USING HAAR TRANSFORMED VOCAL SOURCE INFORMATION FOR AUTOMATIC SPEAKER RECOGNITION

Nengheng Zheng and P.C. Ching

Department of Electronic Engineering, the Chinese University of Hong Kong,
Shatin, N.T., Hong Kong, PRC.
Email: {nhzheng, pcching}@ee.cuhk.edu.hk

ABSTRACT

This paper attempts to investigate the effectiveness of incorporating vocal source information for enhancing automatic speaker recognition accuracy. We propose a new method to extract discriminative features from the linear prediction (LP) residual signal, which are closely related to the glottal excitation of individual speaker. A complementary parameter set in addition to the commonly used linear predictive cepstral coefficients (LPCC), called Haar Octave Coefficients of Residue (HOCOR), is obtained by applying Haar transform to the LP residue. This additional feature vector retains the spectro-temporal characteristics of the source excitation sequences that are related to the fundamental frequency, harmonics as well as their phases. Experimental evaluation over the YOHO corpus demonstrates the high speaker discriminative power and high inter-speaker variability of HOCOR. Speaker recognition tests with both vocal tract feature (LPCC) and vocal source information (HOCOR) outperform the conventional methods of using LPCC only.

1. INTRODUCTION

Automatic Speaker Recognition (ASR) is a biometric identification process, in which personal identity is recognized on the basis of speaker information obtained from speech. According to the speech production model, human speech is the output of the vocal tract system driven by the source excitation

$$s(n) = u(n) * h(n), \tag{1}$$

where $u(n)$ denotes the excitation source and $h(n)$ represents the impulse response of the vocal tract system [1]. State-of-the-art ASR systems typically extract features carrying vocal tract chara-cteristics, such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC). Recently, some experimental results have shown that features bearing vocal cord characteristics, such as pitch, harmonics, *etc*., can work as supplementary features to those vocal tract ones and can improve speaker recognition performance [2][3].

In addition to exploiting prosodic and harmonic features, researchers have also examined the usefulness of the LP residue for speaker recognition [4][5]. As shown in Figure 1, if the vocal tract transfer function of the speech production model can be characterized by the predictive coefficients, *i.e.* $\hat{a}_k \approx a_k$, the prediction error, referred as the LP residue, will approximate the excitation signal, $e(n) \approx u(n)$. Thus the LP residue can be used to derive the source information [1]. However, the spectral envelop of $e(n)$ is nearly flat, features extracted from the Fourier spectrum do not make much contribution [4]. In [5], auto-associative neural network was applied to obtain source information from the residue. This method is not widely used because of its expensive computational cost.

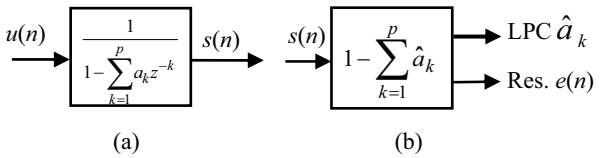

Figure 1. Synthesis (a) and analysis (b) models of LPC system

Plumpe *et al.* applied temporal modeling of the glottal flow derivative waveform to speaker identification [6]. The drawback of this method is the difficulty in locating the closed glottis interval which is crucial for estimating the glottal flow derivative.

In this article, we present a new technique to extract source information from the LP residue. Rather than taking Fourier transform, Haar transform [7] is applied to the residual signal, which is computationally simpler. More importantly, Haar transform is effective in detecting the bursts within the residue for voiced sounds [8]. The Haar spectrum essentially represents a time-frequency analysis of the residual signal. While it does not provide the true glottal flow, information related to the pitch and its harmonics, as well as the spectro-temporal features of the excitation source could be characterized. In order to reduce the feature dimension, the energy of each individual Haar octave (or the time-indexed sub-groups of each octave) is computed to form the so called Haar Octave Coefficients of Residue (HOCOR), and this is used as a feature complementary to the LPCC.

The remainder of this paper describes the proposed technique and experimental results that demonstrate the effectiveness of HOCOR for speaker recognition. In Section 2 we introduce the Haar transform and illustrate how we generate the HOCOR feature set. The experimental set up and recognition results are presented in Section 3. Finally we draw a conclusion of this work in Section 4.

This work was partially supported by a research grant awarded by the Hong Kong Research Grant Council.

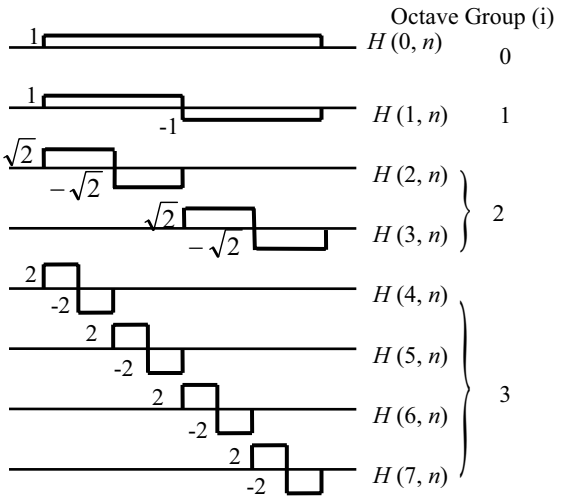

Figure 2. The first 4 octave groups of Haar Function

2. FEATURE EXTRACTION

2.1. Haar Function and Haar Transform

Discrete Haar transform of signal $x(n)$ and its inverse transform can be formulated as

$$\begin{cases} X(k) = \dfrac{1}{N}\sum_{n=0}^{N-1} x(n)H(k,n), k = 0,1,\cdots,N-1 \\ x(n) = \sum_{k=0}^{N-1} X(k)H(k,n), n = 0,1,\cdots N-1 \end{cases} \quad (2)$$

The Haar function $H(k, n)$ is a completely orthogonal function set of rectangular waveforms

$$H(0,n) = 1, \qquad 0 \le n \le N-1$$

$$H(k,n) = H(2^{i-1} + j - 1, n)$$

$$= \begin{cases} \sqrt{2^{i-1}}, & \dfrac{j-1}{2^{i-1}}N \le n < \dfrac{j-\frac{1}{2}}{2^{i-1}}N \\ -\sqrt{2^{i-1}}, & \dfrac{j-\frac{1}{2}}{2^{i-1}}N \le n < \dfrac{j}{2^{i-1}}N \\ 0, & elsewhere \end{cases}, \quad (3)$$

$$i = 1,2,\cdots \qquad j = 1,2,\cdots,2^{i-1}$$

where ***i*** denotes an octave subset having a zero-crossing in a given width $N/2^{i-1}$, and ***j*** gives the position of the function within this subset. Spectral decomposition with such a rectangular base function is more appropriate than that with sinusoidal function due to the noise like, burst mode changing characteristics of the residue [7]. Figure 2 shows the first 4 groups of Haar Function.

If we define the *frequency* to be the number of zero-crossing within a time interval, then the Haar transform provides a kind of time-frequency analysis of the signal. We can define the Haar spectrum as

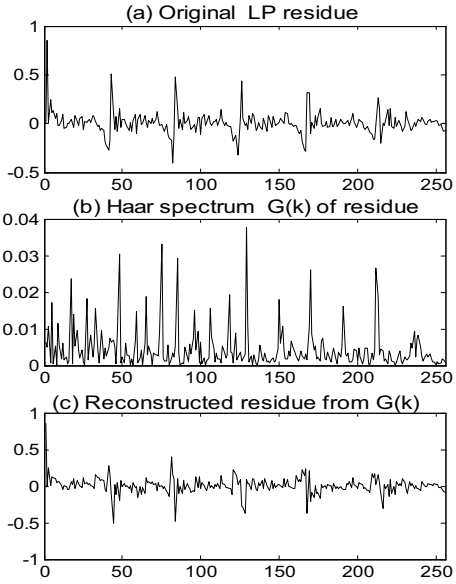

Figure 3. Haar spectrum of a length-256 LP residue. The 6 periodic peaks of residue (a) were represented by the peaks within different octave (b), *i.e.* 6 periodic peaks within the last octave. The reconstructed signal (c) well matches the original one in both the frequency and phase of the bursts.

$$G(k) = |X(k)|, \qquad k = 0,1,\cdots,N-1. \quad (4)$$

All $G(k)$'s within an octave group can be considered as the result of scanning the signal with a specific Haar function. Thus the Haar spectrum retains the spectro-temporal characteristics of the residue. Figure 3 shows the Haar spectrum (b) of a segment of the residue (a). The peaks of $G(k)$'s within an octave are position sensitive to the bursts within the signal. The reconstructed signal (c) from $G(k)$ keeps the burst properties of the original one, in both periodicity (frequency) and position (phase).

2.2. Generating HOCOR

To generate the Haar Octave Coefficients of Residue, HOCOR, we first group the $G(k)$'s within different octaves, *i.e.*

$$\begin{aligned} &\mathbf{H}_0 = \{G(0)\} \\ &\mathbf{H}_i = \{G(k) \mid k = 2^{i-1},\cdots,2^i - 1\}. \\ &i = 1,2,\cdots,\log_2 N \end{aligned} \quad (5)$$

Each octave essentially corresponds to a specific frequency decomposition of the signal. For example, for a length-256 signal segment with 8 kHz sampling frequency, the $\mathbf{H}_i$ correspond to the Fourier frequency components as follows,

$$\begin{aligned} &\mathbf{H}_0 \rightarrow D.C. \\ &\mathbf{H}_1 \rightarrow f = 32\,Hz \\ &\mathbf{H}_2 \rightarrow f = 64\,Hz \\ &\quad\vdots \\ &\mathbf{H}_8 \rightarrow f = 4\,kHz \end{aligned} \quad (6)$$

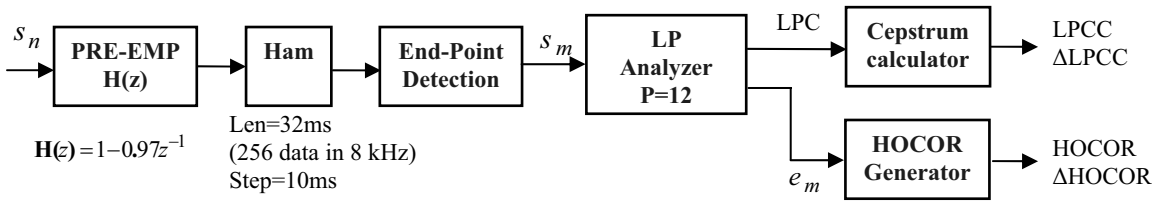

Figure 4. The front-end processing of feature extraction

Thus $\mathbf{H}_1$ will give frequency component at 32 Hz, while $\mathbf{H}_8$ shows the presence of frequency as high as 4 kHz. For speech processing, the fundamental frequency is seldom less than 64 Hz, and therefore the first three octaves can be ignored.

To compose a HOCOR feature vector, the simplest set we can derive is given by

$$\mathbf{HOCOR}_0 = \left\{ \sum_{\mathrm{G(k)} \in \mathrm{H_i}} G^2(k) \;\middle|\; i = 3,4\cdots, \mathbf{log}_2\, N \right\}. \qquad (7)$$

In this case, the feature vector has just 6 elements containing only information of the fundamental frequency and the harmonics, but not the temporal information since all $G(k)$'s within an octave are added together. To retain the temporal information, each octave can be equally divided into 2 sub-groups and then the energy of each sub-group is computed to generate a double sized HOCOR. For convenience, we call it the first-ordered HOCOR, noted as HOCOR_1. There are now 12 elements in the HOCOR_1 feature vector and contains approximate temporal information of the constituent frequency components. To extend further so as to obtain more detailed temporal information, each octave can be divided into 4, 8 and up to 2^{i-1} sub-groups, noted as

$$\mathbf{H}_i^{\alpha} = \left\{ \mathbf{H}_i^{\alpha}(j) \;\middle|\; j = 0,1,\cdots,2^{\alpha} - 1 \right\}, \quad (\alpha \le i - 1), \qquad (8)$$

where

$$\mathbf{H}_i^{\alpha}(j) = \left\{ G(k) \;\middle|\; \begin{array}{l} k = 2^{i-1} + j \cdot 2^{i-1-\alpha}, \cdots, \\ 2^{i-1} + (j+1) \cdot 2^{i-1-\alpha} - 1 \end{array} \right\}. \qquad (9)$$

And the αth-ordered HOCOR is given by

$$\mathbf{HOCOR}_{\alpha} = \left\{ \sum_{G(k) \in \mathbf{H}_i^{\hat{\alpha}}(j)} G^2(k) \;\middle|\; \begin{array}{l} i = 3, \cdots, \mathbf{log}_2 N \\ \hat{\alpha} = \mathbf{min}(i-1, \alpha) \\ j = 0, \cdots, 2^{\hat{\alpha}} - 1 \end{array} \right\} \qquad (10)$$

2.3 Properties of HOCOR

In summary, HOCOR bears the following properties.

- HOCOR is uncorrelated to the LPCC in a large extent, since the residue is theoretically orthogonal to the vocal tract system.
- The rectangular base function and the time-frequency properties of Haar transform result in better spectral decomposition of the noise like, burst mode changing residual signal.
- HOCOR_{α} with $\alpha > 0$ represents pitch and harmonics as well as their phase information of the vocal source, which will be useful for speaker recognition.
- Computational simplicity.

3. EXPERIMENTS

3.1 System Design

Experiments were conducted using the male subset of the YOHO corpus [9]. Both the identification and verification experiments were carried out. The speaker models were trained by 128 components GMM [10] with the first three enrollment sessions. The fourth enrollment session was used for background selection in verification tasks. Testing trials used all the 10 sessions with an utterance of about 2.5 seconds for each trial. For comparison, a baseline system with LPCC_D feature, which contains LPCC and its first order time difference, ΔLPCC, was implemented. The feature extraction procedure was shown in Figure 4. The speech samples were first pre-emphasized, and then weighted with Hamming window function. The silent segments were excluded by endpoint detection. After LP analysis, the LP residue was passed through the HOCOR generator to generate the HOCOR_{α} as described in section 2.2. And 12 dimensional LPCC can be obtained from the linear prediction process. Finally, the first order time difference elements, ΔLPCC and ΔHOCOR, were produced to include dynamic characteristics.

3.2 Experimental Results

For identification test, we measured the identification error rate (IDER). For verification, we measured the equal error rate (EER) and detection cost ($\mathrm{C_{Det}}$). Each of them represents a point on the detection error tradeoff (DET) plot. While the EER corresponds to the point where the false rejection and false acceptance error are the same, the $\mathrm{C_{Det}}$ measures the system performance with different weights for these two types of error [11],

$$\mathrm{C_{Det}} = C_{FR} \cdot P_{FR} \cdot P_{Target} + C_{FA} \cdot P_{FA} \cdot P_{NonTarget}. \qquad (11)$$

Here we set $C_{FR} = 10 \cdot C_{FA}$, which corresponds a point in the lower-right region that user convenience is required.

The contribution of temporal information for speaker recognition was demonstrated in Figure 5. As shown, the recognition performance was improved when more temporal information was incorporated, *i.e.* the IDER curve declined as α increased from 0 to 2, and then went up as α greater than 3. The EER and $\mathrm{C_{Det}}$ curves had the similar trend to the IDER. The degeneration of performance with larger α may be due to the correlation between sub-groups and the very large feature vector size in which the training data may be insufficient. In the following experiments, the 12 dimensional HOCOR_1 feature was used.

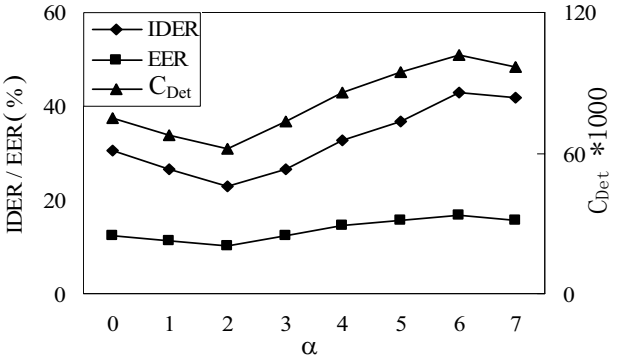

Figure 5. Recognition results with $HOCOR_\alpha$

Performance	LPCC_D (24 dim.)	$HOCOR_1$ (12 dim.)	$HOCOR_1$_D (24 dim.)
IDER (%)	1.51	26.63	17.64
EER (%)	1.04	11.14	8.74
$C_{Det} \times 1000^*$	5.05	67.84	52.78

Table 1. Recognition results for vocal tract and source features.

Performance	Feature Combination	Score Combination
IDER (%)	1.16	1.06
EER (%)	0.99	0.89
$C_{Det} \times 1000$	4.69	4.03

Table 2. Recognition results for fusing complementary features.

As evident from Table 1, $HOCOR_1$ resulted in 26.63%, 11.14% and 67.84 for IDER, EER, and C_{Det}, respectively. Also, the first order delta elements provided additional discriminative power which improved the performance by 34%, 22% and 22%, respectively. But the discriminative power of $HOCOR_1$ and $HOCOR_1$_D are not convincing compared to that of LPCC_D. However, we expected to gain some improvement by using $HOCOR_1$_D as complementary feature to LPCC_D. Since theoretically the residual signal and the vocal tract system should be uncorrelated, combination of source and vocal tract information should reduce recognition error.

To determine how and how much the fusing information can improve the performance, two fusing methods were used. One is called *Feature Fusion*, where the $HOCOR_1$_D was appended to the LPCC_D to form the 48 dimensional LPCC_D_$HOCOR_1$_D. The second is called *Score Fusion*, where the final score was the summation of scores calculated from LPCC_D and $HOCOR_1$_D,

$$S = \frac{S_l + S_h}{2}, \tag{12}$$

where the subscripts l and h correspond to LPCC_D and $HOCOR_1$_D, respectively. Table 2 lists the recognition results of the information fusion system. Obviously, when taking into account of both vocal tract and source excitation characteristics, the system performed better when compared with the baseline system where only vocal tract information was used. And using *Feature Fusion*, the relative reductions of IDER, EER and C_{Det} were about 23.2%, 4.8% and 7.1%, respectively. While with *Score Fusion*, the improvement was more significant. We obtained a relative error or cost reduction of 29.8%, 14.4% and 20.2%. The inferiority of *Feature Fusion* to *Score Fusion* may be partly due to the insufficient training data for the large sized feature vector.

[*] C_{Det} is multiplied by 1000 for convenient display

4. CONCLUSION

This paper presents a new technique to extract discriminative features from the source excitation. Instead of extracting information from the Fourier spectrum of the LP residue, our method applied Haar transform to the residual signal. The Haar spectrum provides a time-frequency analysis of the residue which retains the spectro-temporal characteristics of the excitation sequences. The feature set $HOCOR_\alpha$ with $\alpha > 0$ offers speaker-specific information that is related to the fundamental frequency, harmonics as well as their phases. The recognition tests showed the discriminative power of $HOCOR_\alpha$ for speaker recognition. Especially, when served in supplement to the vocal tract feature (LPCC_D), $HOCOR_1$_D relatively improved the recognition performance of 29.8%, 14.4% and 20.2% for IDER, EER, and C_{Det}, respectively.

5. REFERENCES

[1] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, Vol. 63, pp. 561-579, 1975.

[2] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling Dynamic Prosodic Variation for Speaker Verification," *ICSLP* 1998, Sydney, pp. 3189-3192.

[3] B. Imperl, Z. Kacic, and B. Horvat, "A Study of Harmonic Features for Speaker Recognition," *Speech Communication*, Vol. 22, pp. 385-402, 1997.

[4] P. Thevenaz and H. Hugli, "Usefulness of the LPC Residue in Text-independent Speaker Verification," *Speech Communication,* Vol. 17, pp. 145-157, 1995.

[5] B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and System Features for Speaker Recognition Using AANN Models," *ICASSP* 2001, Salt Lake City, pp. 409-413.

[6] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification," *IEEE Tans. Speech Audio Processing,* Vol. 7, No. 5, pp. 569-585, 1999.

[7] K. G. Beauchamp, *Walsh Functions and Their Applications,* Academic Press, London *etc.,* 1975.

[8] D. W. Thomas, "Burst Detection Using the Haar Spectrum," *Proc.: Theory and Application of Walsh and Other Non-sinusoidal Functions,* 1973.

[9] http://www.ldc.upenn.edu.

[10] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication,* Vol. 17, pp. 91-108, 1995.

[11] A. Martin and M. Przybocki, "The NIST 1999 Speaker Recognition Evaluation: An Overview," *Digital Signal Processing*, Vol. 10, pp. 1-18, 2000.