INTERSPEECH 2009 BRIGHTON

Exploration of Vocal Excitation Modulation Features for Speaker Recognition

Ning Wang, P. C. Ching, and Tan Lee

Department of Electronic Engineering,
The Chinese University of Hong Kong, Hong Kong, China
{nwang, pcching, tanlee}@ee.cuhk.edu.hk

Abstract

To derive spectro-temporal vocal source features complementary to the conventional spectral-based vocal tract features in improving the performance and reliability of a speaker recognition system, the excitation related modulation properties are studied. Through multi-band demodulation method, source-related amplitude and phase quantities are parameterized into feature vectors. Evaluation of the proposed features is carried out first through a set of designed experiments on artificially generated inputs, and then by simulations on speech database. It is observed via the designed experiments that the proposed features are capable of capturing the vocal differences in terms of F0 variation, pitch epoch shape, and relevant excitation details between epochs. In the real task simulations, by combination with the standard spectral features, both the amplitude and the phase-related features are shown to evidently reduce the identification error rate and equal error rate in the context of the Gaussian mixture model-based speaker recognition system.
Index Terms: speaker recognition, vocal source features, amplitude and frequency modulation

1. Introduction

The speech production mechanism has long been intensively explored for the steady and reliable indicators of speech units and speaker identities. Speech spectrum and the parametric model of the Linear Prediction Coding (LPC) have been successfully employed as the feature resources for speech-related statistical pattern recognition applications. Despite good performance has been achieved using standard cepstral features, such as Mel-frequency cepstral coefficients (MFCC), it is found to have bias towards the content of the speech unit [1] and is sensitive to the environmental variations [2], [3]. This motivates the exploration of more robust features in the development of a reliable speaker recognition system. Previous studies have shown that LP residual signals contain useful information about the excitation source and can be exploited for speaker recognition applications [4], [5], [6].

The vocal tract system is known to be a modulation system, where the individual formant modulates the excitation source, and produce a voiced sound. This property of speech has been used to synthesize and code speech data [7]. Features based on the frequency modulation (FM) of speech have also been investigated by employing several methods. For example, the segmental average instantaneous frequencies of signals are employed in [8]. Method capturing the spectral centroids that depends on the FM of signals is proposed by Paliwal *et al.* [9]. Dimitriadis *et al.* in [10] employed the average of instantaneous frequencies weighted by amplitudes. An all-pole FM extraction approach is suggested in [11], which claims greater reliability in FM features estimation. Published results indicate that the phase information captured by the FM features can offer assistance to enhancing recognition performance when jointly used with the conventional amplitude-based features. However, most of these features focus on the temporal modulation properties of the formants of the vocal tract system, and exclude the excitation characteristics. The sinusoidal model [7], [12], [13], on the other hand, fits the speech waveform by composing a set of sinusoids which are harmonically related to the fundamental frequency (F0) of the speech signal. This method is primarily employed for pitch tracking [12], and is used in excitation coding technique which is evaluated via the Minimum Mean Squared Error (MMSE) metric. These methods take advantage of the modulation properties of the excitation source but pay little attention to their speaker discrimination potentials. In this paper, we attempt to explore the speaker-relevant characteristics of the modulation phenomena in the excitation source of speech. Unlike the synthesis and coding systems, whose focus are speech intelligibility, waveform matching or transmission load, our method concentrates on characterizing the slow temporal (envelope and frequency) modulations in the LP residual signal, by using multi-band analysis and the nonlinear signal processing method. We propose a new set of modulation features, which are estimated from the multi-band AM-FM model of the residual signal. The parameters are noted as Averaged Instantaneous Envelope (AIE) and Averaged Instantaneous Frequency (AIF), respectively. In a multi-stream speaker recognition system, these features are used as the complementary speech features to MFCC.

This paper is organized as follows. In Section 2, we described the extraction procedures of the proposed feature AIE and AIF, as well as the designed experiments to evaluate these features. In Section 3, the experimental setup and analysis of the results on speaker identification and verification tasks are given. Section 4 concludes the paper.

2. Feature Extraction

In this section, we introduce the AM-FM speech model, and then present the derivation of the proposed features. Thereafter, we evaluate the proposed features in terms of several typical speech properties.

2.1. The AM-FM Speech Model

An AM-FM signal is described by (1) [11],

$$p(n) = a(n)cos(\frac{2\pi}{f_s}[f_c n + \sum_{r=1}^{n} q(r)] + \theta), \qquad (1)$$

where f_c is the center frequency value, f_s is the sampling frequency, $q(n)$ is the frequency modulating (FM) signal, and $a(n)$ is the time-varying amplitude. The instantaneous phase

Copyright © 2009 ISCA

$\phi(n)$ is $f_c n + \sum_{r=1}^{n} q(r)$, with a backward difference between $\phi(n)$ and $\phi(n-1)$, the instantaneous frequency sequence is defined as $f(n) = f_c + q(n)$. In multi-band demodulation analysis, a speech signal $s(n)$ can be modelled as the summation of N such AM-FM signals [11], one from each frequency band, that is,

$$\begin{aligned} s(n) &= \sum_{k=1}^{N} p_k(n) \\ &= \sum_{k=1}^{N} a_k(n) cos(\frac{2\pi}{f_s}[f_c^k n + \sum_{r=1}^{n} q_k(r)]), \end{aligned} \quad (2)$$

where k is the frequency band index, and the instantaneous envelope and frequency sequences are noted as $a_k(n)$ and $f_k(n)$, respectively. $f_k(n)$ is related with the FM component $q_k(n)$ by $f_k(n) = f_c^k + q_k(n)$.

2.2. Estimation of AIE and AIF Parameters

We employ the multi-band AM-FM model on the excitation signal to extract the AIE and AIF feature vectors. The process of computing the AIE and AIF features is summarized as follows:

1. Voicing decision: The AIE and AIF features are extracted from voiced speech only. The voicing status is detected using Talkin's Robust Algorithm for Pitch Tracking [14].
2. Computing linear predictive (LP) residual signal. In this paper, we use the LP residual signal as the representative of the excitation source. Each voiced segment is divided into overlapping frames with 30 msec duration and 10 msec frame shift. The residual signal $e(n)$ is obtained for each frame by taking LP inverse filtering. To diminish intra-speaker variation, the amplitude of the residual segment is normalized to the range of $[-1, 1]$.
3. Applying Gammatone filterbank to decompose the residual signal. The Gammatone filterbank [15] models the cochlea through a bank of overlapping bandpass filters whose center frequencies are equally spaced on the ERB scale. A Gammatone filterbank which contains N filters is applied on $e(n)$ to produce the subband signals. We set N to be 20, and the center frequency $f_c(k)$ ranges from 4KHz to 80Hz with a k increase from 1 to 20.
4. Estimating instantaneous envelope and instantaneous frequency. Teager's energy separation algorithm [16] is employed in obtaining the instantaneous envelope (IE) sequence $|a(n)|$ and the instantaneous angular frequency $\frac{2\pi}{f_s} f(n)$ (IF) on a frame basis for each subband signal.
5. Smoothing of the IE and IF sequence. A 21-point median filter is applied to remove the abrupt impulses in the frame IE and IF sequence.
6. Frame averaging of the smoothed IE and IF. An averaging operation is done on the smoothed IE and IF sequences for the frames in each subband. In this step, we remove the fluctuations of the IE and IF sequences, and track the amplitude and frequency of the most significant frequency components in each subband frame by frame.

At the end of the above procedures, the N dimensional feature vectors AIE and AIF are derived. Given one particular frame, the AIE and AIF vectors are viewed as the amplitude and frequency distributions of the principal components over the frequency bands. The AIE and AIF feature vectors are described as

$$AIE = \{AIE(1), AIE(2), \ldots, AIE(N)\},$$

and

$$AIF = \{AIF(1), AIF(2), \ldots, AIF(N)\}.$$

The speech data used in the experiments of this paper are sampled at 8 KHz, and thus the speech has the highest frequency of 4 KHz. Since there is no strict derivation to determine an optimal subband number N, and the frequency resolution by the 20-channel Gammatone filterbank can separate the harmonically related frequency components for the data we used, we consider not to use a larger N here.

2.3. AIE and AIF Parameter Analysis

The AIE and AIF features derived above are considered to have captured the amplitude and frequency characteristics of the principal components of the different subbands. In this part, we would like to evaluate these parameters by examining some typical speech properties. These properties cover the variation in F0, the difference in pitch epoch shape, and the relevant excitation details existing between the adjacent pitch epochs. With this logic, we have designed experiments specifically to evaluate the features in the following two aspects:

- Fundamental frequency (F0) value

The fundamental frequency (F0) is one of the most primary properties in distinguishing different speakers by human ears. B. S. Atal in [17] used pitch contour to identify speakers, and there have been pitch-related features proposed throughout the years for similar purposes. In this experiment, we generate two impulse trains to approximate the pitch-periodicity of the excitation signal. By taking these different pitched signals as input, we intend to see whether the AIE and AIF feature vectors can produce discriminative information.

In Figure 1, there are two impulse trains which are denoted as $e_1(n)$ and $e_2(n)$, respectively, they are both sampled at 8 KHz. Their waveforms are shown on the top row. $e_2(n)$ has chosen F0 to be 172Hz, which is the same as $f_c(18)$ (center frequency of the 18th filter), it is equivalent to 2.1 times of $f_c(15)$ and 3.2 times of $f_c(13)$. $f_c(k)$ decreases as k increases. The F0 of signal $e_1(n)$ is set to be half of $e_2(n)$. On the middle row, the AIE of $e_1(n)$ and $e_2(n)$ are illustrated. It is observed that in the lower frequency region, for $e_2(n)$, the $k = 18, 15, 13$ frequency bands have small peaks for AIE values, and there are no peaks for $e_1(n)$. In the high frequency region, most k in $e_2(n)$ give much higher AIE values than $e_1(n)$, and the AIE values of $e_1(n)$ are more evenly distributed among different k's. On the bottom row, the AIF values are revealed to have different distributions over k for $e_1(n)$ and $e_2(n)$, while the higher frequency bands tend to have a larger AIF value for $e_2(n)$.

- Pitch epoch shape and the details between epochs

Aside from the pitch period variation among speakers, it is believed that the shape of the pitch epoch and the details between adjacent epochs also play roles in discriminating different speakers. In this experiment, these properties will be focused on separately.

In Figure 2, in addition to signal $e_2(n)$, two other signals $e_3(n)$ and $e_4(n)$ and their AIE vectors are illustrated in the three rows, respectively, where all three share the same F0=172

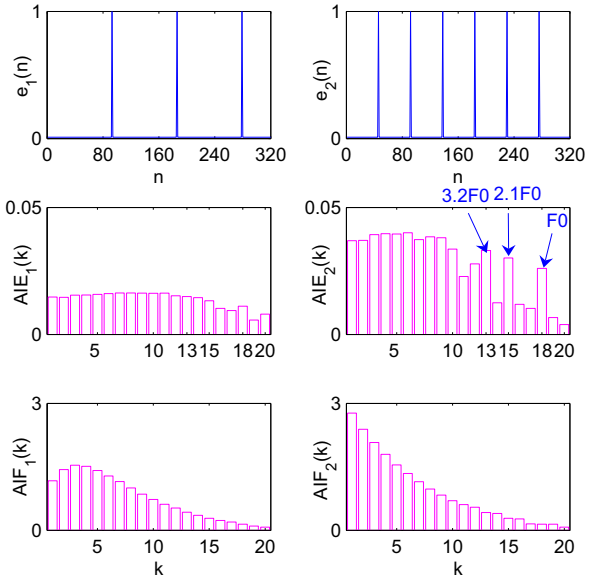

Figure 1: *AIE and AIF for artificial excitation signals with different F0.*

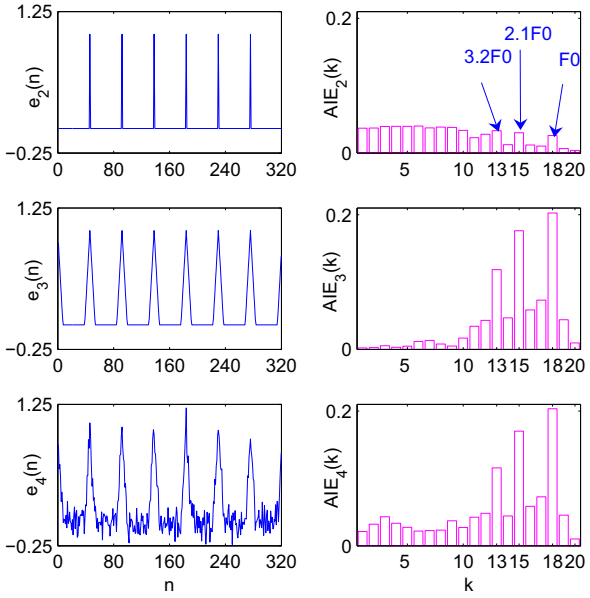

Figure 2: *AIE and AIF for artificial excitation signals: (1). with different epoch shapes: $e_2(n)$ and $e_3(n)$; (2). with and without details between adjacent epochs: $e_3(n)$ and $e_4(n)$.*

Hz. For the comparison between $e_2(n)$ and $e_3(n)$, it is revealed that although they have the same F0 value, their difference in pulse shape makes their AIE vectors differ a lot. We observe that the lower frequency amplitude is emphasized in $e_3(n)$, and the peaks at $k = 18, 15, 13$ bands are also more prominent for $e_3(n)$ than that for $e_2(n)$. The comparison between $e_3(n)$ and $e_4(n)$ focuses on the effects of the embedded details among the epochs. It is seen that the peaks at $k = 18, 15, 13$ appear for $e_4(n)$ as well, and the AIE values for the higher frequency regions are also enlarged.

Indicated by the above experiments and compared with the conventional spectral analysis results, we can see that the AIE vector can reveal the amplitude modulation information in different frequency bands which is absent from the flat spectrum of the excitation signal. On the other hand, AIF provides phase-related information for vocal excitation.

3. Speaker Recognition Experiments

In this section, we will evaluate the proposed features AIE and AIF as the complementary parameters to the MFCC features through speaker identification (SID) and speaker verification (SV) experiments on a speech database CU2C. The standard MFCC features we used contain 39 components: the log energy, 12 static coefficients, and their dynamic and acceleration coefficients.

3.1. Database: CU2C

CU2C is a Cantonese speech database developed for speaker recognition research at the Chinese University of Hong Kong [18]. We use a part of CU2C, which contains the speech data from 50 male speakers. Each speaker has 12 sessions of recordings which were collected over a time span of 4-9 months. There are 6 utterances in each session. Each of them contains a sequence of 14 randomly generated digits. The utterance length is about 5 to 6 seconds. The utterances were recorded via a head-worn microphone in a quiet room. The original sampling frequency was 16 KHz. The speech data we used in this paper were down-sampled to 8 KHz.

3.2. Experimental Setup

For all speakers, 6 out of the 12 sessions are used to train the speaker models, and the remaining data are used for performance evaluation. The standard approach for UBM-GMM training is adopted. Three separate systems are built based on MFCC, AIE, and AIF, respectively. The features AIE and AIF are used as complementary counterparts of MFCC, respectively, in the two sets of experiments. The score-level fusion technique is used to combine the contributions of the two systems by MFCC and AIE in one set, and by MFCC and AIF in another set. The final decision is determined by the overall combined score. In the identification tasks, the log-likelihood score of each test is a linear combination of the log-likelihood scores from the MFCC and AIE/AIF features, with weighting parameters w_M and w_A (i.e., $L = w_M L_M + w_A L_A$). Meanwhile, in the verification tasks, the fusion is performed on the log-likelihood ratio scores, that is, $\lambda = w_M \lambda_M + w_A \lambda_A$. In both tasks, w_M and w_A are related by $w_M + w_A = 1$. The weighing strategy is described as follows: initially, let $w_M = 0$, and $w_A = 1$. Next, we empirically increase w_M by a step of $\frac{1}{50}$, and repeat this for 50 times, with $w_A = 1 - w_M$ satisfied. Finally, we identify the optimum parameter set $[w_M, w_A]$ with the best recognition results.

For SID and SV tasks, the identification error rate (IDER) and equal error rate (EER) are used as the primary performance indicators, respectively.

3.3. SID and SV Evaluation Results

In Table 1, the SID and SV results are listed separately in two columns. Different sets of feature configurations are indicated in the five rows. There are three experimental tasks that employ individual features: MFCC, AIE, and AIF, and two tasks giving the fused results from MFCC and AIE, or MFCC and AIF.

Table 1: *Speaker recognition results: IDER and EER (in %).*

feature configuration	IDER	EER
MFCC	0.9	1.5
AIE	33.9	11.9
AIF	28.3	10.1
MFCC+AIE	0.6	**1.2**
MFCC+AIF	**0.4**	1.3

3.4. Results Analysis

3.4.1. Recognition accuracy analysis

From the SID and SV results shown in Table 1, the performance of individual features AIE and AIF are far from comparable with those of MFCC features. However, both the combined results of MFCC with AIE and MFCC with AIF produce considerable improvements over the individual MFCC performance. As indicated by the best SID and SV results achieved here, IDER and EER showed improvements of 55.6% and 20.0%, respectively, over the MFCC results. Considering that the dimension of the excitation-related features AIE and AIF are both 20, nearly half of that of MFCC, we believe that the AIE and AIF features represent a promising direction in helping enhance the reliability of the MFCC-based speaker recognition system.

3.4.2. Feature complementarity analysis

Concerning the combined SID and SV results recorded in Table 1, the averaged weighting factor ratios are $w_M : w_A = 48 : 52$ for the SID experiment, and $w_M : w_A = 57 : 43$ for the SV experiment, respectively. The 0.4% IDER and 1.2% EER are achieved with w_A= 0.46 and 0.48, respectively. This reveals the high complementary relation between the MFCC and AIE/AIF features in representing the acoustic characteristics of an individual speaker.

4. Conclusion

This work aims to explore the vocal excitation source properties in terms of the temporal modulations in both amplitude and frequency. Under the framework of multi-band demodulation, the excitation-related amplitude and frequency components are first separated over the time, their distribution across the multiple bands are then parameterized into feature vectors. These spectro-temporal parameters are evaluated through analytical comparisons and simulation results, and we found that (1) multi-band amplitude and frequency modulation parameters are capable to capture the time-frequency vocal excitation characteristics; (2) modulation-related source parameters are complementary with the relevant vocal tract features; and (3) speaker recognition accuracy provided by the spectral-based features can be further improved by combining the proposed source features in a multi-stream recognition system.

5. References

[1] W. N. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 15, no. 6, pp. 1884–1892, Aug. 2007.

[2] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, pp. 261–291, 1995.

[3] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, pp. 275–294, 1991.

[4] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Commun.*, vol. 17, no. 1-2, pp. 145–157, 1995.

[5] S. R. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction of speech," *Speech Commun.*, vol. 48, pp. 1243–1261, 2006.

[6] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 181–184, Mar. 2007.

[7] R. J. Mcaulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[8] Y. Wang, J. Hansen, G. K. Allu, and R. Kumaresan, "Average instantaneous frequency (AIF) and average log-envelopes (ALE) for ASR with the Aurora 2 database," in *Proceedings of EUROSPEECH*, 2003, pp. 25–28.

[9] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258–261, Feb. 2004.

[10] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letter*, vol. 12, no. 9, pp. 621–624, Sept. 2005.

[11] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using all-pole model," *Electronics Letters*, vol. 44, no. 6, March 2008.

[12] R. J. Mcaulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proceedings of IEEE*, 1990, pp. 249–252.

[13] Y. Stylianou, "Decomposition of speech signals into a deterministic and a stochastic part," in *Proceedings of ICSLP'96*, vol. 2, 1996.

[14] D. Talkin, *Speech coding and synthesis*. Elsevier Science, 1995, ch. A robust algorithm for pitch tracking (RAPT), pp. 495–518.

[15] M. Cooke, *Modeling auditory processing and organization*. Cambriage, UK: Cambriage University Press, 1993.

[16] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transaction on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[17] B. S. Atal, "Automatic speaker recognition based on pitch contours," *Journal of the Acoustical Society of America*, vol. 52, no. 6, pp. 1687–1697, 1972.

[18] N. Zheng, C. Qin, T. Lee, and P. C. Ching, "CU2C: A dual-condition cantonese speech database for speaker recognition applications," in *Proc. Oriental-COCOSDA*, 2005, pp. 67–72.