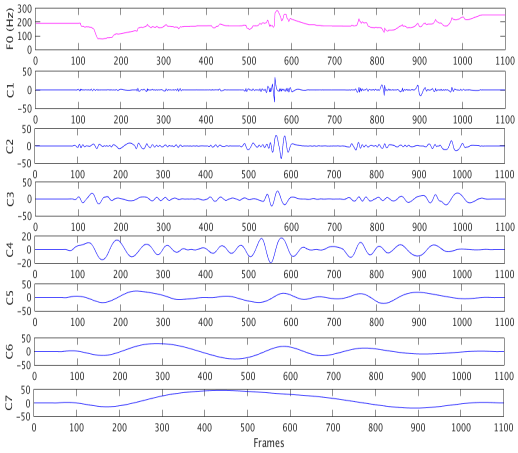

Figure 1: *IMFs obtained from the decomposition of the F_0 of a speech signal. Here, "C#" denotes the different level of IMF components from upper order (C1) to lower order (C7).*

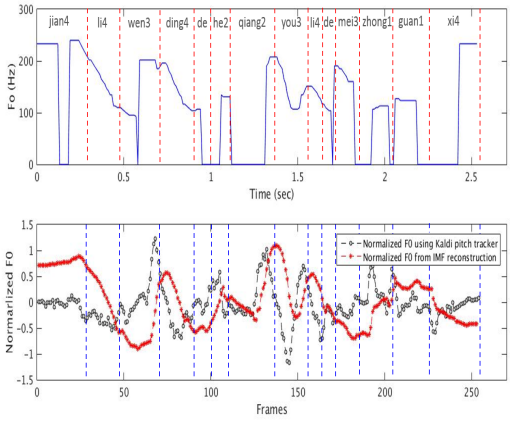

Figure 2: *Raw F_0 contour (top) and normalized F_0 features (bottom) using the Kaldi pitch tracker and IMF reconstruction for acoustic modeling in ASR systems. The vertical dashed lines show the toneme boundaries.*

The EMD behavior can be regarded as a filter bank with overlapping band-pass filters [20]. In an EMD analysis of the F_0 contour, the upper order IMFs are interpreted as the output of a high-pass filter, which usually represents the short temporal variation of the F_0 contour, such as quick co-articulation. The middle-order IMFs are interpreted as the output of the upper half-band part, which usually represents lexical tonal patterns. The remaining IMFs are roughly composed of long-term temporal variations caused by intonation. Among the decomposed F_0 variation, we therefore infer that the middle-order IMFs have the prominent lexical tonal information needed for acoustic model training.

2.3.3. Normalization of reconstructed F_0

A short-time smoothing method called moving window normalization [21] was used to reduce the noise of F_0 in this study. The difference between the baseline systems and our proposed system is that we normalize the reconstructed F_0 related tonal features with a POV-weighted mean subtraction for Mandarin ASR. The weighted average reconstructed pitch value was subtracted at each time t. The F_0 contour is computed over a window of width 151 frames centered at time t and weighted by the POV value p mentioned in Section 2.2.

Figure 2 shows the original raw F_0 contour of utterance from the data corpus, and the normalized F_0 features using the Kaldi pitch tracker (i.e., the baseline tonal feature), and the normalized F_0 related tonal features by recombining the IMF components for acoustic modeling. The vertical dashed lines show the toneme boundaries. As shown in Figure 2, the normalized F_0 from IMF reconstruction presents a more correct citation form of the lexical tone compared with the normalized original F_0 using the Kaldi pitch tracker. According to the study of [9], intonation frequently produces lexical tones at the initial and final positions of a sentence that behave differently from those at other positions, and the reconstructed F_0 improved the shape and scale of the lexical tones perturbed by these intonation, such as the initial tone 4 ("jian4") and the final one ("xi4") in Fig. 2.

3. Experiments

Compared to state-of-the-art GMM-HMMs, deep neural network (DNNs) have well-documented advantages for large vocabulary continuous speech recognition. In this section, we describe both GMM-HMM and DNN-HMM based Mandarin systems with typical tonal features, as these serve as our baseline systems.

3.1. Data corpus

Training sets for the acoustic modeling were constructed using 1997 Mandarin Broadcast News Speech data (HUB4-NE, LDC98S73), which contains about 27 h of speech, and the GALE Phase 2 Chinese Broadcast Conversation Speech (LDC2013S04), which contains about 97 h of speech. Development sets were taken from the training data, and consisting of 1 h of HUB4-NE and 3 h of the GALE Phase 2 Speech.

In addition to the transcripts of the acoustic training speech (3.7M characters), the GALE Phase 1 Chinese Broadcast News parallel text parts 1, 2, and 3 (LDC2007T23, LDC2008T08, and LDC2008T18) were added to the training data for the language modeling (LM) [22]. Meanwhile, the training texts provided by the HUB4 task (LDC95T13), containing 186M characters, were used to extend coverage of the LM. Two 4-gram based LMs were used in the evaluation, one was the LM trained by all the transcripts of the acoustic training speech, and the other is trained by the LM training text mentioned above. A linear interpolation of the two LMs was trained using the SRILM toolkit [23] with Kneser-Ney smoothing. The pronunciation lexicon of the LM consisted of about 100K words.

In this paper, we report the experiments on broadcast conversation speech, which was randomly selected from LDC2013S04 included about 3 h of speech.

3.2. Acoustic feature extraction

The acoustic models are trained on Mel-frequency cepstral coefficients (MFCC), which were extracted using a 25ms window and 10ms frame shift. Each frame of speech data was represented by a 39-dimensional feature vector including 13 MFCCs with their first and second derivatives. The three-dimensional

Table 1: Recognition results (in %WER) for GMM-HMM and DNN-HMM systems using various models with different types of features.

Feature	Model	WER
MFCC	ML SAT GMM	27.63
	MMI SAT GMM	25.40
	DNN (cross-entropy)	21.66
MFCC + pitch (baseline)	ML SAT GMM	25.86
	MMI SAT GMM	24.32
	DNN (cross-entropy)	19.87

Table 2: Recognition results (in %WER) for GMM-HMM and DNN-HMM systems using acoustic modeling with proposed tonal features. The relative error rate reduction in percent is given in parentheses. "Individual" means individual middle order IMFs, and "reconstructed" means F_0 reconstructed by recombining a collection of the middle order IMFs.

Feature	Model	WER
MFCC + pitch (individual)	ML SAT GMM	24.93 (3.60)
	DNN (cross-entropy)	19.28 (2.97)
MFCC + pitch (reconstructed)	ML SAT GMM	24.71 (4.45)
	DNN (cross-entropy)	19.17 (3.52)

F_0 features mentioned in Section 2.2 were appended to the spectral features, resulting in a 42-dimensional feature vector (MFCC + pitch). The MFCC and MFCC + pitch vector were spliced in time, taking a context size of seven frames (three on each side of the current frame), and projecting the dimension of the concatenated vector down to 40 dimensions using linear discriminate analysis (LDA). The resulting features were further de-correlated with a maximum likelihood linear transformation (MLLT), which is also known as the global semi-tied covariance transform. Moreover, speaker adaptive training (SAT) was performed using a single feature-space maximum likelihood linear regression transform, estimated per speaker.

3.3. Subsystem descriptions and performance

Both GMM-HMM and DNN-HMM systems were built in this study. We adopt 181 tonal phonemes (tonemes) consisting of consonants and five tonal vowels, which are distinct from the five types position boundaries, as the basic HMM units. The GMM-HMM systems were trained on the LDA+MLLT+SAT features described above. The HMM models were constructed with a maximum of 8,000 tied triphone states and each state had 16 Gaussian mixture components. We compared the results of the models trained using ML with those using discriminative training applied with a feature space boosted MMI (FBMMI) followed by model space boosted MMI (BMMI) training.

The DNNs were trained on the same LDA + MLLT + SAT features as the GMM-HMM, and the only difference with GMM-HMM was that the features were globally normalized to have zero mean and unit variance. The input to the network was a 15 frame (seven frames on each side of the current frame) context window of the 40-dim features, and projected down to 300 dimensions using LDA. The network had 300 nodes as input and six layers (i.e., five hidden layers), where each hidden layer had 2100 neurons and 8000 nodes as output. The DNNs were trained using cross-entropy at frame-level. Mini-batch stochastic gradient descent (SGD) was used to minimize the cross-entropy between the labels and network output. The utterance frames were presented in a randomized order, and the SGD used mini-batches of 256 frames. An exponentially decaying schedule was used that started with an initial learning rate of 0.008. In this schedule, rate is halved when the improvement in the frame accuracy on the cross-validation set between two successive epochs falls below 0.01%. The stop criterion occurs when the frame accuracy increases by less than 0.001%. A single GPU (Tesla K20m) was used to accelerate the training process.

The GMM-HMM and DNN-HMM systems that we built with MFCC + pitch features serve as the baseline systems for verifying the proposed method. Table 1 presents the recognition results of the baseline systems with the original tonal features and compares them with those using MFCC features only. A comparison of the recognition results for each systems shows that the appended original tonal features significantly improve performance in Mandarin ASR for the task of broadcast conversation speech.

3.4. Performance of the EMD-based tonal features

Based on the findings of Section 2.3, we applied the middle order IMFs of F_0 as augmented acoustic features to investigate their effect on the recognition performance in both GMM-HMM and DNN-HMM systems. In this study, either reconstructed F_0 based on the middle order IMFs or an individual middle order IMF were used as the expanded acoustic features.

The recognition results of the two systems using acoustic modeling with reconstructed F_0 related tonal features are shown in Table 2. Here, "individual" indicates that individual middle order IMFs has been used as the expanded acoustic features, and "reconstructed" means that tonal features reconstructed by recombining a collection of the middle order IMFs were used as the expanded acoustic features. In comparison with the results of Table 1, both individual and reconstructed GMM-HMM and DNN-HMM systems achieved better recognition performance than those with the original tonal features.

4. Discussion and Conclusion

In this study, we explored the EMD method on the F_0 contour of Mandarin Chinese speech for ASR. Based on our analysis, we proposed to reconstruct F_0 related tonal features with a view to removing the irrelevant components of tonal information by recombining a collection of IMFs. The reconstructed F_0 contour was constructed to capture the more prominent lexical tonal patterns as complementary tonal information for acoustic modeling in Mandarin ASR systems.

The experimental results show that the acoustic modeling with reconstructed F_0 related tonal features was able to further improve the recognition performance for Mandarin Chinese compared with that using typical tonal features. Even directly appending the decomposed tonal features as individual components to the acoustic features improved recognition performance. Although the DNN automatically helps to determine which features are useful for learning, the potential for variance still exists. The proposed method should further alleviate variance problems in acoustic modeling not only for Mandarin ASR systems but also for the ASR systems of other tonal languages.

5. References

[1] E. Chang, J. L. Zhou, S. Di, C. Huang, and K. F. Lee, " Large vocabulary Mandarin speech recognition with different approaches in modeling tones," *Proceedings of INTERSPEECH*, pp. 983–986, Oct. 2000.

[2] H. C. H. Huang and F. Seide, " Pitch tracking and tone features for Mandarin speech recognition," *Proceedings of ICASSP*, pp. 1523–1526, Jun. 2000.

[3] X. Lei, M. H. Siu, M. Y. Hwang, M. Ostendorf and T. Lee, " Improved tone modeling for Mandarin broadcast news speech recognition," *Proceedings of INTERSPEECH*, 2006.

[4] W. Gu and T. Lee, " Effects of tonal context and focus on Cantonese F_0," *Proceedings of ICPhS*, pp. 1033–1036, Aug. 2007.

[5] C. Wang, "Prosodic modeling for improved speech recognition and understanding," *Ph.D. thesis, Massachusetts Institute of Technology*, 2001.

[6] T. Lee, W. Lau, Y. W. Wong, and P. C. Ching, "Using tone information in Cantonese continuous speech recognition," *ACM Trans. Asian Lange. Inf. Process*, vol. 1, no. 1, pp. 83–102, 2002.

[7] L. Cheng and L. Lee, "Improved large vocabulary Mandarin speech recognition by selectively using tone information with a two-stage prosodic model," *Proceedings of INTERSPEECH*, 2008.

[8] Y. Xu, "Effects of tone and focus on the formation and alignment of F_0 contours," *Journal of phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[9] X. N. Shen, "Interplay of the four citation tones and intonation in Mandarin Chinese," *Journal of Chinese Linguistic*, vol. 17, no. 1, pp. 61–74, 1989.

[10] Y. B. Wang, S. W. Li, and L. S. Lee, "An Experimental Analysis on Integrating Multi-Stream Spectro-Temporal, Cepstral and Pitch Information for Mandarin Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 710, pp. 2006–2014, 2013.

[11] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," *Speech coding and Synthesis*, pp. 495–518, 1995.

[12] B. S. Lee and D. PW. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," *Proceedings of INTERSPEECH*, 2012.

[13] A. De Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America* , vol. 111, no. 4, pp. 1917–1930, 2002.

[14] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," *Proceedings of ICASSP*, pp. 2494–2498, 2014.

[15] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shinh, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–994, 1998.

[16] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "A new algorithm for instantaneous F 0 speech extraction based on ensemble empirical mode decomposition," *Proceedings of Signal Processing, 17th European*, IEEE, pp. 2347–2351, 2009.

[17] D. Povey, A. Ghoshal, et al, "The Kaldi speech recognition toolkit," *Proceedings of ASRU*, 2011.

[18] E. Gårding "Speech act and tonal pattern in Standard Chinese: constancy and variation," *Phonetica*, vol. 44, no. 1, pp. 13–29, 1987.

[19] Y. Xu, and Q. E. Wang "What can tone studies tell us about intonation?" *Intonation: Theory, models and applications*, pp. 337–340, 1998.

[20] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[21] X. Lei, "Modeling lexical tones for Mandarin Large Vocabulary Continuous Speech Recognition," *Ph.D. thesis, University of Washington, 2006*.

[22] X. Hu, X. Lu, and C. Hori, "Mandarin speech recognition using convolution neural network with augmented tone features," *Proceedings of ISCSLP*, pp. 15–18,2014.

[23] A. Stolcke, "SRILM-an extensible language modeling toolkit," *Proceedings of INTERSPEECH*, Sep. 2002.