Improved Tonal Language Speech Recognition by Integrating Spectro-temporal Evidence and Pitch Information with Properly Chosen Tonal Acoustic Units

Shang-wen Li[1], Yow-bang Wang[2,3], Liang-che Sun[1] and Lin-shan Lee[1,2,3]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2] Graduate Institute of Electrical Engineering, National Taiwan University, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan

b94901123@ntu.edu.tw, {eric, lgsun}@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

We propose an improved Tandem system for tonal language speech recognition. Three different types of features, cepstral, spectro-temporal and pitch features, are integrated for modeling tone and phoneme variation simultaneously. Tonal phonemes (or tonemes) are used for MLP posterior estimation, and tonal acoustic units for HMM recognition. In our experiments conducted on Mandarin broadcast news, a 19.3% relative CER reduction was achieved over the conventional MFCC Tandem baseline. With different training acoustic units, we analyze the complementarity among the three types of features in tone, phoneme, and toneme classification.

Index Terms: spectro-temporal features, pitch, Tandem system, LVCSR

1. Introduction

Mandarin Chinese is a tonal language, in which every syllable is assigned a tone, and the tone has lexical meaning that helps distinguishing ambiguous words. In recent years, much work has been reported which incorporated tone information in large vocabulary continuous Mandarin speech recognition and achieved significant improvement. Such work can be roughly divided into two categories: explicit and embedded tone modeling [1]. In the former, tones and phonemes were modeled independently and then combined through including likelihoods of tone models in acoustic scores [2] or post-processing on the word lattices [1, 3]. In contrast, embedded tone modeling appended the tone-related features to spectral features and modeled the tonal acoustic units with the existing automatic speech recognition (ASR) framework [4, 5]. The introduction of Tandem systems [6] greatly improved the performance of embedded modeling [7]; the smoothed log-pitch features were concatenated with PLP features, a context window of the augmented features was used to capture the long-term information of tones and phonemes, a multi-layer perceptron (MLP) then transform those features to posteriors for tonal acoustic units, and the posteriors serve as observations for ASR systems based on Hidden Markov Models (HMMs). Recently, the original input features of MLP were incorporated with the modulation spectrum representations extracted from a bank of temporal filters [8] and showed even larger performance gain [9].

Contrary to temporal modulation extracted by context window or temporal filters, spectro-temporal modulation also plays an important role in speech signals. Intonation, coarticulation, and transition across phonemes naturally produce sloped patterns on the 2-D spectrogram. This is supported by recent findings in physiological experiments which showed that a large percentage of neurons in the primary auditory cortex of mammal species respond to signals with different spectro-temporal modulations [10]. These findings motivated substantial efforts in parameterizing those behaviors: 2-D Gabor filters were used for filtering the spectro-temporal modulation from spectrogram [11, 12]; autoregressive model and frequency domain linear prediction were utilized to extract the information [13]; approaches using independent component analysis and non-negative sparse coding were also proposed [14]. These researches all resulted in significant improvement. Lately, even more powerful spectro-temporal features with increased feature dimensions were shown to improve further [15, 16].

Previously, we parameterized the spectro-temporal modulation information following Zhao and Morgan's 2-D Gabor filter approach [12] and integrated it with MFCCs at the phonetic posterior level [17]. We analyzed the complementarity between the two posteriors and demonstrated the improvement from integrating the two. Here, we integrate spectro-temporal features with pitch and cepstral features, and use tonal phonemes (or tonemes) for MLP posterior estimation and tonal acoustic units for HMM recognition in Tandem system. Because each tone varies with different sloped pattern in spectrogram, the spectro-temporal features offer additional information. Thus, improved performance in task with Mandarin Chinese was achieved. We also investigate the benefit respectively offered by the three types of features (i.e. spectro-temporal, pitch, and cepstral features) when used with tonal acoustic units.

2. Proposed approach

Here we introduce the three types of features and the framework of using tonal acoustic units in our Tandem recognition systems.

2.1. MFCC features

The cepstral features we use here are MFCCs obtained with a 25ms window and a 10ms shift. The 39-dimension feature vectors include c0 to c12 plus derivatives. Each vector is then concatenated with its previous and following four feature vectors as the input for MLP.

2.2. Pitch features

Smoothed log-pitch estimate plus its first and second derivatives are extracted [18]. Nine neighboring frames are concatenated for a 27-dimension vector as the input of MLP (referred to as pitch features F_0 below).

2.3. Spectro-temporal features

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

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

where σ_t, σ_f, ω_t, and ω_f are the four parameters that decide the shape of each filter. Per previous work [17], we selected the parameter set to divide the temporal modulation frequency from 1 to 16 Hz and the spectral modulation frequency from zero to two cycles per octave equally on a logarithmic scale [12], which was found to most closely correspond to human speech recognition [19]. These Gabor filter parameters were further divided into four streams, each corresponding to one Gabor filter bank, to cover from low to high spectro-temporal modulation frequency bands [12, 17]. As shown in the upper part of Fig. 1, we convolve the log mel-spectrogram of speech signals with the four filter banks and thus obtain four streams of spectro-temporal features for MLP training.

2.4. Tandem systems

In the lower part of Fig. 1, we show how to estimate the posteriors for the tonemes using features mentioned above with the aid of artificial neural networks (ANN). Each stream of spectro-temporal features are respectively concatenated with pitch features F_0 mentioned above at the frame level. Each stream of the augmented features are fed into an MLP with the tonemes as its training target. The MLP output is a vector of posteriors; each element corresponds to the probability of a specific toneme given the input features for the present time frame. 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 F_0+Gabor posteriors and use them in the following system. We similarly augment the MFCC features with F_0 and train an MLP. The output vectors are denoted as F_0+MFCC posteriors. Due to the complementarity between Gabor and MFCC features investigated previously [17], we further merge the F_0+Gabor and F_0+MFCC posteriors by taking the geometric mean over the two, obtain an even better estimation for posteriors, and denote them as F_0+Gabor+MFCC posteriors. In addition, Gabor, MFCC and Gabor+MFCC (geometric-mean merging of the two) posteriors are also obtained where we trained the MLP without pitch feature F_0 for comparison.

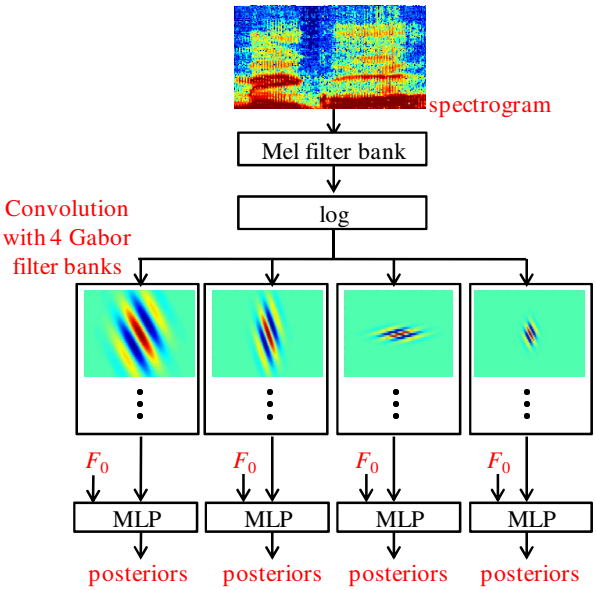

Figure 1: *The generation of F_0+Gabor posteriors.*

The posteriors are first transformed by a logarithm function which maps the range of the value between zero and one to negative real. 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 HMM system. As pointed out in the previous study [16, 17], we further perform the mean and variance normalization (MVN) on the after-LDA log-posteriors. Although the noise level is relative low in the corpus used here, the MVN step helps reducing the mismatch between training and testing data due to speaker variation and also results in distributions better modeled as Gaussians.

The resulting normalized posteriors are then concatenated with the MFCCs and used in HMM recognition. In this way, we have different Tandem systems which differ in the MLP input features. The units for MLP posterior estimation and HMM recognition are also varied (tonal or not) for analysis.

3. Experiments

3.1. Experiment setup

The experiments were conducted on the MATBN (Mandarin Across Taiwan-Broadcast News) corpus [17]. The training set includes 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 phonemes (monophone) and expanded each vowel to four tonal vowels while kept the consonant part unchanged. Thus, we obtained a 75 toneme set as the MLP training target. Only the four lexical tones are included in the toneme set with the neutral tone mapped to tone 3. Because there are few syllables with the neutral tone, adding it to the toneme set caused data-imbalance problem [7]. The 147 intra-syllable right-context-dependent Initial-Final (RCDIF) units were similarly expanded to 257 tonal RCDIFs including Initials and tonal Finals as the recognition target for HMM training. RCDIF is widely used for Mandarin speech recognition. The Initial is the initial consonant of a syllable, while the Final is the vowel (or diphthong) part plus an optional medial or nasal ending. Each RCDIF unit (tonal or not) was modeled as an HMM of three states. A trigram language model was used in decoding. The toneme-level training labels were based on forced alignment results.

3.2. LVCSR result

We list in Table 1 the character error rate (CER) for each LVCSR experiment with different configurations. Row (a) is the conventional HMM system with MFCC features and serves as the non-Tandem baseline. The rest are all Tandem systems and we show what posteriors are used in the first column. The second and third columns show the target of MLP (36 phonemes or 75 tonemes) and HMM (147 RCDIFs or 257 tonal RCDIFs) in each system as explained in detail in Sec. 3.1. Due to different cardinality in the sets of acoustic units used, the dimension of posterior vectors and the number of HMM models are varied among the experiments. Therefore, we change the mixture number of HMMs in each system to obtain similar number of parameters for reasonable comparison.

In rows (b) to (g), the MLP targets are the 36 phonemes disregarding the tone. (b) is the conventional MFCC Tandem system and (c) is the Gabor Tandem system [12]. The two systems achieved comparable results and integrating them in posterior level (row (d)) improved the performance. This is because the complementarity of the cepstral and spectro-temporal features as pointed out in previous works [17, 20].

Replacing HMM units with tonal RCDIFs improved consistently among the three posteriors; (e) and (f) yield statistically significant improvements over (b) and (c); (g) achieve statistically significant improvements over (b) to (d). Here, the confidence level of significant test is set to 95%. This fact suggests that the different tones affect the MFCCs and posteriors, although the MFCCs smooth out much signal level fluctuation and the posteriors are obtained mainly to discriminate among phonemes, not tones. Using tonal RCDIFs allows HMMs to learn the variation of MFCCs and posteriors across tones and benefits the performance.

In rows (h) to (j), we replace the MLP target with tonemes. Gabor toneme posteriors, (i), statistically significantly achieved better performance than the phoneme counterpart, (f), but the MFCC toneme posteriors, (h), did not. Combining the two posteriors, (j), yielded statistically significant improvement over (g). Because training MLP with phoneme targets is mainly aimed to discriminate between phonemes and some tonal clues may be lost in posteriors, Gabor toneme posteriors retain more information to model the tonal targets in HMM and provide improvement. However, owing to that every posterior set is concatenated with MFCCs as the observations of HMM, the tonal clues in cepstral features are already retained and thus the MFCC toneme posteriors offer little additional tonal information as compared to (e). Besides, comparing (i) and (j) to (f) and (g), we see more improvement (1.8% to 2.1%) than (f) and (g) to (c) and (d) (0.8% to 1.0%). Because Gabor features extract much more modulation information than cepstral features, they are more suitable for modeling tonal acoustic units than MFCCs or phoneme posteriors; the latter focus on discriminating between phonemes and retain little tonal variation. Hence, changing MLP targets to tonemes helps learning more tonal information at the front-end ((i) and (j)). Without such information, what the tonal HMM units can do in discriminating the tones is limited. However, the improvement from the tonemes for MLP and tonal RCDIFs for HMM units are additive.

Finally, we augment the MLP input features with F_0 and report the result in (k) to (m). Adding the pitch features offered consistent improvement over (h) to (j); the improvements from (h) to (k) and (j) to (m) are statistically significant while (i) to (l) are not. Additional to the information from mel-spectrogram where the MFCC and Gabor features are extracted, the F_0 provide clues directly from speech signals and they are beneficial. In our experiments, the system integrating pitch, cepstral and spectro-temporal features and modeling tone and phoneme variation simultaneously (tonemes for MLP and tonal RCDIFs for HMM), (m), achieved the best performance, 4.1% absolute or 19.3% relative CER improvement compared to the conventional Tandem system baseline, (b).

3.3. Analysis

Here, we investigate the frame-level tone and phoneme classification accuracy for the posteriors with different MLP input features. Firstly, we show the frame accuracy of the four tones and their total in Fig. 2. The six bars in each tone type correspond to the accuracy for six toneme posterior sets (used in (h) to (m) in Sec. 3.2): MFCC, Gabor, Gabor+MFCC, F_0+MFCC, F_0+Gabor, and F_0+Gabor+MFCC. To calculate the frame-level tone accuracy, we excluded consonants and summed the posterior probabilities of all tonemes having the same tone but varied phoneme parts, frame by frame. Thus, only four probabilities were obtained for the four tones although there are 75 tonemes. We then classified the tone following the maximum a posteriori (MAP) criterion on the tone posteriors.

Table 1. *CER for each recognition system with varied features and training targets. 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	HMM unit	CER(%)
(a) MFCC (non-Tandem)	×	147	24.6
(b) MFCC	36	147	21.2
(c) Gabor	36	147	21.3
(d) Gabor+MFCC	36	147	20.4
(e) MFCC	36	257	20.3
(f) Gabor	36	257	20.3
(g) Gabor+MFCC	36	257	19.6
(h) MFCC	75	257	20.4
(i) Gabor	75	257	18.2
(j) Gabor+MFCC	75	257	17.8
(k) F_0+MFCC	75	257	19.5
(l) F_0+Gabor	75	257	17.9
(m) F_0+Gabor+MFCC	75	257	17.1

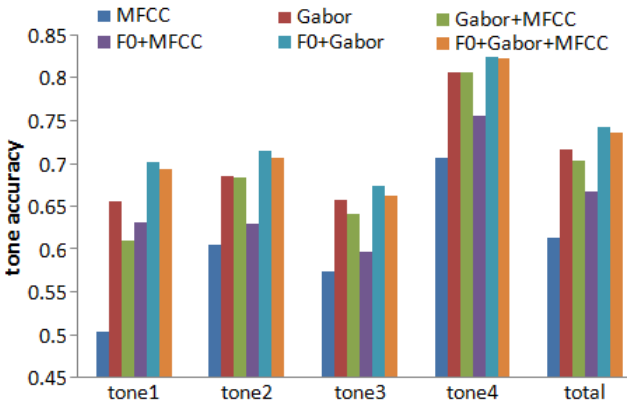

Figure 2: *The frame-level tone accuracy of different features used with MLP for tonemes*

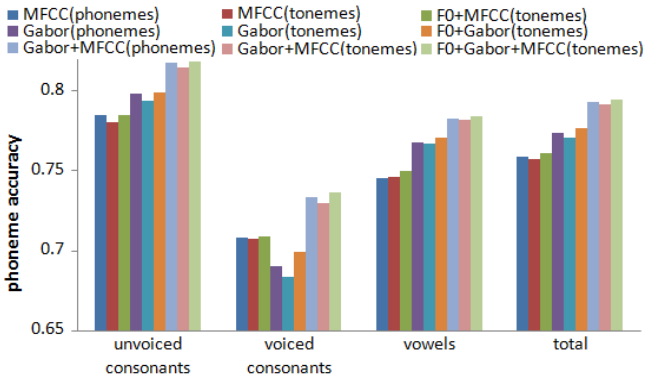

Figure 3: *The frame-level phoneme accuracy for different MLP targets and input features.*

There is a common trend for all the four tones. In the first three bars, the Gabor features resulted in the best accuracy while MFCC the worst. Because the 2-D Gabor filters have different sloped pattern in spectrogram according to their parameters and respond to different spectro-temporal modulation, the Gabor features are more sensitive than cepstral features to frequency component changing across time and frequency band. Therefore, Gabor features retain more information needed for tone classification and improve accuracy. Owing to that the performance of MFCC features was much worse than

Gabor, and combining the two sets of posteriors slightly deteriorated the performance. In the last three bars, we additionally appended F_0 to features in the first three. With the additional pitch features, the accuracy is consistently improved in every tone and every feature set. As mentioned above, even though Gabor filters parameterize considerable spectro-temporal modulation components, the mel-spectrogram, where the Gabor features are extracted, still somewhat smooth the spectrogram. The F_0 features provide additional information from the speech signal directly. From the above two facts, Gabor features outperform MFCC in tone classification, and augmenting them with F_0 (5th bar) further achieved the best performance.

In Fig. 3, we show the frame-level phoneme accuracy of different phoneme types (unvoiced consonants, voiced consonants and vowels) and their total. The first three bars correspond to MFCC features (posteriors used in (e), (h), and (k)), the next three to Gabor (used in (f), (i), and (l)), and the last three to MFCC+Gabor (used in (g), (j), and (m)). In each group of the three bars, the first bar is with MLP for 36 phonemes, the other two for 75 tonemes, and the third with F_0 used in addition. To calculate the phoneme accuracy, for each frame, we summed the posterior probabilities of all tonemes with the same phoneme but varied tone. Thus, the frame-wise phoneme posteriors are obtained for MAP classification.

We see in Fig. 3 that for unvoiced consonants, vowels, and the total, Gabor posteriors (the middle three bars) outperform MFCC (the first three), and their integration (the last three) improve further; for voiced consonants, MFCC are superior to Gabor, and their integration still perform even better. Consistent to previous results [17], there is complementarity among Gabor and MFCC features in phoneme recognition, and integrating those offers improvement. In addition, comparing the 1st, 4th, and 7th bars to 2nd, 5th, and 8th respectively, we see the phoneme targets result in higher phoneme accuracy than tonal targets. Because in the former, the MLP was aimed at discriminating among the phonemes while the MLP was also affected by tones in the latter. Although the former offers better phoneme accuracy, it leaves less tone information for the following HMM and thus causes higher overall CER ((e) to (g) vs (h) to (j)). Interestingly, the 3rd, 6th, and 9th bars (with F_0) outperform 2nd, 5th, and 8th (without F_0) correspondingly. This indicates the pitch features helps discriminating not only the tones, which is well known, but the phonemes as well.

Table 2 list the toneme accuracy of different features used with MLP for tonemes. It simultaneously considers the effect of features on both tone and phoneme accuracy. The six cases correspond to the six bars in Fig 2 and the 2nd, 3rd, 5th, 6th 8th, and 9th bars in Fig. 3. There is highly correlation between toneme accuracy in Table 2 and CER ((h) to (m)) in Table 1.

In summary, both Gabor and F_0 features help in tone classification. To classify the phonemes, all the three features contain complementary information, and integrating them improves the performance. Although phoneme MLP targets achieve higher phoneme accuracy than toneme MLP targets, the phoneme posteriors hold less tonal clues and result in higher CER in LVCSR. The F_0+Gabor+MFCC toneme posteriors, giving the best phoneme classification with slight performance degradation in tone, offer the greatest improvement in both toneme accuracy and CER.

Table 2. *Frame-level toneme accuracy (%) of different features used with MLP for tonemes*

Features	Acc.	Features	Acc.
MFCC	62.9	F_0+MFCC	65.1
Gabor	68.0	F_0+Gabor	69.4
Gabor+MFCC	69.3	F_0+Gabor+MFCC	70.7

4. Conclusion

We utilize MLP for modeling tones and phonemes with MFCC, Gabor, and pitch features. The resulting posteriors were integrated in a Tandem system and yielded a 19.3% relative CER reduction over an MFCC Tandem baseline. Gabor and pitch features are shown useful for tone classification. In classifying phonemes, MFCC and Gabor features contribute differently and depend on phoneme types, while adding pitch features provides consistent improvement. These complementarities among the three types of features originate naturally from the varied way these features are extracted, and explain the CER reduction achieved by combinatory posteriors in our experiments in Mandarin Chinese.

5. References

[1] T. Lee, W. Lau, Y.-W. Wong and P.-C. Ching, "Using tone information in Cantonese continuous speech recognition," *ACM Trans. Asian Language Info. Process*, vol. 1, pp. 83–102, 2002.

[2] H.-L. Wang, Y. Qian, F. K. Soong, J.-L. Zhou and J.-Q. Han, "A Milti-Space Distribution (MSD) Approach to Speech Recognition of Tonal Languages," in *Proc. Interspeech*, 2006.

[3] L.-W. Cheng, and L.-S. Lee, "Improved Large Vocabulary Mandarin Speech Recognition by Selectively Using Tone Information with a Two-stage Prosodic Model," in *Proc. Interspeech* 2007.

[4] H.-C. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proc. ICASSP*, 2000.

[5] E. Chang, J.-L. Zhou, S. Di, C. Huang and K.-F. Lee, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," in *Proc. ICSLP*, 2000.

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

[7] M.-Y. Hwang, G. Peng, W. Wang, A. Faria, A. Heidel and M. Ostendorf, "Building a Highly Accurate Mandarin Speech Recognizer," in *Proc. ASRU*, 2007.

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

[9] F. Valente, M. Magimai.-Doss, C. Plahl, S. Ravuri and W. Wang, "A comparative large scale study of MLP features for Mandarin ASR," in *Proc. Interspeech*, 2010.

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

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

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

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

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

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

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

[17] S.-W. Li, L.-C. Sun and L.-S. Lee., "Multi-stream spectro-temporal and cepstral features based on data-driven hierarchical phoneme clusters", in *Proc. ICASSP*, 2011

[18] X. Lei, M.-H. Siu, M.-Y. Hwang, M. Ostendorf and T. Lee, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," in *Proc. Interspeech*, 2006.

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

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