Time –Frequency Representation of Vocal Source Signal for Speaker Verification

Musala.Venkateswara Rao
Department of Electronics and computer Engineering
K.L.Univeristy A.P.,India

Kallakunta. Ravi Kumar
Department of Electronics and computer Engineering
K.L.Univeristy A.P.,India

Podila.Manoj
(Y8EM297)
Department of Electronics and computer Engineering
K.L.Univeristy A.P.,India

ABSTRACT
We propose an effective feature extraction technique for obtaining essential time-frequency information from the linear prediction (LP) residual signal, which are closely related to the glottal vibration of individual speaker. With pitch synchronous analysis, wavelet transform is applied to every two pitch cycles of the LP residual signal to generate a new feature vector, called Wavelet Based Feature extraction (WBFE), which provides additional speaker discriminative power to the commonly used linear predictive Cepstral coefficients (LPC).WBFE out performs the LPCs coefficients. In this paper, we have demonstrated the effectiveness of using vocal source features to supplement vocal tract features for improved speaker verification and the verification results are displayed in the MATLAB.

Keywords: LP residual, WBFE, LPC.

1. INTRODUCTION
Current speaker recognition systems are dominated by the vocal tract characteristics represented using spectral features such as MFCC and linear prediction coefficients (LPCC) derived through short-time spectral analysis. System based on spectral features perform well in acoustically matched and noise free conditions. However, they fail to model information about speaker that might contribute to speaker recognition. It has been shown that spectral features are affected by channel and noise. Therefore, researchers try additional features to capture the speaker-specific characteristics of excitation source [1], which is obtained by passing the speech signal through inverse filter designed with LPC coefficients. The main source of excitation for production of speech is the glottal vibration. In each glottal cycle, the instant of glottal closure is the instant at which significant of vocal tract takes place. Hence, the small region around the instant of glottal closure contains significant information about the speaker for developing speaker verification systems, which is focused in this paper.

According to the speech production model, human speech is the convolution output of the source excitation signal $u(n)$ and the impulse response of the vocal tract system $h(n)$ [2],

$$s(n) = u(n) * h(n) \quad (1)$$

The importance and applicability of vocal tract characteristics have been extensively acknowledged [2]. However, the usefulness of the vocal source excitation, as well as its effective retrieving technique as not been thoroughly studied. Many of the studies on vocal source information for speaker recognition have been focused on pitch and cepstral coefficients of the excitation signal (*i.e.*, the LP residual signal) [4][5][6]. More recently, explicit temporal modeling of the glottal flow derivative waveform has also been studied, and the importance of glottal flow wave shape of the voiced sound for speaker recognition was demonstrated [7]. However, none of these studies revealed the time-frequency properties of the source excitation signal, which may be more appropriate to characterize the relatively fast time-varying excitation signal and could be helpful for speaker recognition.

In this paper, instead of doing time-frequency analysis of the LP residual signal for each frame of a fixed duration, we adaptively select the analysis window length to be exactly equal to two pitch cycles by pitch synchronous analysis. Wavelet transforms with the Daubechies wavelet and 4 dyadic (or octave) dilated baby wavelets is applied to the pitch synchronized LP residual signal. The wavelet transform coefficients corresponding to each baby wavelet are grouped together to form an octave group. And the energy of each octave group (or the time-indexed subgroups of each octave) is computed to form the so called Wavelet Based Feature extraction of LP Residual (WBFE), while it does not provide the true glottal flow, information related to the frequency composition as well as the dynamic evolution within each pitch cycle can be characterized. The remainder of this paper will give a detailed explanation of proposed method and experimental results will be presented to demonstrate the effectiveness of WBFE for speaker verification. Most existing speech/speaker recognition systems adapt a frame-based feature extraction approach where vocal tract transfer function is assumed to be stationary with in voicing [9]. The dynamic evolution of glottal waveform is highly related to the speaker specific larynx characteristics, and thus should be useful for speaker recognition [8]. To exploit the time-frequency information of the source excitation, we present a wavelet transform based feature extraction technique, as described in Figure 1.

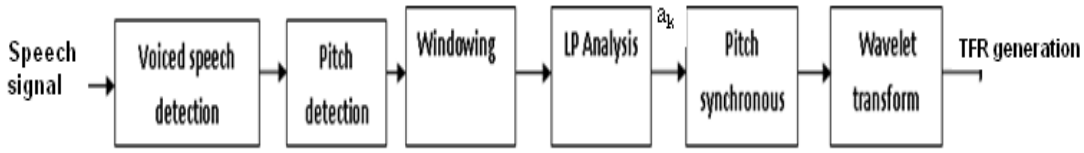


Fig 1. Feature extraction using Daubechies Wavelet Transform

2. TIME- FREQUENCY FEATURE EXTRACTION:

2.1 Voiced Speech Detection:

Only the voiced segment is concerned in our method since the unvoiced speech segment does not contain too much information of the vocal source production mechanism. A two step algorithm is used to determine the voiced sound. First, an energy detector is applied to remove the silence and most of the unvoiced portions that have considerably lower energy than the voiced segments. Then, the' filtered' speech signal is fed through a zero-crossing detector to eliminate the remaining unvoiced sounds.

2.2 Pitch Estimation:

Pitch is estimated using cepstrum analysis for every 32 ms of voiced speech [9]. The pitch period is used for window length selection and pitch pulses detection in the following steps.

2.3 Windowing:

The voiced speech is segmented with a rectangular window with variable length equals to 2.5times of the estimated pitch period.

2.4 LP Inverse Filtering:

The windowed speech frame is inverse filtered to generate the LP residual signal. The12th order LP coefficients are computed by autocorrelation method [1].

2.5 Pitch Synchronous Analysis:

The residual signal of the voiced speech essentially represents bursts at the vocal cord closing. If we define the period between two successive bursts to be a pitch cycle, then pitch synchronous analysis can be achieved by detecting these bursts.

2.6 Wavelet Transform of u(n):

With the above process, the analysis window for wavelet transform is now strictly constrained to be exactly 2 pitch cycles and 1pitch cycle overlap with every window starting at the excitation epoch. The wavelet transform of $u(n)$ can be expressed as

$$W_u(a,b) = \frac{1}{\sqrt{|a|}} \sum u(n) \psi^* \left(\frac{n-b}{a} \right) \quad (2)$$

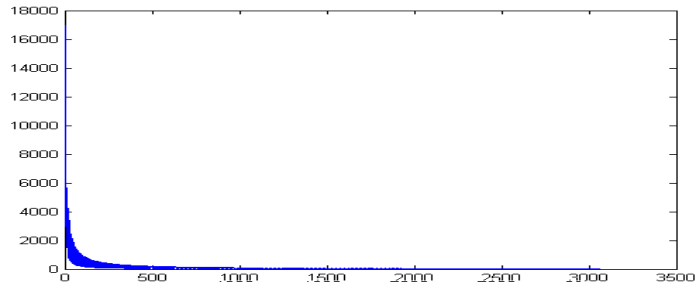
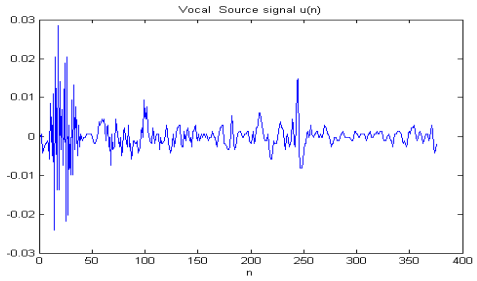
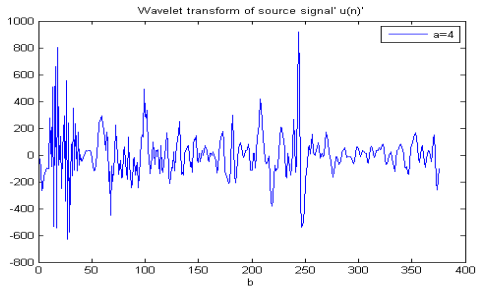


Fig 2a: Mother wavelet function Ψ(*n*)

Where **Ψ(*n*)** is the mother wavelet function is shown in fig 2., *a* is the dilation (or scaling) parameter, and *b* is the translation parameter. The resolution in time can be trade-off for resolution in frequency by selecting various scaling parameters. For a specific resolution, the time-varying characteristics can be measured as the translation parameter *b* changes. Figure 3 shows a segment of two pitch cycles residual signal (top panel) and its 3 wavelet transforms in different scales (the bottom 3 panels). As shown, as *a* increases, time resolution decreases, while the frequency resolution improves. Also, the time varying characteristics of $u(n)$ can be measured from W **(*a*, *b*)** u at different translation parameters.





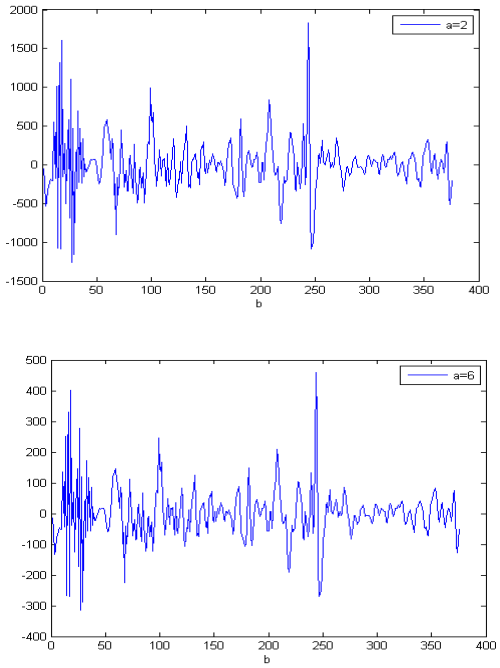


Fig 3.Analyzed LP residual signal (top panel) and its wavelet transforms (bottom three panels) with scaling parameters *a*=2, 4 and 8, respectively.

3. Time-Frequency Feature Generation:

To generate the time-frequency feature vector from *u*(*n*), we first do wavelet transform using Daubechies wavelet with dyadic(or octave) scaling Parameter,

$$a = \{2^k \mid k = 1,2\cdots,6\} \qquad (3)$$

and translation parameter,

$$b = 1,2,\cdots,N \qquad (4)$$

Where *N* equals to the length of *u*(*n*). All the $W_u(a,b)'s$ with a specific scaling parameter can be considered as the frequency analysis of the signal with a particular time-frequency resolution, and can be grouped together as

$$W_k = \{W_u(2^k,b) \mid b = 1,2,\cdots N\}, k = 1,2,\cdots,6 \qquad (5)$$

Each W_k is called an octave group. Finally, we derive the new feature vector as

$$\mathrm{WBFE} = \{\|W_k\| \mid k = 1,2.....6\} \qquad (6)$$

Where $\|\cdot\|$ denotes the 2-norm operator. In this case, the feature vector has 6 elements containing only frequency information, but not the temporal characteristics. To retain the temporal details, each octave group can be equally divided into 2 subgroups and then the energy of each subgroup is computed to generate a double sized feature vector noted as WBFE2. There are now 12 Elements in the WBFE and provides a certain degree of temporal information of the constituent frequency components. To extend further so as to obtain more detailed temporal characteristics, each octave group can be further divided into (α>2) subgroups,

$$W_k^{\alpha}(j) = \{W_u(2^k,b) \mid b \in (j-1:j] \times Round(N/\alpha)\} \qquad (7)$$
$$j = 1,2,\cdots\alpha$$

Note that the final subgroup of each octave may have more or less components than *Round*(*N* /α) Finally, a 8α-dimensional feature vector can be generated

$$\mathrm{WBFE} = \{\|W_k^{\alpha}\| \quad j = 1,2 \ldots.\alpha \qquad (8)$$

K=1,2……..,6

4. EXPERIMENTS

4.1 System Design

Experiments were conducted using female speech signals with different speaking rates and the speech data were recorded through a microphone in a reasonably quiet recording room. The sampling frequency is 8000 Hz. In this paper speaker verification experiments were carried out in which the classification is a binary choice decision. That is given the matching score of the input feature vectors against the claimed speaker model and the system should make a decision between two hypotheses: the input speech is from the claimed speaker H_1 or from an impostor H_0. It has been shown that the performance can be greatly improved by normalizing the female speaker model scores over the background speaker model scores (imposter H_0). Taking the stochastic models as an example, the decision can be made upon the log-likelihood ratio

$$\wedge = \log\frac{P\left(x_i/H_1\right)}{P\left(x_i/H_0\right)} = \log P\left(x_i/H_1\right) - \log P\left(x_i/H_0\right) \rangle \theta \qquad (9)$$

Where $P(x_i|H_1)$ is the probability of an observation *xi* generated by the claimed speaker, and $P(x_i|H_0)$ is the probability of the observation NOT generated by

1. Setting $\Theta = p_1/p_0$, where p_1 and p_0 are the *a priori* probabilities that input speech is from the true speaker and from the impostor, respectively.

2. Choosing μ to satisfy a fixed false accept rate (FAR) or false reject rate (FRR) according to the Neyman-Pearson criterion.

3. Experimentally determined in developing stage. That is, varies Θ to find different FAR and FRR and choose Θ to give the desired FAR/FRR ratio. This method has been adopted in most speaker verification systems.

The different speech utterances were trained by 256 components using GMM [11]. Verification is done for same speaker with different speech utterances. In this the proposed feature

extraction technique is compared with the prevalent feature extraction techniques like LPC's, LPC_DIFF feature.

4.2 Experimental Results:

Generally, the overall performance of speaker verification system [12] depends on the false acceptance rate (FAR) and false rejection rate (FRR). For verification, we measured the equal error rate (EER), which corresponds to where the false rejection rate and false acceptance rate are the same. In this system, performance of speaker verification is improved by fusing WBFE along with traditional feature vector (LPC,LPC_DIFF) is shown in fig 4(a,b and c).

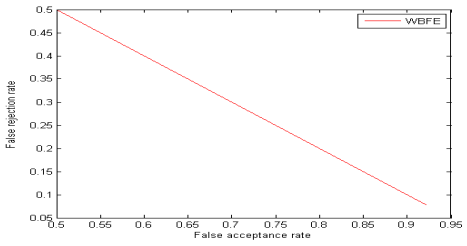


Fig 4. (a).

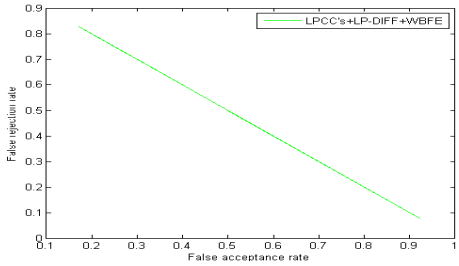


Fig 4 (b)

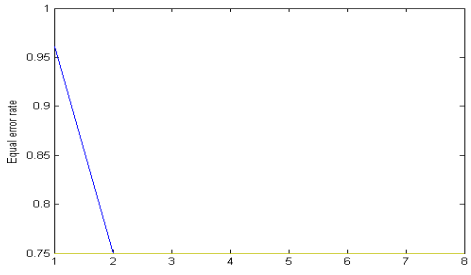


Fig 4. (c).

Fig4: Speaker verification results with WBFE, LPC, LPC DIFF and EER

Table 1 gives the respective verification results of the vocal tract information LPC+LPC_DIFF), the vocal source information (WBFE), and their information fusion. The contribution of temporal information for speaker verification is demonstrated in fig4.

Table 1:Verification results for different features

Performance	LPC+LP-DIFF	WBFE	WBFE+LPC+LP-DIFF
FRR	0.37	0.5	0.82
FAR	0.84	0.92	0.92
EER	0.68	0.96	0.91

5. CONCLUSIONS

In this paper, we presented a feature extraction technique to effectively retrieve the time-frequency information from the excitation signal. Upon pitch synchronous analysis, the wavelet transform is applied to the LP residual signal with exactly 2 pitch cycles. The feature set WBFE with (α >1) characterizes the speaker specific spectro-temporal properties of each pitch cycle, as well as the time evolution over successive pitch cycles. In testing phase, the discriminative power of WBFE is more effective in the speaker verification. Especially, when employing WBFE, as a complementary characteristic to the vocal tract features (LPC-DIFF).The speaker verification system with fused information relatively outperforms that with only LPC coefficients for verification [3].

6. REFERENCES

[1] G.Chenchamma "Speech Coding with Linear Predictive Coding," Proceedings of the ICMEE2009, Chennai, India was published in www.worldscibooks.com/etextbook.

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

[3] Furui, S., "Cepstral Analysis Technique for AutomaticSpeaker Verification," *IEEE Trans. Acoust.,Speech, Signal Processing*, Vol. ASSP-29, pp. 254-272, 1981.Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Campbell, J. P., "Speaker Recognition: A Tutorial,"*Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437-1462, 1997.

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

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

[7]] Plumpe, M. D., Quatieri, T. F., and Reynolds, D. A.,"Modeling of the Glottal Flow erivative Waveformwith

Application to Speaker Identification,"*IEEE Trans. Speech Audio Processing*, Vol. 7, No.5, pp. 569-585, 1999.

[8] ZhengNengheng, and Ching, P.C., "Using HaarTransformed Vocal Source Information for AutomaticSpeaker Recognition," to appear in the *Proc.ICASSP*2004, Montreal, Canada

[9] Quatieri, T. F., *Discrete-Time Speech Signal Processing*, Prentice-Hall, 2001

[10] Noll, A. M., "Cestrum Pitch Determination," *J.Acoust. Soc. Am.*, Vol. 41, pp. 293-309, 1967.

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

[12] Higgins, A., Bahler, L., and Porter, J., "Speaker Verification Using Randomized Phrase Prompting,"*Digital Signal Processing*, Vol. 1, pp.89-106. 1991.