A practical approach to real-time application of speaker recognition using wavelets and linear algebra

Duc Son Pham, Michael C. Orr, Brian Lithgow and Robert Mahony

Department of Electrical and Computer Systems Engineering
Monash University VIC 3168

Email: morr@ieee.org

Abstract:

A continuous wavelet transform, with Morlet wavelets as the basis functions, is used to map speech into the time-frequency domain. Forward and inverse FFT routines are used to implement the wavelet transforms. A coefficient covariance matrix is defined and an Eigenvalue decomposition is used to optimally determine significant wavelet based filters that accurately represent speech and potentially identify different speakers.

Key words: speaker recognition, wavelet transform, Morlet, Eigenvalue decomposition

1. Introduction:

Human perception of speech is, in part, reliant on detection of formant frequencies and their transitions times. Subjects utilize formant transition length to categorise sounds [2]. Real time speaker recognition is an active area of speech research. One possible explanation for the failure of existing algorithms to identify speakers is the accuracy of the signal analysis used, in characterising, for example, stop consonants [3]. A simple speaker recognition system generally consists of two stages. Firstly, the purely time series data filtered and mapped into a domain more suited to feature extraction. Many existing schemes use a windowed FFT algorithm to map speech data into a time-frequency domain. Secondly, the transformed data is compared with prior knowledge about a library of speakers to find a match. In order to obtain reasonable speech recognition the transformed signal must have both localised time and frequency resolution, i.e. the ability to capture non-stationary aspects of speech such as stop-constants. The windowed FFT process used by many speech recognition systems provides excellent information on formant frequencies and slower transitions, however, it has serious deficiencies in representing stop consonants and other key sounds used by humans in speech recognition.

In this paper we propose a wavelet-based analysis of speech using an Eigenvalue decomposition of the covariance matrix of the wavelet coefficients. The authors believe that the excellent time-frequency resolution of the Morlet will provide a good representation of human speech. The abstract nature information obtained in the wavelet transform domain is overcome by using covariance analysis on the wavelet coefficients to determine the best representation of speaker characteristics. In this preliminary study, it is demonstrated that the proposed method can be used for efficient reconstruction using either a full set or a significantly reduced set of coefficients.

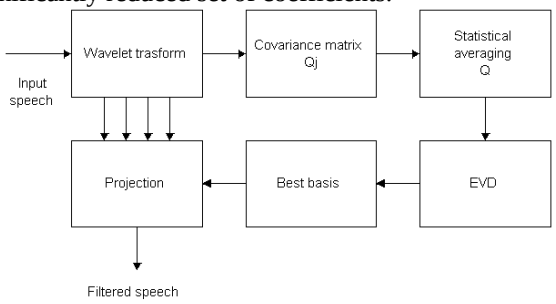


Figure 1: Proposed speech-processing system

2. Selection of wavelet and filter bank:

Many speech-processing techniques have tried to mimic the auditory filter bank through the use of cepstral filters [4]. These techniques are often solely based on the spectrum of a signal, unlike human hearing which also utilises temporal information [7]. Wavelet-based analysis provides better localisation in both time and frequency domains. In speech processing, this is essential for tracking vowels and consonants. The Morlet wavelet provides best time and frequency localisation.

$$\psi(t) = (2\pi)^{-1/2} e^{-t^2/2} e^{j\omega t}$$

To design of wavelet filter bank, the main considerations are the bandwidth (of individual wavelet filters) and the number of wavelet filters per octave. In this paper, the signal is sampled at 20kHz. The speech spectrum considered is the frequency range 100Hz – 8kHz. To provide separation of formant frequencies, each filter in the filter bank has a Q of approximately 6. The table below shows how the spectrum was split into octaves.

Octave	Starting frequency	Ending frequency
1	62.5 Hz	125 Hz
2	125 Hz	250 Hz
3	250 Hz	500 Hz
4	500 Hz	1 kHz
5	1 kHz	2 kHz
6	2 kHz	4 kHz
7	4 kHz	8 kHz

Table 1: Frequency range of filter bank

Additionally, in order to get a reasonably flat response across the spectrum, the filters were designed to to overlap at 3dB points. The wavelet filter bank can be then described by the filter coefficients

$$\boldsymbol{\beta}=[\beta_1, \beta_2,.. \beta_N]$$

in which $\beta_1 = 1$, $\beta_N < 2$ and the coefficient β_i corresponds to the filter

$$\psi_j(t) = (2\pi)^{-1/2} e^{-t^2/2} e^{j\beta_j \omega t}$$

Experiments on the response of individual filters indicate that 4 filters per bank is a good choice. A discussion of the derivation of the filter coefficient values is given in an appendix. The $\boldsymbol{\beta}$ values used are [1.07 1.37 1.65 1.94]. Figures 2 and 3 show the frequency response of the filters in the octave between 250Hz – 500 Hz

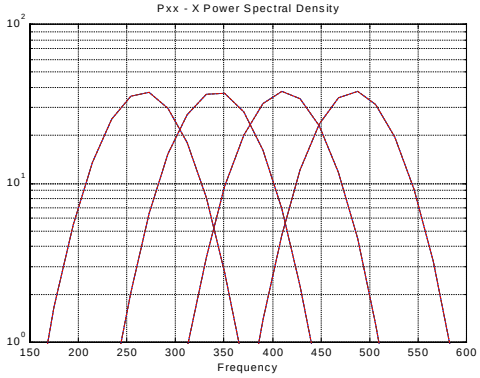


Figure 2: Frequency response of 4 wavelet filters 250Hz-500Hz

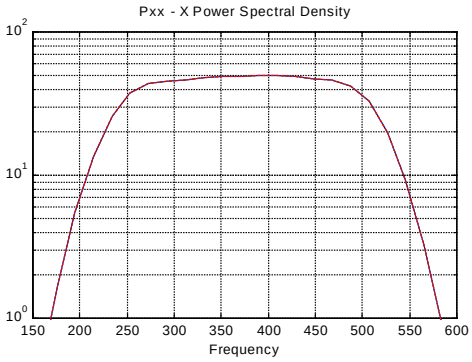


Figure 3:. Band pass filter250Hz- 500Hz ripple

3. Continuous wavelet transform:

The continuous wavelet transform is

$$CWT(a,\tau) = \int x(t) \frac{1}{\sqrt{a}} \psi(\frac{t-\tau}{a}) dt \qquad (1)$$

The wavelet function $\psi(t)$ used is the Morlet wavelet. The discrete version of the wavelet transform can be written as

$$CWT(m,n) = \sum_{k=0}^{N} x[k] 2^{m/2} e^{-(\alpha k/2)^2} e^{-j\frac{2\pi}{N} 2^m (k-n)} \qquad (2)$$

where
$$\begin{cases} \text{m : frequency scale} & \text{m} = 0-6 \\ \text{n : time shift} & \text{n} = 0\text{-}1023 \\ \text{k : sample index} & \text{k} = 0\text{-}1023 \end{cases} \qquad (3)$$

and α= 0.004 is the scale factor to fit the Morlet wavelet to the time frame used.

Direct implementation of (2) and (3) requires a large number of computations. If (2) is viewed as the response of a signal to a linear system, i.e. a simple convolution, computationally it is advantageous to calculate (2) in the frequency domain [1, 5]. In addition, if the term *(k-n)* in (2) is considered as the frequency term in the FFT, the wavelet transform for an entire octave can be evaluated as the inverse Fourier transform

$$CWT(m, n:0...N-1) = IFFT\{FFT(x(k)) * FFT(\psi_m(k))\}$$

where the scaled wavelet for a particular octave is

$$\psi_m(k) = 2^{m/2} \psi_0(k/2^m)$$

Even though the Morlet wavelet bases are not orthogonal, the reconstruction can be evaluated from the following formula [5]

$$\{x_n : n = 0...N-1\} = \frac{\delta j}{C_\delta \psi(0)} \sum_{m=0}^{M} \frac{CWT(m,n)}{2^m} \qquad (4)$$

where C_δ is a constant (0.776) for Morlet wavelet and δj will be evaluated from experiments [5]. To some extent, δj is the compensation factor because the continuous wavelets are not orthogonal.

An algorithm similar to [5] is proposed for fast implementation of the wavelet transform. The algorithm is to be hardware implementable:

1. Design the fundamental wavelet bases to cover the lowest frequency range
2. FFT the sampled signal on a 1024 sample frame basis
3. At each scale and for each sub-band, decimate the fundamental wavelet, padding zeros to have equal number of points as the signal and FFT the baby wavelets

4. Using linear vector space algorithm, process the time frequency information.
5. Inverse FFT the processed result backs to time domain to obtain the approximate signal of a speaker.

Figure 3 depicts the system response for a short speech sample by a male speaker. The maximum magnitude of the error is within 5% of the peak magnitude of the original signal.

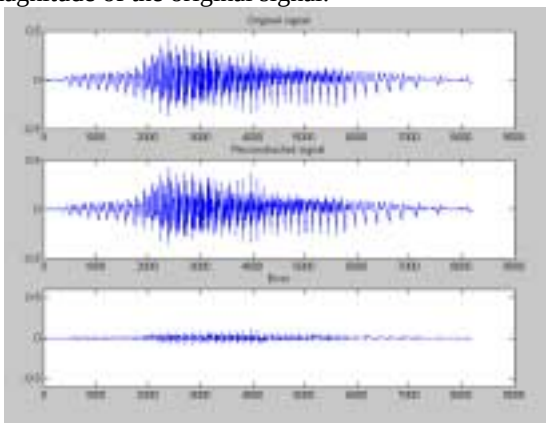


Figure 3: Word "one" by a male speaker and the reconstructed signal with full set of wavelet coefficients

4. Best vector basis and speaker recognition:

Mathematically, the wavelet coefficients associated with the frequency response and localized in time can be viewed as a vector in linear vector space:

$$\omega_j(k) = \{\omega_{j1}, \omega_{j2}, ...\omega_{jK},\} = \{CWT(m,n) : m = 1..M, n = 1:N\}$$

This vector may be thought of as a time-sequence itself that varies in magnitude and direction in the linear vector space according to the time-varying characteristics of the processed speech. If the vector is randomly distributed, such as would be obtained if Gaussian noise were substituted for the speech signal, the vector will span the all directions with equal probability. The authors hypothesise that the characteristic sounds made by individual speakers will lead to a unique subspace structure in the time-varying response of the wavelet filter coefficients averaged over time. The structure may be the result of a number of physical attributes of the speaker, such as structure of the vocal tract, etc. To use this structure to design a speech recognition algorithm it is sufficient to characterise the subspace associated with a specific speaker or class of speakers. To illustrate, suppose that $\Omega_P = \{\omega_{jP}\}$ and $\Omega_Q = \{\omega_{jQ}\}$ denote the subspaces in the linear vector space associated with the average male and female pitch tones. If a signal is processed and it is found that the average wavelet filter coefficients lie consistently in Ω_Q and are significantly disparate from Ω_P then the speaker is more likely to be female than male.

If a vector time-sequence has a structure that carries sufficient information to be identified and extracted, it should be possible to identify the subspace associated with this structure by studying the covariance of this signal. Suppose this subspace has dimension P where (0<P<K), then the subspace is represented by a sequence of P vectors that span the subspace directions

$$\hat{\omega}_j(k) = \{\omega_{j1}, \omega_{j2} ... \omega_{jP}\}$$

This is the best basis that can optimally represent a given speaker. It is expected that for our speaker recognition, (1) the P value will be small and (2) different speakers will have different sets of code-vector basis. Of course, the number of speakers that can be differentiated depends upon the total number of subspaces in the linear vector space and the number of subspaces needed to represent each speaker.

To find out the structure or the best basis, Eigenvalue decomposition (EVD) is used on the covariance matrix [1, 6]. The covariance matrix Q is essentially the matrix that describes the correlation between the subspaces of a given vector.

$$Q_j = \omega_j(k)\omega^H{}_j(k) = \begin{pmatrix} \omega_{j1} \\ \omega_{j2} \\ .... \\ \omega_{jK} \end{pmatrix} \begin{pmatrix} \omega_{j1}{}^* & \omega_{j2}{}^* & .... & \omega_{jK}{}^* \end{pmatrix}$$

The structure of a vector will develop for a statistical averaging of the covariance matrix

$$\overline{Q} = \frac{1}{N}\sum_{j=1}^{J} Q_j$$

The covariance matrix can be decomposed in to the diagonal and the Eigen vector matrices

$$\overline{Q} = VDV^{-1}$$

If the structure in the Q matrix is represented by a small number of subspaces, the diagonal matrix should have a small number of significant Eigenvalues and the rest are insignificant

$$D = \begin{pmatrix} \lambda_1 & & & & & ... \\ & ... & & & & \\ & & \lambda_P & & & \\ & & & \varepsilon_1 & & \\ & & & & ... & \\ & & & & & \varepsilon_{K-P} \end{pmatrix}$$

The best basis will be the Eigen vectors that are associated with the P significant Eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_P\}$

Suppose the best basis is $V_P = \{v_1, v_2, ...v_P\}$, the "rectified" speech vector is the projection of the original vector on the subspace of the best basis

$$\hat{\omega}_j(k) = \sum_{i=1}^{P} v_i < v_i, \omega_j(k) >$$

An algorithm to approximate a signal by best vector subspaces can be described as follow:

1. Process the signal and map the signal to a time-frequency vector.
2. Normalise the vector
3. Calculate the covariance matrix Q
4. Average the covariance matrix Q over J times (J should be large enough to ensure the structure develops).
5. EVD the average Q matrix
6. Pick out the best P vector basis corresponding to the P largest Eigenvalues
7. Project the speech vector onto the best basis subspace
8. Using inverse wavelet transform to bring back to time domain to obtain the "rectified" signal.

We describe here our first successful attempt to filter a real speech signal using best vector basis approach. In this experiment, a structure between the audio bands of a speech signal is to be examined. It defines which sub-bands over seven octaves are significant in representing the unique characteristics of each speaker. The experiment aims at the possibility of representing a speaker with a smaller number of linear vector subspaces than the full set of 28 as in our design.

Figure 4 shows the graphical representation of the diagonal matrix and the Eigen vector matrix. The diagonal matrix suggests that there are a small number of significant Eigen values. Figure 5 depicts the signal as reconstructed from the best 10 linear vector subspaces, hence it appears possible to represent speech with the reduced subspace.

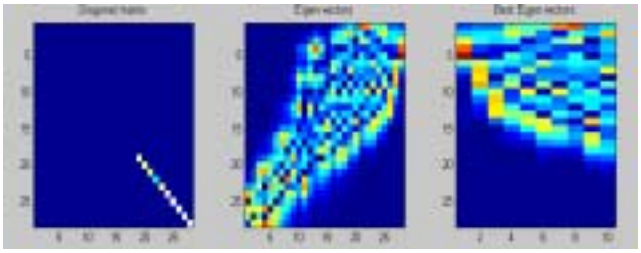


Figure 4. The diagonal matrix (left), Eigen vector matrix (middle) and best vector basis matrix (right)

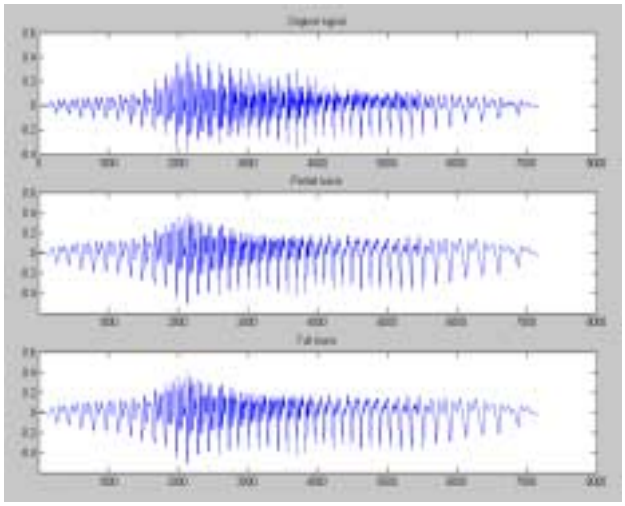


Figure 5. The signals original (top) and as rectified from best 10 (middle) and full subspace (bottom)

Another series of tests have also been conducted. In most cases, as far as speech intelligibility is concerned, the representation of speech by the best basis approach provides good intelligibility. The longer the samples over which the average statistical Q matrix is measured, the better the representation.

Based on this success, it appears theoretically and practically possible to improve speaker recognition. The critical point in future extensions of the work is the definition and estimation of structure necessary to set up a proper framework for vector subspace analysis. The next stage is to develop the covariance matrix algorithm for short-term mapping of target speakers. It is envisaged that in doing this, an algorithm for filtering overlapping speakers could be developed. Such analysis would require careful and detailed work and is beyond the scope of this paper.

5. Conclusion:

A novel approach for speech characterization and reconstruction using wavelets and linear algebra was described. The design of wavelet filters and the wavelet transform have been proven efficient and sufficient. The authors believe that the algorithm described in this paper could be adapted in the future to be the front end of a speaker recognition system. Even though real time requirements are not strictly imposed in this development, the proposed approach is promising.

6. References:

[1] Delmas, J.P., "On adaptive EVD asymptotic distribution of centro-symmetric covariance matrices," *IEEE Transactions on Signal Processing*, vol. 47, pp. 1402 - 1406, 1999.

[2] Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C., "The discrimination of

speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, pp. 358 - 368, 1957.

[3] Smits, R., "Accuracy of quasistationary analysis of highly dynamic speech signals," *Journal of the Acoustical Society of America*, vol. 96, pp. 3401 - 3415, 1994.

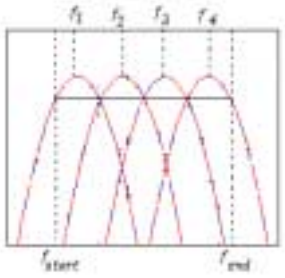
[4] Tokuda, K., Kobayashi, T., and Imai, S., "Adaptive cepstral analysis of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 481 - 489, 1995.

[5] Torrence, C. and Compo, G. P., "A practical guide to wavelet analysis," *Bulletin of the American Meterological Society*, vol. 79, pp. 61 - 78, 1998.

[6] Yu, K.-B., "Efficient, parallel adaptive eigen-based techniques for direction of arrival estimation and tracking," *IEEE Transactions on Signal Processing*, pp. 347 - 351, 1990.

[7] Zwicker, E. and Fastl, H., *Psychoacoustics: Facts and models*, 2nd ed: Springer-Verlag, 1999

Appendix Calculation of filter coefficients



The Q factor is defined as the ratio of frequency and the bandwidth associated with that frequency

$$Q = \frac{f}{BW} \quad (1)$$

Assume we are designing two filters centred at f_1 and f_2 with bandwidth BW_1 and BW_2. Then in order to have f_1 and f_2 intersect at the 3dB point, we would require

$$f_2 - \frac{1}{2}BW_2 = f_1 + \frac{1}{2}BW_1 \quad (2)$$

(assume that $f_2 > f_1$)

In addition, we would like to have these two filters have constant Q

$$Q = \frac{f_1}{BW_1} = \frac{f2}{BW_2} \quad (3)$$

Hence, from (1) (2) and (3)

$$f_2 = f_1\left(\frac{1+ 1/2Q}{1- 1/2Q}\right) = f_1\left(\frac{2Q+1}{2Q-1}\right)$$

Now if we want to design a bank of four filters of four filters at f_1, f_2, f_3, f_4 such that the lower 3dB of f_1 is at $f_{start\text{-}band}$, the high 3dB of f_4 is at $2*f_{start\text{-}band}$ or $f_{end\text{-}band}$ and they intersect each other at 3dB points, then

$$f_3 = af_2 = a^2 f_1$$

$$f_4 = af_3 = a^2 f_2 = a^3 f_1$$

where

$$a = \left(\frac{2Q+1}{2Q-1}\right)$$

is a function of Q

Hence

$$f_4 = f_1\left(\frac{2Q+1}{2Q-1}\right)^3$$

The start band frequency and the end-of-band frequency are found from

$$f_{start-band} = f_1 - \frac{1}{2}BW_1$$

$$f_{end-band} = f_1 + \frac{1}{2}BW_4$$

From the relation between BW, Q and f as derived above and given that

$$2f_{start\text{-}band} = f_{end\text{-}band}$$

We have

$$f_4(1+\frac{1}{2Q}) = 2\times f_1(1-\frac{1}{2Q})$$

Then from the result for f_4 above

$$\left(\frac{2Q+1}{2Q-1}\right)^4 = 2$$

This equation yields

$$Q \approx 5.7852$$

$$\left(\frac{2Q+1}{2Q-1}\right) = \sqrt[4]{2} = 1.189^2$$

In terms of the start-band frequency

$$f_1 = \frac{f_{start-band}}{1- 1/2Q} \approx 1.095\, f_{start-band}$$

$$f_2 = 1.189\, f_1 = 1.3017\, f_{start-band}$$

$$f_3 = 1.189^2\, f_1 = 1.548\, f_{start-band}$$

$$f_2 = 1.189^3\, f_1 = 1.8409\, f_{start-band}$$

Therefore, the theoretical array of coefficients of filters is

[1.095 1.3017 1.5480 1.8409]