TOKENIZING FUNDAMENTAL FREQUENCY VARIATION FOR MANDARIN TONE ERROR DETECTION

Rong Tong, Nancy F. Chen, Boon Pang Lim, Bin Ma and Haizhou Li

Institute for Infocomm Research, Singapore

{tongrong,nfychen,bplim,mabin,hli}@i2r.a-star.edu.sg

ABSTRACT

Tone error is commonly observed in tonal language acquisition. Correct tone production is especially challenging for native speakers of non-tonal languages. In this paper, we exploit the fundamental frequency variation (FFV) feature for Mandarin tone error detection. We propose to use FFV through two approaches: (1) Concatenating FFVs along side with standard speech recognition features; (2) Token FFV: Characterizing pitch variation with longer temporal context through GMM tokenization and n-gram language modeling. Our results show that tone error detection improves by incorporating FFV features and the two approaches are complementary to each other.

Index Terms— computer assistant language learning (CALL), computer-assisted pronunciation training (CAPT), tone recognition

1. INTRODUCTION

A computer-assisted language learning (CALL) system provides an easy interface for language learners. Computer assistant language learning systems usually provide segmental and suprasegmental level feedbacks on non-native speech input. The suprasegmental level feedback focuses on the rhythm, stress, and intonation of the speech [1, 2], while the segmental feedback focuses on the pronunciation accuracy of the individual phonetic units [3, 4].

Mispronunciation occurs in both phonetic and prosodic aspects. Tone error is a special case of mispronunciation for tonal languages. How to produce tones correctly is one of the major challenges for non-native language (L2) learners, especially for learners whose native language (L1) is not tonal. Feedback about tonal error is an essential feature in a tonal language learning system, thus this work focuses specifically on the task of Mandarin tone error detection of non-native speech.

Auditory analysis and acoustic analysis have been performed to study the characteristics of Mandarin Chinese tone acquisition and production[5]. A cross-linguistic study on single-word utterances [6] reveals that Mandarin speakers have higher means and larger ranges of F0 than English speakers. Study on the difference of pitch range among native and non-native Chinese speakers [7, 8] suggests a non-native speaker needs to widen his pitch range to produce Mandarin tones correctly.

Fundamental frequency (F0) is one of the most important acoustic cues for tone modeling. A statistics based pitch contour model is proposed [9] for Mandarin TTS. Tone relevant features based on pitch flux [10] is proposed for Chinese dialect identification. F0, duration and energy features are used to capture tone characteristics in automatic speech recognition (ASR) [11] and Mandarin tone recognition [12]. A F0 smoothing method is proposed in [13] to improve the performance in Mandarin tone recognition.

Many studies exploit modeling methods to derive better models to present tone information. Dynamic Bayesian network is used to model MFCC and pitch features [14] for tone recognition. An SVM classifier is used in [15] for tone recognition in continuous Mandarin speech. In [16], context-dependent tone models are built by considering lexical information.

Apart from F0, other features are exploited in Mandarin tone recognition. In [17], acoustic features derived from voice quality analysis are used for Mandarin tone recognition. A deep neural network (DNN) classifier is reported to achieve good tone classification performance without including pitch information [18]. Discrete Cosine Transform Coefficients and Discrete Cosine Series Coefficients are studied [19] on accented Mandarin speech.

In this work, we propose to use fundamental frequency variation (FFV) feature [20] for tone error detection on non-native speech. The idea of FFV is to derive a vector to characterize the within-frame variation of fundamental frequency. It has been shown to be useful in speaker change prediction [21], automatic speech recognition for both tonal and non-tonal languages [22], and low-resource keyword search [23].

In this paper, we first concatenate FFV along side with standard speech recognition features. We also propose an alternative method, Token FFV, motivated by our experience in language and speaker recognition [24, 25, 26, 27] that modeling longer temporal context complements using spectral features. With Token FFV approach, FFV features are tokenized by Gaussian component index based n-grams to model longer time spans of information beyond the frame level.

2. LEXICAL TONES IN MANDARIN CHINESE

Mandarin Chinese is a monosyllabic language, where each character is a single syllable. Each syllable consists of an optional initial, a final and a tone.

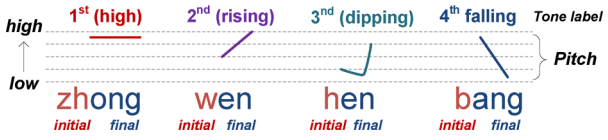

Fig. 1: Mandarin syllable structure

Hanyu Pinyin is the most widely adopted Mandarin phonetic system. Figure 1 shows the pinyin presentation of Mandarin syllable structure. Mandarin Chinese has 4 lexical tones and one neutral tone. The differences of the tones are characterized by their pitch contours: Tone 1 is a high level tone; Tone 2 is rising from mid pitch to high pitch; Tone 3 stats low, it falls slightly then rises; Tone 4

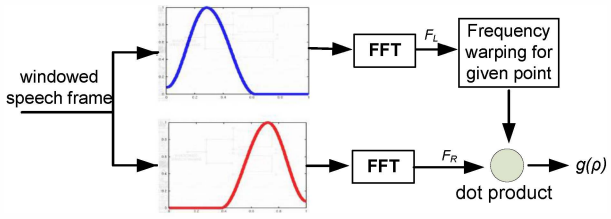

Fig. 2: Deriving the FFV spectrum

starts from high pitch and then falls to low. Neutral tone 5 has no specific contour; it is equivalent to an unstressed syllable. In this work, tone error detection is focused on the detection of tone 1-4.

3. FUNDAMENTAL FREQUENCY VARIATION

3.1. Rationale

Fundamental frequency (F0), as the acoustic correlate of pitch, is one of the most important acoustic cues for tone modeling. F0 and its derivative features are commonly used in Mandarin tone recognition [9, 11, 12]. Unlike F0, a highly post-processed scalar value from frame to frame, fundamental frequency variation (FFV) represents pitch variation per frame in vector-form.

The derivation of FFV feature is based on the following observation: the rate of F0 change of two adjacent speech frames can be inferred by finding the dilation factor required to optimally align the harmonic spacing in their magnitude frequency spectra [28]. Thus FFV extraction relies on the comparison of the frequency magnitude spectra of the left and right halves of each analysis frame.

The FFV feature has three advantages: (1) it is estimated locally from each frame – it does not require peak identification and landmark detection, as required for many pitch tracking algorithms; (2) the vector representation is more flexible for advanced modeling; (3) while F0 is undefined for unvoiced regions, FFV does not suffer from this limitation.

3.2. FFV Feature Extraction Procedure

Given a speech signal after pre-emphasis, the signal is partitioned into 32 ms overlapping frames. The computation of FFV then takes two steps: (1) Deriving the FFV Spectrum; (2) Characterize different speeds of pitch change through filter banks.

3.2.1. Deriving the FFV Spectrum

Figure 2 shows how the FFV spectrum is derived. For each frame, two symmetrically shaped windows with their centers of gravity on the left and right sides of the original speech frame, are overlaid on the signal and used to extract the magnitude spectra of two corresponding left subframe F_L and right subframe F_R. F_L and F_R are 512 point Fourier transforms, computed every 8 ms.

Frequency warping is then applied to the left subframe F_L, with the scaling factor ρ taking on different values to characterize the corresponding rate of change in pitch as the signal progresses from the left to the right subframe. A normalized dot product $g(\rho)$ is derived from the warped left subframe and unwarped right subframe.

The dot product value $g(\rho)$ represents how well the particular rate of pitch change corresponds to the current frame. When the frequency warping precisely accounts for the amount of pitch variation from the left to the right frame, the dot product value $g(\rho)$ takes a value close to 1. Thus, the FFV spectrum vector is computed by varying ρ:

$$[\quad g(\rho_1)\ g(\rho_2)\ g(\rho_3)\ \ldots\ g(\rho_t)\quad] \tag{1}$$

3.2.2. Characterizing different speeds of pitch change

In the second step, the FFV spectrum derived from the first step is compacted using a 7-point filter bank, each filter capturing the varying speed ranges for pitch variation, as shown in Table 1. The shape of the filters depend on the speed of pitch variation it is attempting to characterize. For example, the filters corresponding to very fast pitch changes use rectangular filters to retain the informative behavior of unvoiced frames, which tend to have flat rather than decaying tails in the pitch variation spectrum. After applying the 7 filter banks, the dimension of the FFV spectrum reduces from 512 to 7.

Filter bank	Description
1 trapezoidal filter	capture perceptually flat pitch
2 trapezoidal filters	slowly changing pitch : rising and falling
2 trapezoidal filters	rapidly changing pitch
2 rectangular filters	unvoiced frames which have flat tails

Table 1: Function of 7 Filter banks used to derive FFV vector

4. MANDARIN TONE ERROR DETECTION

In Mandarin speech recognition, tonal phones are commonly used to model phone and its tone variations. The tone information is modeled together with the lexical information during the acoustic model training process. In this work, we attempt to improve tone error detection by two methods: (1) concatenating FFV to the acoustic features in ASR; (2) modeling tones by tokenizing FFV features.

4.1. Error detection using ASR Confidence Measure

The Goodness of Pronunciation (GOP) [29] is a phone level confidence measure to gauge how a particular phone is pronounced differently compared to a native model. Given phone p, the GOP score can be derived by:

$$GOP(p) = \frac{1}{n}\frac{P(O|p)P(p)}{\max_{q\in Q} P(O|q)P(q)} \tag{2}$$

where O is the acoustic observation, which is typically MFCC features but one can also concatenate pitch related features along with MFCC's, Q is the set of all phones, n is the number of frames; $P(O|p)$ stands for the likelihood of the observation on model p, it can be obtained by performing forced alignment with the canonical transcription; $\max_{q\in Q} P(O|q)$ is the maximum likelihood of all the phones in phone inventory, often derived from a phone loop recognition process.

Phone level GOP scores are first normalized by global GOP mean and variance. Syllable level scores are obtained by interpolating the phone level GOP scores, as defined in Equation (3), P_{in} is the normalized GOP score of the initial phones and P_{fi} is the normalized GOP score of the final phones. The interpolation parameter α is determined empirically from the development set.

$$S = \alpha * P_{in} + (1-\alpha) * P_{fi} \tag{3}$$

4.2. Token FFV: Exploiting Sub-Syllabic Pitch Variation

4.2.1. Motivation

We propose a Token FFV method using GMM tokenization and n-gram language modeling technique. The proposed Token FFV approach is inspired by our experience in automatic language and speaker recognition where tokenization and language modeling have shown to be successful in capturing acoustic characteristics across a

relatively longer temporal context. Since Mandarin tones are characterized by pitch contours of the final portion of the syllable (a *final* consists of either a vowel, or vowel plus a nasal consonant), a longer temporal span beyond the frame level is desirable in modeling Mandarin tones. We aim to characterize pitch variations of time units at the sub-syllabic level instead of the syllabic level or beyond since lexical tone distribution in Mandarin is relatively uniform. (Recall that each Mandarin syllable is attached to one lexical tone.)

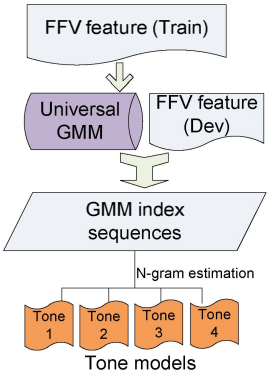

Fig. 3: Token FFV for tone error detection

4.2.2. GMM Tokenization followed by n-gram Language Modeling

Figure 3 shows the steps of proposed Token FFV approach. Given a speech signal, the phonetic boundary for each syllable can be obtained by performing forced alignment using an ASR system. With the phonetic boundaries, the corresponding FFV features for each syllable are extracted and the syllables are labeled as tone 1-4 according to the phonetic transcription. The set of FFV features are separated into train and development set.

A GMM universal background model (UBM) is built using all the FFV feature vectors of the train set. Two gender dependent models are derived by adapting the UBM using FFV feature vectors of corresponding genders. The GMM tokenization process [24, 25] is as follows. For each frame i in the training set, label j is assigned: $j = \arg\max_j P(i|c_j)$, where c_j is the Gaussian mixture component, $j = 1, ..., M$. In this way, each syllable is converted into a GMM index sequence. This GMM index sequence presents the pitch variation of the given syllable.

The GMM index sequences for each of the 4 tone classes are combined and used to derive a tone model using n-gram language modeling approach. The n-gram language modeling process captures the pitch variation information among n consecutive frames. Compared with the frame based F0, the proposed Token FFV method captures the pitch variation in relatively longer time spans.

5. EXPERIMENT

5.1. Corpus

Two native Mandarin speech corpus are used in the experiments. The King-ASR-118 corpus [30] is used for acoustic model training. To further model microphone channel effects and reading-style speech, an internal corpus is used. This corpus is recorded from Mandarin speakers in Beijing and Shanghai in China. Each speaker read 350 utterances; on average each test utterance is 8 syllables. The corpus is split into train and test portions. The training set has about 450 speakers and the test set consists of 1406 utterances from 4 speakers.

The non-native speech corpus used in this study is iCALL corpus [2, 31]. In this corpus, 300 beginning learners of Mandarin Chinese were asked to read 300 Pinyin prompts. Each speaker received a different set of utterances. The speechwas recorded in quiet office rooms. The short utterances of the non-native corpus is split into development and test portions. The development set consists of short utterances from 233 speakers, they are used for parameter tuning. The non-native test set consists of 1887 utterances from 59 speakers, where each utterance has 2 syllables.

Table 2 summaries the native and non-native corpora used in this study, each corpus is separated as Train and Test set. There is no speaker overlap between the Train and Test sets.

Type	Train	Test
Native	King-ASR-118 (1175) Beijing/Shanghai (456)	native (4)
Non-native	iCALL dev (233)	non-native (59)

Table 2: Train and Test sets. The number of speakers are in brackets.

5.2. Automatic Speech Recognition

Two ASR systems were trained using two types of acoustic features, they are denoted as DNN-MFCC and DNN-MFCC-FFV. The feature vector of the DNN-MFCC system consists of 13 dimensional MFCC feature in conjunction with 1 dimension of F0, and their derived deltas, acceleration and third-order deltas. The feature dimension of the DNN-MFCC is 56. The DNN-MFCC-FFV system is trained from the same 56 dimensional feature concatenating 7 dimensional FFV features. The feature vector dimension is 63.

Both ASR system follow the same training mechanism using Kaldi toolkit: a baseline acoustic model is trained with Maximum Mutual Information (MMI) criterion, then DNN training is performed using the phone level alignment obtained from the MMI model. There are 5 hidden layers in the DNN models. In both systems, there are 175 phones and 8537 tied states.

ASR setup	Native	Non-native
DNN-MFCC	48.68	63.30
DNN-MFCC-FFV	47.66	62.31

Table 3: Syllable error rate of native and non-native test sets

Table 3 shows the speech recognition results of the two ASR systems on native and non-native test sets. For all the experiments, the tonal syllable loop grammar is used. The ASR performance of both native and non-native test sets are slightly improved by incorporating FFV features. Although the DNN-MFCC system includes 1 dimensional F0 in acoustic feature vector, incorporating FFV still gives a small improvement in the speech recognition accuracy.

5.3. Tone error detection

5.3.1. Performance measure

Two types of errors are examined for tone error detection performance: false rejection rate (FRR) and false acceptance rate (FAR), where FAR= n_{fa}/n_{tn} and FRR= n_{fr}/n_{tp}. n_{fr} is the number of correct tones that are mis-classified, n_{tp} is the total number of correctly pronounced tones, n_{fa} is the number of wrong tones that are mis-classified as correct, and n_{tn} is the total number of wrongly pronounced tones. FAR and FRR have a trade-off relationship.

5.3.2. Tone error detection with ASR confidence measure

The two ASR systems used for deriving GOP score to detect tone errors are implemented as in 4.1. The raw GOP score is normalized

by the mean and standard variance of each speaker. For each final, a decision threshold is tuned from the non-native development data.

	DNN-MFCC		DNN-MFCC-FFV	
	*F*AR	*F*RR	*F*AR	*F*RR
Tone 1	0.28	0.27	0.26	0.27
Tone 2	0.29	0.30	0.28	0.27
Tone 3	0.29	0.30	0.30	0.30
Tone 4	0.20	0.21	0.20	0.19
All	0.27	0.27	0.26	0.26

Table 4: Tone error detection results using ASR output

Table 4 shows the tone error detection results. Though FFV gives only limited improvement on the speech recognition performance, the overall tone error detection accuracy is consistently improved by using the DNN-MFCC-FFV system. This suggests that FFV provides additional tonal information compared to F0.

5.3.3. Token FFV

We evaluate the proposed Token FFV approach (section 4.2) for tone error detection. The non-native development set is further separated into two portions with a ratio of 6:4. The first portion is used to train two gender dependent universal GMMs, each consisting of 256 mixture components.

The FFV features from the second portion are evaluated on the corresponding universal GMM and tokenized GMM index sequences. The GMM index sequences derived from the same tone class are used to train n-gram language model using SRILM toolkit. In our experiment, $n = 5$ for each tone class. In the detection process, each test syllable is tokenized using the gender matched universal GMM to derive GMM index sequence; the GMM index sequence is evaluated on each of the 4 tone models, the tone model that gives the best perplexity is assigned as the tone class label.

Token FFV models pitch variation at a time span of 50 ms (since $n = 5$ for the n-gram language modeling step and each frame is 10 ms), which roughly corresponds to the duration of short vowels, making it a sub-syllable time unit. This longer time span could provide complementary pitch variation information to frame-level characterizations of pitch and FFV.

	Tone 1	Tone 2	Tone 3	Tone 4	All
FAR	0.33	0.46	0.52	0.38	0.47
FRR	0.24	0.28	0.31	0.22	0.26

Table 5: Tone error detection results with Token FFV

Table 5 shows the tone error detection results of individual tone class and overall performance using Token FFV method. One observation is the high FAR rate, which might be due to the fact that only native data (correctly produced tones) are considered in the n-gram language modeling and decision, hence the mispronounced tone patterns are not well captured. Detection error is higher for Tone 3, which is not surprising because the pitch contour of Tone 3 is the most complex among the four tones.

5.3.4. Fused system: DNN-MFCC-FFV+Token FFV

The two proposed tone error detection systems are fused. For each syllable, the perplexity score derived from Token FFV system is first normalized to [0-1], then it is interpolated with the GOP score derived from the ASR system (as shown in Eq.3). Figure 4 compares the overall tone error detection performance of the DNN-MFCC-FFV system and the fused system: DNN-MFCC-FFV+Token FFV. The detection error trade-off (DET) curves are plotted, where x and y axis indicate FRR and FAR respectively.

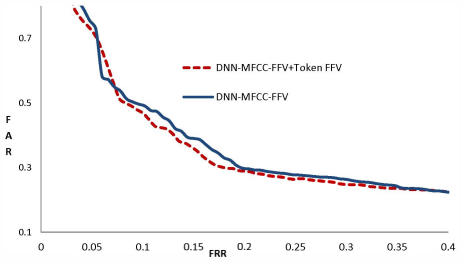

Fig. 4: Tone error detection results with and without Token FFV

Figure 4 shows that tone error detection is further improved by the fusion of the two methods, especially in the low FAR and FRR area. This confirms our assumption that the pitch variation is not fully exploited by concatenating FFV feature in acoustic feature, the explicit tone modeling with Token FFV provides additional tone discriminating information to the GOP based method.

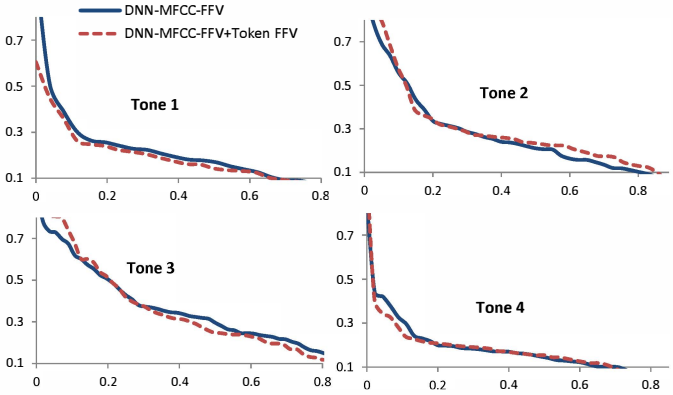

Fig. 5: Individual tone error with and without Token FFV

Figure 5 shows the DET curves for each tone class. We can see that Token FFV improves detection for Tone 1 and 4, but not necessarily for Tone 2 and 3. One plausible explanation is due to the pitch contour characteristics of the four tones. Tone 2 and Tone 3 generally have greater variation in their pitch contours, whereas Tone 1 manifests largely in the unchanged tone category and Tone 4 manifests in the rapidly falling category. In terms of the pitch variation spectrum, it is possible that Tone 1 and Tone 4 fall more cleanly into the categories of pitch variation characterized by the the 7-point filterbank in FFV, and thus their corresponding FFV features are better captured by the Token FFV model.

6. CONCLUSIONS AND FUTURE WORK

In this work, we investigate how to exploit FFV for tone error detection on non-native Mandarin. Our experiment results show that FFV provides additional tonal information when concatenated to GOP scores. In addition, we propose a Token FFV modeling approach, capturing sub-syllabic time spans of pitch variation. The experiment results shows that Token FFV provides complements GOP based tone error detection.

For future work, we plan to refine the Token FFV approach to incorporate discriminative training and further investigate how to improve detection rates for Tone 2 and Tone 3.

7. REFERENCES

[1] Catia Cucchiarini, Helmer Strik, and Lou Boves, "Quantitative assessment of second language learners fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1989–1999, 2000.

[2] Rong Tong, Boon Pang Lim, Nancy F. Chen, Bin Ma, and Haizhou Li, "Subspace Gaussian mixture model for computer assistant language learning," in *ICASSP*, 2014.

[3] Ann Lee and James Glass, "A comparison-based approach to mispronunciation detection," in *SLT*, 2012.

[4] Ke Yan and Shu Gong, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," *International Journal of Information Technology and Computer Science*, pp. 17–23, 2011.

[5] Bei Yang, *A model of Mandarin tone categories – a study of perception and production*, Ph.D. thesis, The University of Iowa, 2010.

[6] Gwang Tsai Chen, *A comparative study of pitch range of native speakers of Midwestern English and Mandarin Chinese: An acoustic study*, Ph.D. thesis, University of Wisconsin-Madison, Madison, 1972.

[7] Patricia Keating and Grace Kuo, "Comparison of speaking fundamental frequency in English and Mandarin," *Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1050–1060, 2012.

[8] Li Juan Guo and Liang Tao, "Tone production in Mandarin Chinese by American students: A case study," in *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)*, 2008, pp. 123–138.

[9] Sin-Horng Chen, Wen-Hsing Lai, and Yih-Ru Wang, "A statistics-based pitch contour model for Mandarin speech," *Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 908–925, 2005.

[10] Bin Ma, Donglai Zhu, and Rong Tong, "Chinese dialect identification using tone features based on pitch flux," in *ICASSP*, 2006.

[11] Hong Xiu Wei, Xin Hao Wang, Hao Wu, Ding Sheng Luo, and Xi Hong Wu, "Exploiting prosodic and lexical features for tone modeling in a conditional random field framework," in *ICASSP*, 2008.

[12] Hussein Hussein, Hansjorg Mixdorff, and Rudiger Hoffmann, "Real-time tone recognition in a computer-assisted language learning system for German learners of Mandarin," in *ICCL*, 2012.

[13] Qian Liu, Jinxiang Wang, Mingjiang Wang, Panpan Jiang, Xirui Yang, and Jiayuan Xu, "A pitch smoothing method for Mandarin tone recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 4, 2013.

[14] Xin Lei, Gang Ji, Tim Ng, Jeff Bilmes, and Mari Ostendorf, "DBN-based multi-stream models for Mandarin toneme recognition," in *ICASSP*, 2005.

[15] Shui Ping Wang, Zhen Ming Tang, Ying Nan Zhao, and Sai Ji, "Tone recognition of continuous Mandarin speech based on binary-class svms," in *Fourth International Conference on Natural Computation (ICISE)*, 2009.

[16] Conggui Liu and Jinxu Tao, "Mandarin tone recognition considering context information," in *Signal Processing, Communication and Computing (ICSPCC)*, 2013.

[17] Surendran Dinoj and G-A Levow, "Can voice quality improve Mandarin tone recognition?," in *ICASSP*, 2008.

[18] Neville Ryant, Jia Hong Yuan, and Mark Liberman, "Mandarin tone classification without pitch tracking," in *ICASSP*, 2014.

[19] Jiang Wu, Stephen A Zahorian, and Hongbing Hu, "Tone recognition for continuous accented Mandarin Chinese," in *ICASSP*, 2013.

[20] Kornel Laskowski, Jens Edlund, and Mattias Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversation dialogue system," in *ICASSP*, 2008.

[21] Kornel Laskowski and Qin Jin, "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum," in *ICASSP*, 2009, pp. 4541–4544.

[22] Florian Metze, Zaid A. W. Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, and Van Huy Nguyen, "Models of tone for tonal and non-tonal languages," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 261–266.

[23] Nancy F. Chen, Sunil Sivadas, Boon Pang Lim, Hoang Gia Ngo, Haihua Xu, Van Tung Pham, Bin Ma, and Haizhou Li, "Strategies for Vietnamese keyword search," in *ICASSP*, 2014.

[24] Bin Ma, Donglai Zhu, Rong Tong, and Haizhou Li, "Speaker cluster based GMM tokenization for speaker recognition," in *INTERSPEECH*, 2006.

[25] Pedro A Torres-Carrasquillo, Douglas A Reynolds, and JR Deller Jr, "Language identification using Gaussian mixture model tokenization," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1.

[26] Rong Tong, Bin Ma, Haizhou Li, and Eng Siong Chng, "A target-oriented phonotactic front-end for spoken language recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 1335–1347, 2009.

[27] Haizhou Li, Kong Aik Lee, and Bin Ma, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, pp. 1136 – 1159, May 2013.

[28] Kornel Laskowski and Jens Edlund, "A snack implementation and Tcl/Tk interface to the fundamental frequency variation spectrum algorithm," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[29] Silke Maren Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, Cambridge University, 1999.

[30] Chinese Mandarin Mobile Speech Recognition Database, "http://www.speechocean.com/en-news/783.html," .

[31] Nancy F. Chen, Vivaek Shivakumar, Mahesh Harikumar, Bin Ma, and Haizhou Li, "Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages," in *INTERSPEECH*, 2013, pp. 803–806.