International Journal of the Physical Sciences Vol. 6(14), pp. 3398-3408, 18 July, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS10.309
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Information optimization for speaker recognition using correlation functions

Alireza Salahshour Vaskas[1]*, Shahaboddin Shamshirband[2], Mohsen Gholami[3] and Mohsen Amiri Besheli[4]

[1]Department of Computer Science, Islamic Azad University, Ghemshahr Branch, Mazandaran, Iran.
[2]Department of Basic Science, Islamic Azad University, Chalous Branch, Mazandaran, Iran.
[3]Iranian academic center for education, culture and research, Mazandaran Branch, Sari, Iran.
[4]Andishe Pars Education and Research Institute, Iran.

Accepted 08 June, 2011

In this article, a method that optimizes the information of an analyzed signal is described. This method is practicable for various analyses and in fact is an additional process that is used after the analysis. We tendered our idea in our previous article. In this method, we first analyzed the signal, after which the first frame and some other frames were put in a matrix, where each row was a frame. Then the likeness of this matrix was computed with itself. Subsequently, this process was done for the second frame and the other frames following it, after which it was done for all frames.

Key words: Cestrum, MODGDF, vocal tract, autocorrelation.

INTRODUCTION

Speaker recognition usually consists of the three stages.First, Preemphasize the speech signal X[n] followed by frame blocking (usually at a frame size of 20 ms and frame shift of 10 ms). A Hamming window is applied on each frame of the speech signal.

Next, analyze the signal with an authentic analysis for speech and finally speaker models are constructed from the features extracted from the speech signal. Then a match score that is a measure of the similarity between the input feature vectors and some other model was computed. But sometimes an additional process combines with the main analysis that is called Information optimization.

In this paper is described the speech production and is discovered that extraction of similarity between frames can optimize the information for speaker recognition.

SPEECH PRODUCTION

Modeling process is usually divided into two parts: the excitation source of vocal tract and the vocal tract.

The excitation source of vocal tract

The airstreams from the lungs passes through the glottis to the vocal tract. The action of the vocal folds determines the phonation type, whose major types are voiceless, whisper and voicing.

During whisper and voiceless phonation, the vocal folds are apart from each other, and the airstreams from the lungs will pass through the open glottis. The difference between whisper and voiceless phonation is determined by the degree of the glottal opening. In whisper, the glottal area is smaller. This results in a turbulent airstreams, generating the characteristic “hissing" sound of whispering. In voiceless phonation, the area of the glottis will be larger and the airstreams is only slightly turbulent when it enters the vocal tract.

Voicing is a more complex mechanism than voiceless phonation and whisper. Voicing is a result of periodic repetitions of the vocal folds opening and closing. During the opening phase, the respiratory effort builds up the sub glottal pressure until it overcomes the muscular force which keeps the vocal folds together. The glottis opens, and the compressed airstreams bursts into pharynx with a speed of 2 to 5 m/s

This relatively high speed causes a local drop of air pressure at the glottis, and as a consequence of this so-

*Corresponding author. E-mail: shahab.sham@gmail.com

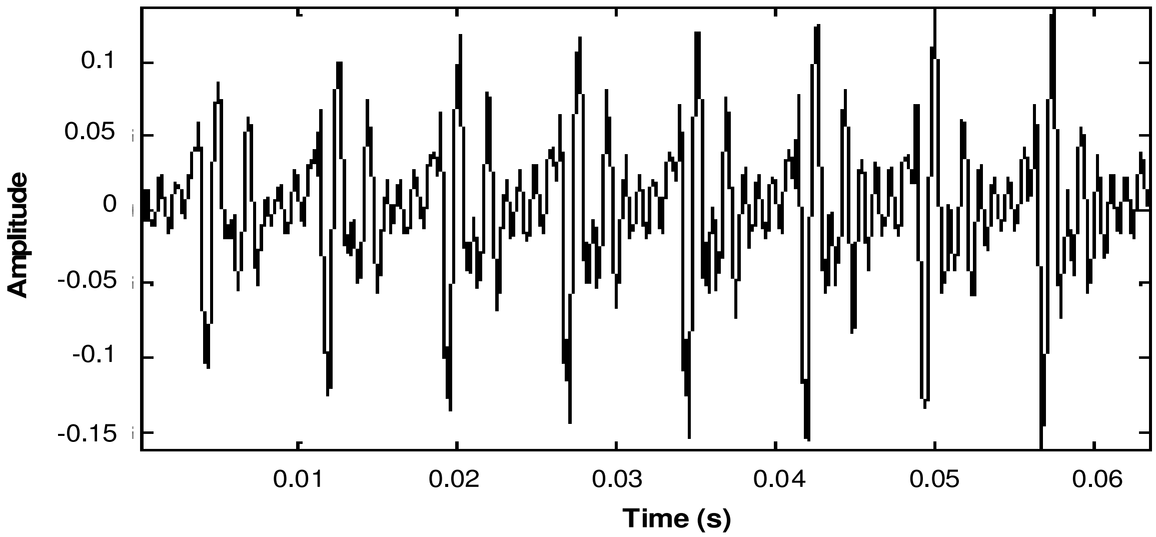

Figure 1. Excitation signal for voicing.

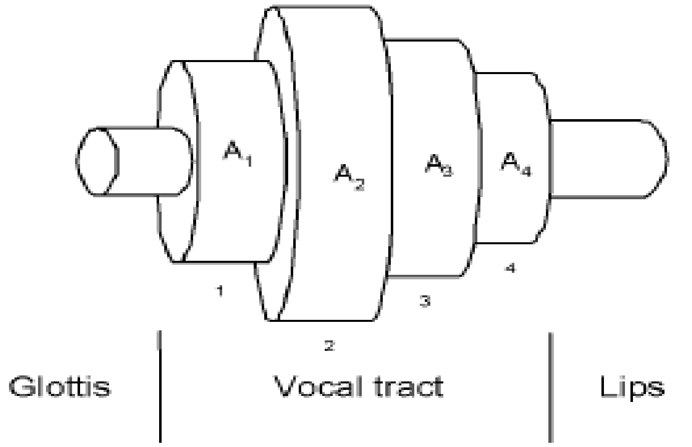

Figure 2. Vocal tract.

called Bernoulli effect, the vocal folds start to close. The combined effort of the Bernoulli effect and muscular tension overcomes the force of respiratory pressure very quickly, and the vocal folds are pulled together. The coupling of the opening and closing phases continues, and the result is a periodic stream of air puffs which serves as the acoustic source signal for the voiced sounds. We refer to this signal as train of impulse (Figure 1).

The vocal tract

Often the acoustic filter is modeled as a hard-walled tube resonator. In this so-called lossless tube model, the vocal tract is considered as a cascade of N lossless tubes with varying cross-sectional areas (Figure 2).

For this kind of resonator, the resonances can be computed analytically. In the case of a single tube (N = 1), the resonances of the tube (formant frequencies) are given by the following Equation 1:

$$F_n = \frac{(2n - 1)\,c}{4L} \tag{1}$$

Where Fn is the nth formant frequency [Hz], c is the speed of sound in air [m/s], and L is the total length of the tube [m].

As a result, the speech signal is formed with multiplying the excitation function in the vocal tract and with

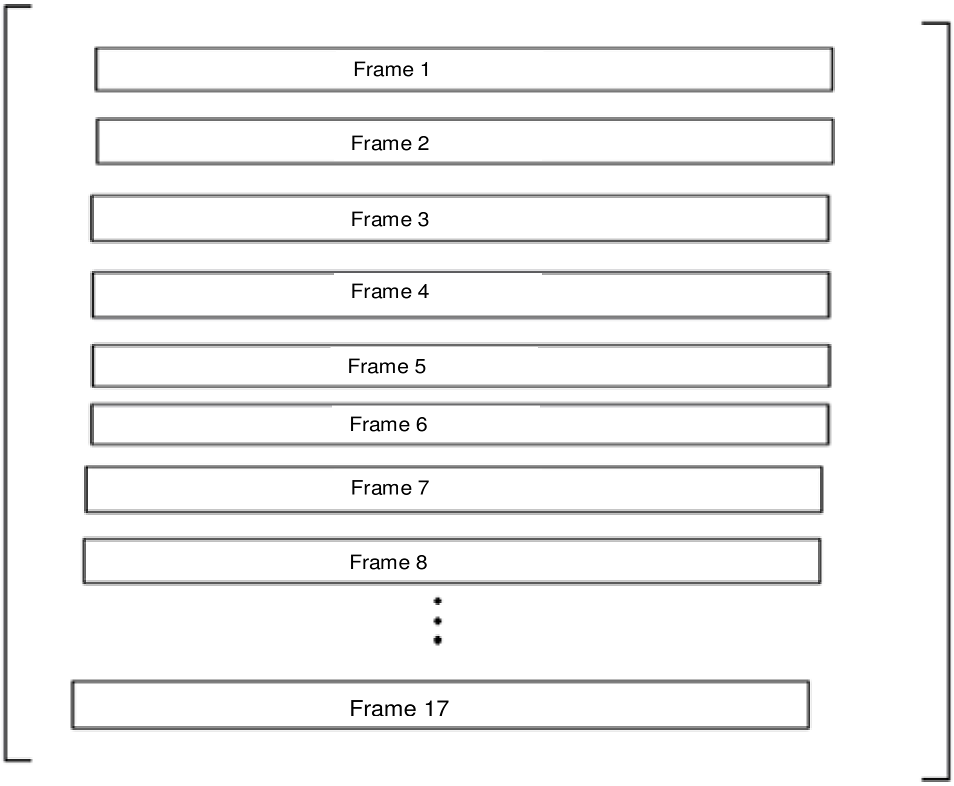

Figure 3. The matrix contains the first frame and its next 16 frames.

reference to this signal; we can extract information from the vocal tract.

So we can say the voicing phonemes signal is impulse response of vocal tract and the voiceless phonemes signal conveys the white noise and the vocal tract information together. As a result of this, the voicing phonemes are often used for speaker recognition.

INTRODUCTION OF INFORMATION OPTIMIZATION USING EXTRACTION OF SIMILARITY BETWEEN FRAMES

Here, our idea is described using a tentative example. Assume a speaker that you know his/her speech, if the speaker says "door" you will identify his/her speech and if the speaker says "pack" you will identify his/her speech again.

Maybe this is usual for you, but if this topic is surveyed microscopically, you will discover that however the words (pack and door) have different phonemes but you identify the speaker. So we can say the speech convey two types of information, first the information of phonemes said by the speaker and secondly, the information of the speaker. The first type of information in various frames is different because the phoneme that is spoken in the various frames is different. But the second type of information in various frames is same because the speaker in the various frames is same.

Therefore, extraction of similarity between frames only extracts the same information between frames, and because of this, the obtained information is only about the speaker and is very good for speaker recognition.

INFORMATION OPTIMIZATION USING SIMILARITY EXTRACTION

For information optimization in this article, first analyze the speech frames with the group delay function analysis that will be described later. Then make a matrix containing the first frame and its next 16 frames (Figure 3). Then using the correlation function formula (2) computes the autocorrelation coefficients in various directions.

$$Rf(a,b) = \sum_{-\infty}^{+\infty} \sum_{-\infty}^{+\infty} f[x,y].f[x+a,y+b] \qquad (2)$$

The directions that are used in this article for computing the autocorrelation coefficients are selected tentatively.

In many experiments, it is found that the horizontal autocorrelation coefficients will get best results for

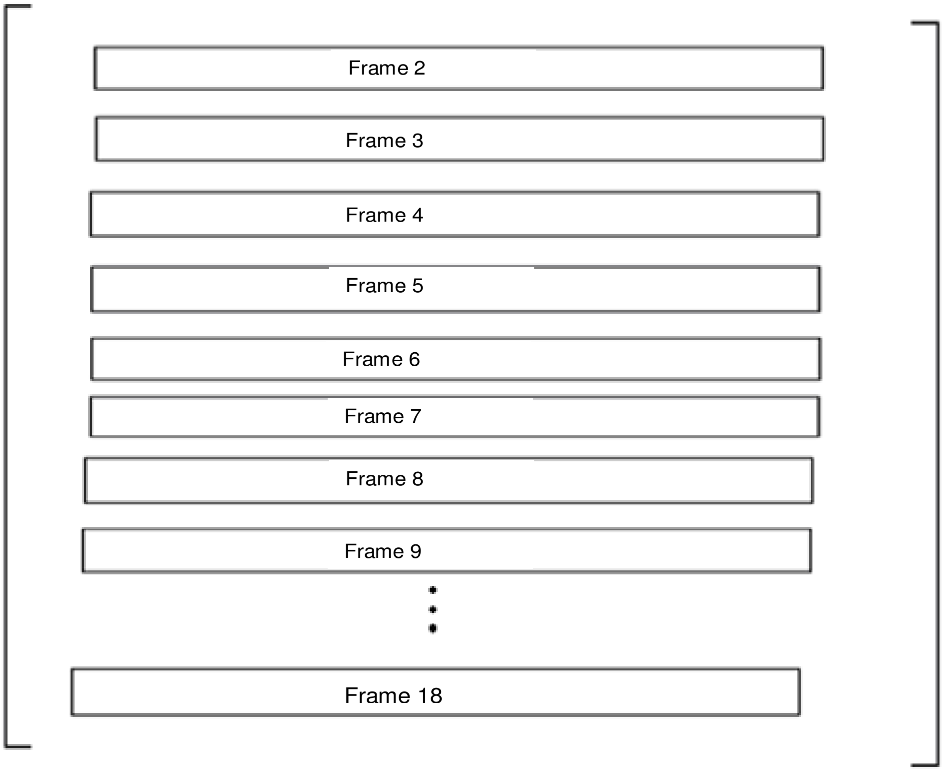

Figure 4. The matrix contains the second frame and its next 16 frames.

speaker recognition. For computing the horizontal autocorrelation coefficients, fix the "b" and vary the "a" from 0 to Nf in the autocorrelation formula that the Nf is the length of each frame (Negheng, 2005).

Subsequently, some coefficients were obtained from the analyzed frame whose length is as long as the frame. The autocorrelation coefficients length is as long as the length of the analyzed signal.

Then take the discrete cosine transform from obtained coefficient and extract the first coefficients as feature vectors. Then make another matrix with the second frame and its 16 next frames (figure 4). After this step, the discreet cosine transform is used to extract some coefficients.

Thus, after the optimization process, we take a feature vector for each frame with some coefficients that is ready to apply to the neural networks.

MODIFIED GROUP DELAY FUNCTION ANALYSIS

Group delay is defined as the negative derivative of the Fourier transform phase. Mathematically, the group delay function is defined as:

$$GDF_{(\omega)} = -\theta'_{(\omega)} \tag{3}$$

The Fourier transform phase and the Fourier transform magnitude are related. The group delay function can also be computed from the signal using:

$$\theta(\omega) = \tan^{-1} XI(\omega) / XR(\omega) \tag{4}$$

$$GDF(\omega) = -\theta'(\omega) = -\partial / \partial(\omega)(\tan^{-1} XI(\omega) / XR(\omega)) \tag{5}$$

$$GDF_{(\omega)} = \frac{H_{R(\omega)} Y_{R(\omega)} + Y_{I\omega} H_{I(\omega)}}{| H_{(\omega)} |^2} \tag{6}$$

Where the subscripts R and I, denote the real and imaginary parts of the Fourier transform. X(ω) and Y(ω) are the Fourier transforms of X[n] and nX[n] , respectively (Hegde et al., 2007).

Properties of group delay functions

The two main properties of the group delay functions of relevance to this work are as follows:

(1) Additive property;
(2) High-resolution property.

Additive property

The group delay function exhibits an additive property. Let

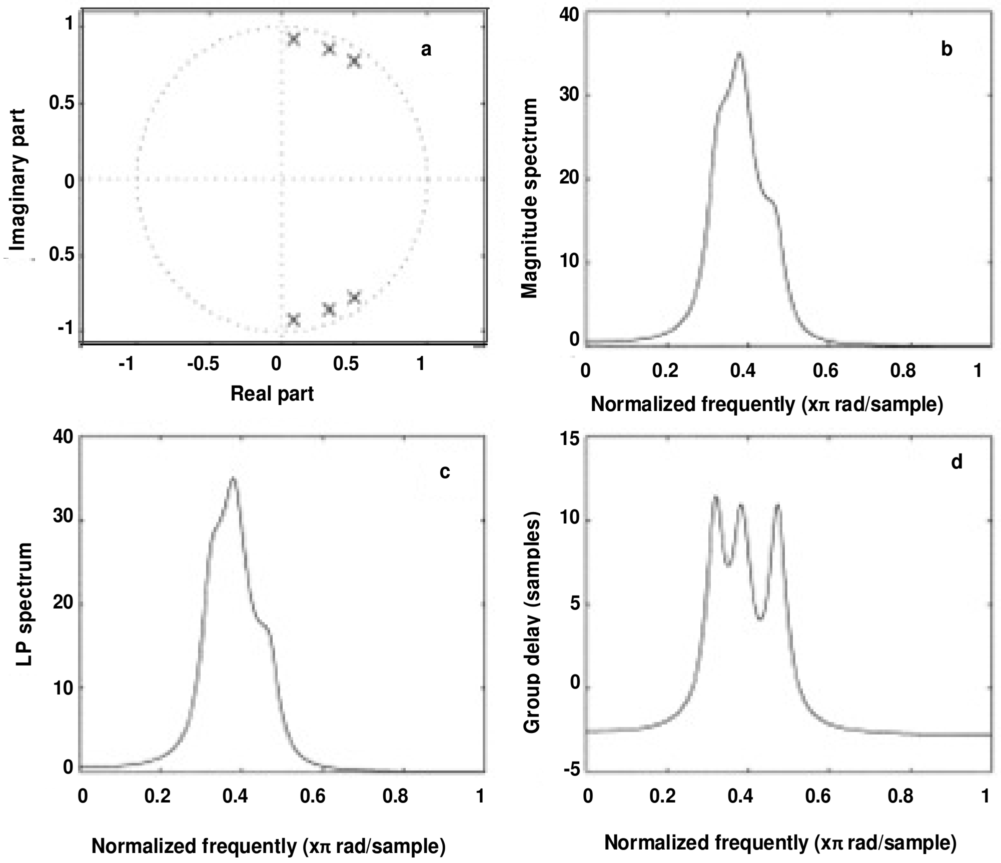

Figure 5. Comparison of the minimum phase group delay function with the magnitude and linear prediction (LP) spectrum.

$$H(\omega) = H_1(\omega).H_2(\omega) \tag{7}$$

Where H_1 and H_2 are the responses of the two resonators whose product gives the overall system response. Taking absolute value on both sides we have

$$| H(\omega) |=| H_1(\omega) | . | H_2(\omega) | \tag{8}$$

Using the additive property of the Fourier transform phase

$$\arg(H(\omega)) = \arg(H_1(\omega)).\arg(H_2(\omega)) \tag{9}$$

Then, the group delay function is given by

$$GDF(\omega) = -\partial(\arg(H(\omega)) / \partial\omega = -\partial(\arg(H_1(\omega)) + \arg(H_2(\omega))) / \partial\omega$$

$$GDF(\omega) = -\partial(\arg(H_1(\omega)) / \partial\omega - \partial(\arg(H_2(\omega))) \tag{10}$$

$$GDF(\omega) = GDF_1(\omega) + GDF_2(\omega)$$

where тh1 and тh2 correspond to the group delay function of H_1 and H_2, respectively. It is clear that multiplication in the spectral domain becomes an addition in the group delay domain (Hegde et al., 2007).

High-Resolution property

The group delay function has a higher resolving power when compared to the magnitude spectrum. The ability of the group delay function to resolve closely spaced formants in the speech spectrum has been investigated in an illustration given in Figure 5 to highlight the high-resolution property of the group delay function over both the magnitude and linear prediction spectrum. Figure 5(a) shows the z–plane plot of the system consisting of three complex conjugate pole pairs. Figure 5(b) is the corresponding magnitude spectrum, while Figure 5(c) illustrates the spectrum derived using LPC analysis, and Figure 5(d) is the corresponding group delay spectrum. It can be clearly observed that the three formats are resolved better in the group delay spectrum when compared to the magnitude or linear prediction spectrum. From these results, it is also evident that the system information in the speech signal is captured relatively better by the group delay spectrum when compared to the magnitude or linear prediction spectrum (Hegde et al., 2007).

Basis for modifying the group delay function

It has been shown in Yegnanarayana et al. (1984) that group delay functions can be used to accurately represent signal information as long as the roots of the z-transform of the signal are not too close to the unit circle in the z-plane. It is also true that the vocal tract

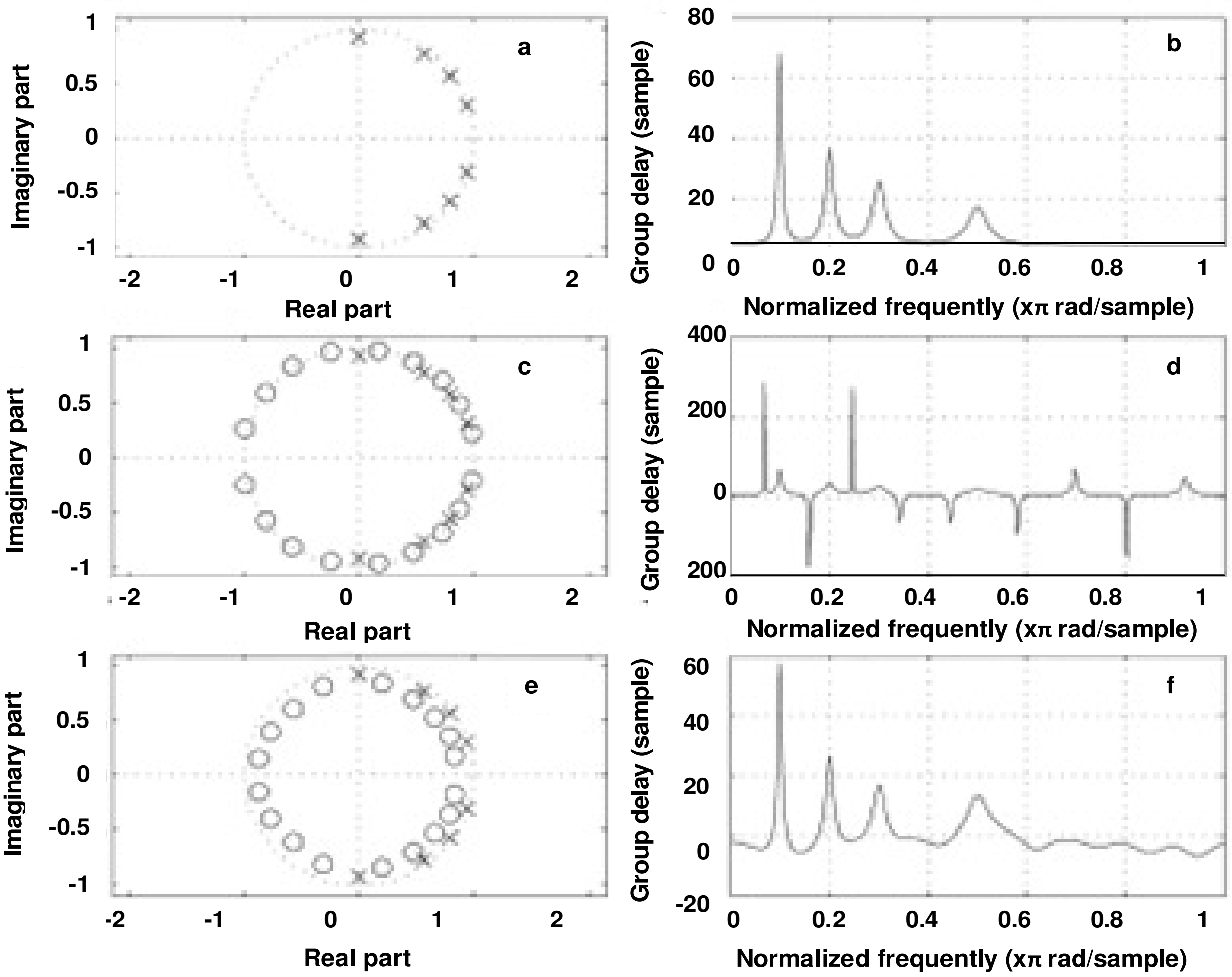

Figure 6. Significance of proximity of zeros to the unit circle.

system and the excitation contribute to the envelope and the fine structure respectively of the speech spectrum. When the Fourier transform magnitude spectrum is used to extract speech features, the focus is on capturing the spectral envelope of the spectrum and not the fine structure.

Similarly, the fine structure has to be de-emphasized when extracting the vocal tract characteristics from the group delay function. The zeros that are close to the unit circle manifest as spikes in the group delay function and the strength of these spikes is proportional to the proximity of these zeros to the unit circle. To illustrate this, a four formant system with four poles and their complex conjugates is simulated. The pole-zero plot of the four formant system is shown in Figure 6(a) while the corresponding group delay spectrum is shown in 6(b). Figure 6(c) shows the pole-zero plot of the same system with zeros added uniformly in very close proximity to the unit circle. It is evident from Figure 6(d) that the group delay spectrum for such a system becomes very spiky and ill defined primarily due to zeros that are added in very close proximity to the unit circle in the z-plane. In Figure 6(e), we manually move all the zeros radially into the unit circle and re-compute the group delay function of such a system. The group delay spectrum of such a system is shown in Figure 6(f). It is clear that this technique of pushing the zeros into the unit circle radially restores the group delay spectrum without any distortions in the original formant locations. The spikes introduced by zeros close to the unit circle form a significant part of the fine structure and cannot be eliminated by normal smoothing techniques. Hence, the group delay function has to be modified to eliminate the effects of these spikes. Here, the considerations discussed so far form the basis for modifying the group delay function (Hegde et al., 2007).

Modified group delay function

As mentioned, for the group delay function to be a meaningful representation, it is only necessary that the roots of the transfer function are not too close to the unit circle in the plane. Normally, in the context of speech, the poles of the transfer function are well within the unit circle. The zeros of the slowly varying envelope of speech correspond to that of nasals.

The zeros in speech are either within or outside the unit circle since the zeros also have nonzero bandwidth. Here, we modify the computation of the group delay function to suppress these effects.

Let us reconsider the group delay function derived directly from the speech signal. It is important to note that the denominator term |X(ω)| in (6) becomes zero, at zeros that are located close to the unit circle. The spiky nature of the group delay spectrum can be overcome by replacing the term |X (ω)| in the denominator of the group delay function as in (6) with its cepstrally smoothed version, |S(ω)| (Hegde et al., 2007).

Significance of cepstral smoothing

Assuming a source system model of speech production, the z-transform of the system generating the speech signal is given by

$$H(z) = \frac{N(z)}{D(z)} \tag{11}$$

Where the polynomial N(z) is the contribution due to zeros and the polynomial D(z) is the contribution due to the poles of the vocal tract system. The frequency response of H(z) is given by

$$H(\omega) = \frac{N(\omega)}{D(\omega)} \tag{12}$$

Where N(ω) and D(ω) are obtained by evaluating the polynomials on the unit circle in z-domain. By using the additive property of the group delay function, the group delay function of the system characterized by H(ω) is given by

$$GDF_H(\omega) = GDF_N(\omega) - GDF_D(\omega) \tag{13}$$

where τ N (ω) and τ D (ω) are the group delay functions of N (ω) and D (ω), respectively. Spikes of large amplitude are introduced into τ N (ω) primarily due to zeros of N (z) close to the unit circle. As already discussed, the group delay function can be directly computed from the speech signal as

$$GDF_{(\omega)} = \frac{H_{R(\omega)}Y_{R(\omega)} + Y_{I\omega}H_{I(\omega)}}{|H_{(\omega)}|^2} \tag{14}$$

The group delay function for τ N (ω) in (13) can be written as

$$GDF_N(\omega) = \frac{\alpha_N(\omega)}{|N(\omega)|^2} \tag{15}$$

where α N (ω) is the numerator term of (14) for τ N (ω). As |N (ω) |² tends to zero (for zeros on the unit circle), τ N (ω) has large amplitude spikes. Similarly, the group delay function for τ D (ω) in (13) can be written as

$$GDF_D(\omega) = \frac{\alpha_D(\omega)}{D(\omega)^2} \tag{16}$$

where α D (ω) is the numerator term of (14) for τ D (ω). The term | D (ω) |² does not take values very close to zero since D(z) has all roots well within the unit circle. Therefore, the term τ D (ω) contains the information about the poles of the system and has no spikes of large amplitude. Substituting (13) and (15) in (16), we have:

$$GDF_H(\omega) = \frac{\alpha_N(\omega)}{|N(\omega)|^2} - \frac{\alpha_D(\omega)}{|D(\omega)|^2} \tag{17}$$

where α N (ω) and α D (ω) are the numerator terms of (14) for τ N (ω) and τ D (ω), respectively. Assuming that the envelope of | N (ω) |² is nearly flat (zero spectrum), multiplying τ x (ω) with | N (ω) |² will emphasize the resonant peaks of the second term. This leads to the initial form of the modified group delay function which is given by

$$MODGDF_H(\omega) = GDF_H(\omega).|N(\omega)|^2 \tag{18}$$

Substituting (17) in (18)

$$= \alpha_N(\omega) - \frac{\alpha_D(\omega)}{|D(\omega)|^2}.|N(\omega)|^2 \tag{19}$$

In (19), an approximation to |N(ω)|² is required, which is a nearly flat spectrum (ideally a zero spectrum). An approximation E (ω) to |N (ω) |² can be computed as

$$E(\omega) = \frac{S(\omega)}{S_c(\omega)} \tag{20}$$

Where S (ω) is the squared magnitude (|N (ω) |²) of the signal X[n] and Sc (ω) is the cepstrally smoothed spectrum of S (ω) (Hegde et al., 2007). Alternately, the modified group delay function can be defined as

$$GDF_{(\omega)} = \frac{H_{R(\omega)}Y_{R(\omega)} + Y_{I\omega}H_{I(\omega)}}{|H_{(\omega)}|^2} \tag{21}$$

Therefore, the modified group delay function is capable of pushing zeros on the unit circle, radially into the unit circle, and thus emphasizing τ D (ω) which corresponds to the contribution from the poles of the vocal tract system (Yegnanarayana et al., 1984).

Definition of the modified group delay function

Since the peaks at the formant locations are very spiky in nature, two new parameters α and γ are introduced to reduce the amplitude of these spikes and to restore the dynamic range of the speech spectrum. The new modified group delay function is defined as:

$$MODGDF_{(\omega)} = \left(\frac{\tau_{(\omega)}}{|\tau_{(\omega)}|}\right)\left(|\tau_{(\omega)}|\right)^{\alpha} \tag{22}$$

Where

$$\tau_{(\omega)} = \frac{H_{R(\omega)}Y_{R(\omega)} + Y_{I\omega}H_{I(\omega)}}{S_{(\omega)}^{2\gamma}} \tag{23}$$

where S (ω) is the smoothed version of | X (ω) |. The parameters α and γ introduced vary from 0 to 1.

Robustness to convolutional and white noise

Assuming a source system model of speech production, the clean speech X c (n), its Fourier transform and the corresponding group delay function (Hegde et al., 2007) is given by

$$x_c(n) = \sum_{k=1}^{p} a_k x_c(n-k) + Ge(n) \tag{24}$$

$$X_c(\omega) = \frac{GE(\omega)}{A(\omega)}. \tag{25}$$

Similarly, the noisy speech signal and its Fourier transform are given by

$$x_n(n) = x_c(n) * h(n) + w(n) \tag{26}$$

$$X_n(\omega) = X_c(\omega)H(\omega) + W(\omega) \tag{27}$$

where h (n) is the time invariant channel response and w (n) is the additive white noise. Taking the Fourier transform of (24) and substituting in (27), X n (ω) and the corresponding group delay function τ N (ω) is given by

$$X_n(\omega) = \frac{GE(\omega)H(\omega) + A(\omega)W(\omega)}{A(\omega)} \tag{28}$$

$$\tau_n(\omega) = \tau_{\mathrm{numerator}}(\omega) - \tau_a(\omega) \tag{29}$$

where τ numerator (ω) is the group delay function corresponding to that of GE(ω)H(ω)+A(ω)W(ω) , and τ a (ω) is the group delay function corresponding to A(ω).

Further, the term GE(ω)H(ω) in τ numerator (ω) dominates in high signal-to-noise ratio (SNR) regions and the term A(ω)W(ω) in τ numerator (ω) dominates in low SNR regions. Since α is chosen such that (0<α<1), the question of noise being emphasized does not arise. In the high SNR case, it is the excitation, and in the second case, it is white noise that makes the group delay spectrum spiky and distorted primarily due to zeros that are very close to the unit circle in the z-domain.

White noise has a flat spectral envelope, and hence, contributes zeros very close to the unit circle. Further, the locations and amplitudes of these spikes are also not known. To suppress these spikes, the behavior of the spectrum where the noise zeros contribute to sharp nulls is utilized. A spectrum with a near flat spectral envelope containing the spectral shape contributed by the zeros is derived using cepstral smoothing as discussed and multiplied with the group delay function to get the modified group delay function as in (22) and (23). The effects due to the excitation can be dealt with by pushing all zeros very close to the unit circle in the z-domain, well inside the unit circle by appropriately selecting values for the two parameters α and γ as defined in (22) and (23).

In Hegde et al. (2007), the log and root compression approaches are compared with the MODGDF in the presence of white noise at different values of SNR and is picked 20 complete sentences from different dialect regions, consisting of both female and male speakers, from the TIMIT database. These sentences are added with white noise scaled by a factor η.

The average error distributions between the clean and the noisy speech across all frames corresponding to the 20 sentences are then calculated for four different values of SNR 0, 3, 6, and 10 dB and is illustrated in Figure 7.

It is clear from Figure 4 that average deviation of the noisy speech cepstra from the clean speech cepstra is the least for the MODGDF when compared to either the spectral root, the energy root, or the log compressed cepstra (Hegde et al., 2007).

COMPUTATION OF VARIOUS FEATURES

Here, the computation of the MODGDF and optimized MODGDF are discussed.

Algorithm for computing the modified group delay cepstra

The following is the algorithm for computing the modified group delay cepstra:

(1) Preemphasize the speech signal x(n) followed by frame blocking at a frame size of 20 ms and frame shift of 10 ms. A hamming window is applied on each frame of the speech signal (IMPEDOVO et al., 2008).
(2) Compute the DFT of the framed and windowed speech signal x (n) as X (k) and the time scaled speech signal nx(n) as Y(k).
(3) Compute the cepstrally smoothed spectra of |X (k)|. Let this be S (k). A low-order cepstral window (lifterw) that essentially captures the dynamic range of |X (k)| should be chosen. In Hegde et al. (2007), it is discussed that the best value for this window is 6.
(4) Compute the modified group delay function by (22) and (23). And in this step we will obtain a spectrum as long as the original frame and because of the spectrum is bilateral we get the first half of the spectrum.
(5) The parameters α and γ introduced vary from 0 to 1. In Hegde et al. (2007), it is discussed that the best value for α is 0.4 and the best value for γ is 0.9.
(6) Compute the modified group delay cepstra as

$$c(n) = \sum_{k=0}^{Nf} f(k)\cos(n(2k+1)\pi / Nf) \tag{30}$$

Where τ m (k) is the modified group delay spectra and Nf is the length of τ m (k).

Algorithm for computing the optimized modified group delay cepstra

The following is the algorithm for computing the optimized modified group delay cepstra:

(1) Preemphasize the speech signal x(n) followed by frame blocking at a frame size of 20 ms and frame shift of 10 ms. A hamming window is applied on each frame of the speech signal figure 8 (IMPEDOVO et al., 2008).
(2) Compute the DFT of the framed and windowed speech signal x (n) as X (k) and the time scaled speech signal nx(n) as Y(k).
(3) Compute the cepstrally smoothed spectra of |X (k)|. Let this be S (k). A low-order cepstral window (lifterw) that essentially captures the dynamic range of |X (k)| should be chosen. In Hegde et al. (2007), it is discussed that the best value for this window is 6.
(4) Compute the modified group delay function by (22) and (23). And in this step we will obtain a spectrum as long as the original frame and because the spectrum is bilateral, we get the first half of the spectrum.
(5) The parameters α and γ introduced vary from 0 to 1. In Hegde et al. (2007), it is discussed that the best value for α is 0.4 and the best value for γ is 0.9.
(6) Make a matrix contains the first frame and its next 16 frames (Figure 3). Then compute the horizontal autocorrelation coefficients as discussed in information optimization using similarity extraction (R[0,0],R[0,1],R[0,2],R[0,3],...,R[0,Nf]) and do the same for all frames.

Note that the Nf is the analyzed frame length.

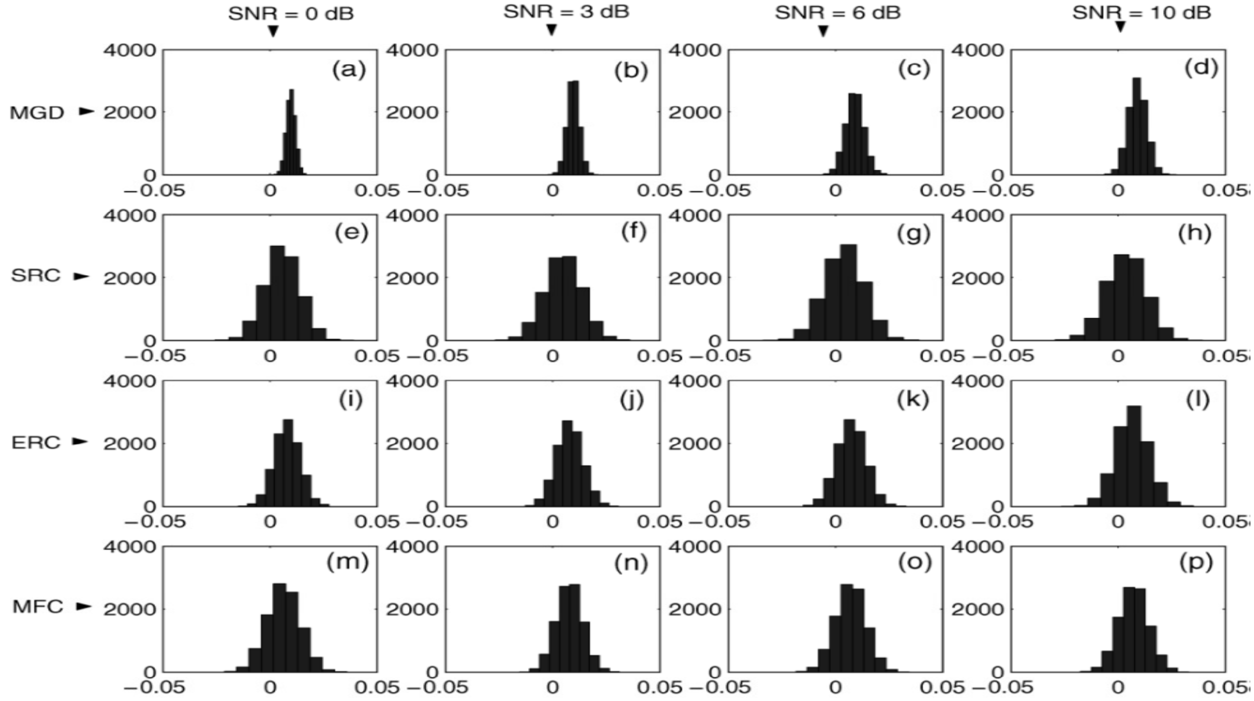

Figure 7. Comparison of the average error distributions of the MODGDF, MFCC, and root compressed cepstra in noise. (a) Error distribution of the MODGDF (α = 0:4, γ = 0:9) at 0-dB SNR. (b) Error distribution of the MODGDF (α = 0:4, γ = 0:9) at 3-dB SNR. (c) Error distribution of the MODGDF (α = 0:4, γ = 0:9) at 6-dB SNR. (d) Error distribution of the MODGDF (α = 0:4, γ = 0:9) at 10-dB SNR. (e) Error distribution of the spectrally root compressed cepstra (root = 2/3) at 0-dB SNR. (f) Error distribution of the spectrally root compressed cepstra (root = 2/3) at 3-dB SNR. (g) Error distribution of the spectrally root compressed cepstra (root = 2/3) at 6-dB SNR. (h) Error distribution of the spectrally root compressed cepstra (root = 2/3) at 10-dB SNR. (i) Error distribution of the energy root compressed cepstra (root = 0.08) at 0-dB SNR. (j) Error distribution of the energy root compressed cepstra (root = 0.08) at 3-dB SNR. (k) Error distribution of the energy root compressed cepstra (root = 0.08) at 6-dB SNR. (l) Error distribution of the energy root compressed cepstra (root = 0.08) at 10-dB SNR. (m) Error distribution of the MFCC at 0-dB SNR. (n) Error distribution of the MFCC at 3-dB SNR. (o) Error distribution of the MFCC at 6-dB SNR. (p) Error distribution of the MFCC at 10-dB SNR.

(7) Compute the optimized modified group delay cepstra as (30) where τ m (k) is the modified group delay spectra and Nf is the length of τ m (k).

EXPERIMENTAL SETUP

Data base

The data base in this project contained from 10 sentences from 20 various speakers from TIMIT that 7 sentences are used for training and 3 sentences are used for testing the network.

The neural network

The used neural network in this project is a three layer back propagation network, that is, the first layer has 8 neurons, the hidden layer has 19 neurons and the output layer has 20 neurons which are equal to the number of speakers and this network has 20 inputs (Muzhir et al., 2007).

The used function in first layer and hidden layer is tansig that is between -1 and 1, and the used function in output layer is logsig that is between 0 and 1 Figure 9.

Apply each frame to input of network and get a number between 0 and 1 in each rows of output of network that each number is probability of each speaker figure 10.

RESULTS

The simulation results for clean signal in timit database that is described in previously sections are illustrated in Table 1. The simulation results for noisy signal with SNR3 in Timit database are illustrated in Table 2.

The recognition percent in these tables is low because the simulation is text independent. In text independent, the network training and data base testing varies (Gish et al., 1994). Usually, in the text independent simulation, the recognition percent is lower than text dependent simulation. Finally, we can see in the tables that the recognition percent of the Optimized GDF is better.

CONCLUSION

In this article, we describe a method for optimization of analyzed frames using extraction similarity between one

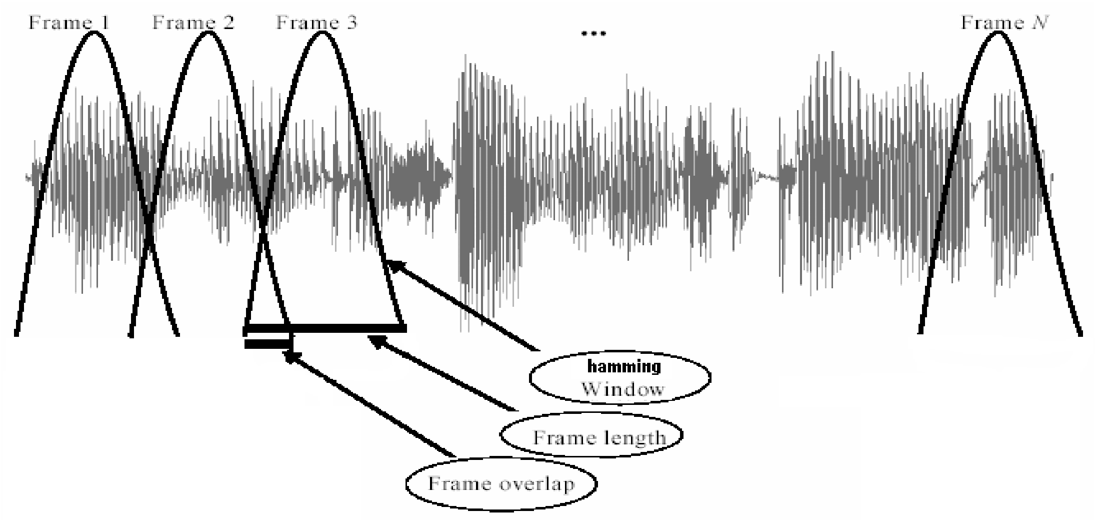

Figure 8. The windowing process.

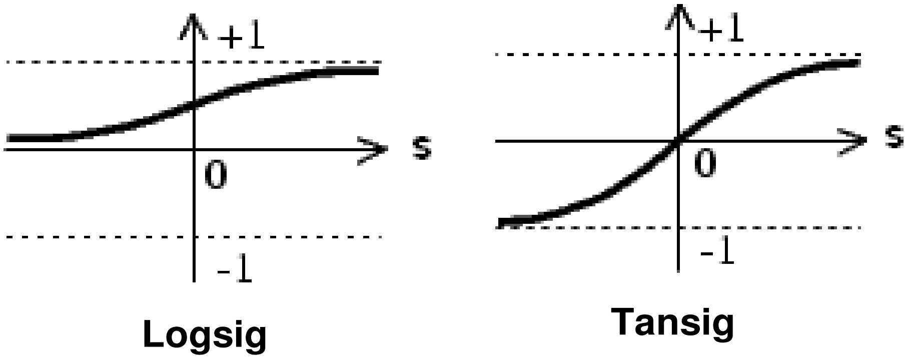

Figure 9. Tansig and logsig functions.

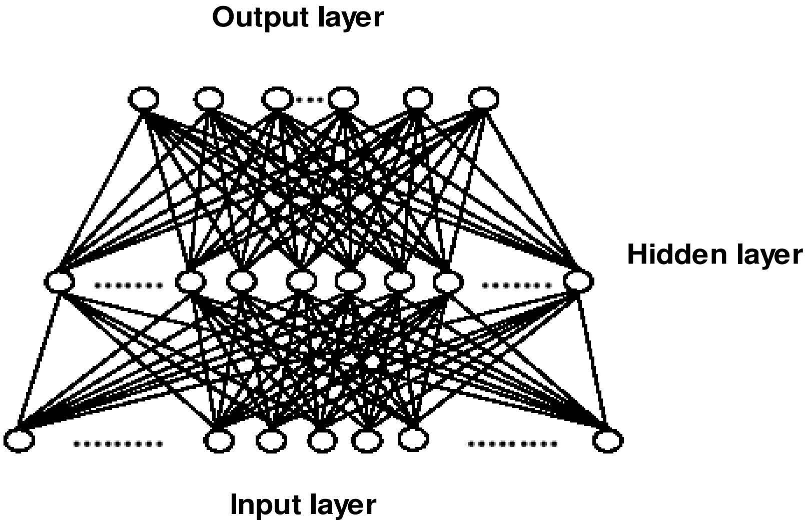

Figure 10. Back propagation neural network.

Table 1. The results of recognition percent for clean signal.

Number of speakers	Feature	Recognition (%)
20	GDF	90
20	Optimized GDF	96.66

Table 2. The results.

Number of speakers	Feature	Recognition (%)
20	GDF	83.33
20	Optimized GDF	90

frame and next frames. We used from correlation function and extracted the correlation coefficient in various direction.

Note that this method can be used for every analysis and not only in extracting the information of speaker and removing the information of phonemes. The used analysis in this text is group delay function and the results of simulations, describes that the correlation method is better than other.

REFERENCES

Gish H, Schmidt M (1994).Text independent speaker identification. IEEE Signal Process, 11(4):18–32.

Hegde R, Hema M, Murthy A, Gadde VRR (2007).Significance of the Modified Group Delay Feature in Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 15(1):190-202.

Impedovo D, Mario R (2008). Frame Length Selection in Speaker Verification Task. WSEAS Transactions on Systems, 7(10):1028-1037.

Muzhir SA, Thabit SM, Karim MA (2007).Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform. J. Comput. Sci., 3(5): 304-309.

Negheng Z (2005). Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Electronic Engineering, The Chinese University of Hong Kong.

Yegnanarayana B, Saikia DK, Krishnan TR (1984). Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Trans, Speech, Signal Process, 32(3): 610–622.