The Text Independent Speaker Recognition Using Modified Group Delay Function Analysis and Correlation Function

Alireza Salahshour Vaskas, Sayed Ahmad Ghasemi, Mohsen Gholami, Shahabodin Shamirband

1 Sama Organization (Affiliated with Islamic Azad University) Sari Branch.
2 Iranian Academic Center for Education, Culture & Research.
3 Young Research Club, Islamic Azad University, Chalous Branch.

Abstract: In this research we describe a method for feature extraction from the analyzed signal that is based on the extracting similarity between one frame and the next frames. In practice this method doesn’t extract feature from one frame instead of doing this for many frames. Also we can use this method for various analyses and in practice this method is used for decreasing the dimensions of the frame for modeling & matching. The analysis used in this project is modified Group delay function that we will know it in this text.

Key words: Cepstrum, GDF, larynx filter, vocal tract, decreasing dimensions, MODGDF, Auto1, Auto2

INTRODUCTION

The speaker recognition consists of following five steps:
- Framing and windowing the speech signal
- Preprocessing
- Main analysis
- Decreasing dimensions
- Modeling & matching

After framing & windowing & preprocessing we analyze the signal of each frame. This step applies to signal for extraction the important information from it that often uses the frequency information. After this step we have to decrease the dimensions of each frame because the length of each frame is long & we can’t use it for modeling & matching. In this text we describe sound production & conclude extraction of similarity between frames can be better than feature extraction from individual frame.

2. Modified Group Delay:

The analysis that we use in this text is called Modified Group delay function.

2.1 Definition of the Group Delay Function:

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

\[ GDF(\omega) = -\frac{d}{d\omega} \phi(\omega) \]

(1)

The Fourier transform phase and the Fourier transform magnitude are related together. Therefore with using of this relationship and (1) the group delay function can be computed directly from the signal:

\[ GDP(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \]

(2)

Corresponding Author: Alireza Salashshoor Vaskas, Islamic Azad University, Sama Organization Branch Sari. E-mail: Ali.selahshoor@gmail.com
2.2 High-resolution Properties:

High-resolution property is one of the important reasons for using the GDF analysis in speaker recognition. An experiment is given in following to highlight the high-resolution property of the group delay function we simulate the system consisting of three complex conjugate pole pairs that the complex conjugate pole pairs is very close to each other & a impulse apply to its input & a signal produces in its output.

We refer to this signal as a speech signal & illustrate the different spectrum drives from the signal.

Fig. 1: The simulated system

Fig. 2: Magnitude of fourier transforms

Fig. 3: Log magnitude of Fourier transforms
It can be clearly observed that the three formants are resolved better in the group delay spectrum when compared to the other analysis.

2.3 Additive Property:
The group delay function exhibits an additive property. Let

\[ H(e^{j\omega}) = H_1(e^{j\omega}) \cdot H_2(e^{j\omega}) \]  \hspace{1cm} (3)

The magnitude of \( H \) is:

\[ |H(e^{j\omega})| = |H_1(e^{j\omega})| \cdot |H_2(e^{j\omega})| \]  \hspace{1cm} (4)

Using the additive property of the Fourier transform phase

\[ \text{ar} = \text{ar}(H(e^{j\omega})) = \text{ar}(H_1(e^{j\omega})) + \text{ar}(H_2(e^{j\omega})) \]  \hspace{1cm} (5)

Then, the group delay function is given by

\[ \text{GDF}_H(\omega) = \text{GDF}_H(\omega) + \text{GDF}_H(\omega) \]  \hspace{1cm} (6)

It is clear that multiplication in the spectral domain becomes addition in the group delay domain.

2.4 the Cause of Modifying the Group Delay Function:
Assuming a source system model of speech production is given by:

\[ H(\omega) = \frac{N(\omega)}{D(\omega)} \]  \hspace{1cm} (7)

Then, by Using of the additive property:

\[ \text{GDF}_H(\omega) = \text{GDF}_N(\omega) - \text{GDF}_D(\omega) \]  \hspace{1cm} (8)
Remember that the group delay function can be computed directly from the signal:

\[
\text{GDF} \left( \omega \right) = \frac{\alpha_{\omega} \left( \omega \right)}{\left| X(\omega) \right|^2} = \frac{\alpha_{\omega} \left( \omega \right)}{\left| X(\omega) \right|^2}
\]

The group delay function for GDFN (\( \omega \)) can be written as:

\[
\text{GDF}_N \left( \omega \right) = \frac{\alpha_{N}(\omega)}{|N(\omega)|^2}
\]

\( aN(\omega) \) is the numerator of (10).

The group delay function for GDFD (\( \omega \)) can be written as:

\[
\text{GDF}_D \left( \omega \right) = \frac{\alpha_{D}(\omega)}{|D(\omega)|^2}
\]

\( aD(\omega) \) is the numerator of the (11). We know in location of zeros that are very close to unit circle, the magnitude of \( N(\omega) \) tends to be zero and the GDFN (\( \omega \)) tends to be "infinity".

\[
|N(\omega)|^2 \rightarrow 0 \quad \frac{\alpha_{N}(\omega)}{|N(\omega)|^2} \rightarrow \omega
\]

And in location of poles that are very close to unit circle, the magnitude of \( D(\omega) \) tend to be zero and the GDFD (\( \omega \)) tend to be "infinity".

\[
|D(\omega)|^2 \rightarrow 0 \quad \frac{\alpha_{D}(\omega)}{|D(\omega)|^2} \rightarrow \omega
\]

And we know the poles of a speech source filter can not be very close to unit circle therefore only in location of zeros that very close to unit circle the GDF has large amplitude spikes.

An experiment is given in following to highlight the importance of modifying the GDF:

We simulate the system consisting of three complex conjugate pole pairs and apply an impulse in its input, and then a signal is generated in its output that we refer to this signal as a speech signal and compute the GDF spectrum of this signal:

Fig. 4: Simulated system.
You can see in this case the GDF describes the system information very good.

Now, we simulate the same system with zeros added uniformly in very close proximity to the unit circle and apply an impulse in its input get a signal from its output:

![Simulated system](image)

**Fig. 6: Simulated system.**

We refer to this signal as a speech signal and compute the GDF spectrum of this signal:

![GDF spectrum](image)

**Fig. 7: GDF of above signal**

You can see in fig.7 that when we added the zeros to the system; GDF of this system doesn't consist of information of system.

2.5 **Modifying the Group Delay Function:**

2.5.1 **First Modifying:**

Remember from previous sections that the zeros that are very close to unit circle produces large amplitude spikes in GDF spectrum that is undesirable for us.

Assuming a source system model of speech production is given by:
According to additive property:

\[ GDF_{H}(\omega) = GDF_{M}(\omega) - GDF_{D}(\omega) \]  \hspace{1cm} (15)

The group delay function can be computed directly from the signal:

\[ GDF_{x}(\omega) = \alpha_{x}(\omega) \left( \frac{X_{I}(\omega)Y_{I}(\omega) + Y_{I}(\omega)X_{I}(\omega)}{|X(\omega)|^2} \right) = \frac{\alpha_{x}(\omega)}{|X(\omega)|^2} \]  \hspace{1cm} (16)

And we combine the (16) and (15):

\[ GDF_{x}(\omega) = \frac{\alpha_{x}(\omega)}{N(\omega)^2} - \frac{\alpha_{D}(\omega)}{|D(\omega)|^2} \]  \hspace{1cm} (17)

Remember that \(|N(\omega)|\) in the denominator of first term of (17), tend to be 0 at location of zeros that is close to unit circle and GDF\(_{N}(\omega)\) tend to be \(\infty\) and GDF\(_{H}(\omega)\) tend to be \(\frac{1}{c}5^2\).

For solving this problem, we can multiply GDF\(_{H}(\omega)\) with \(|N(\omega)|\):

\[ \text{MODGDF}(\omega) = GDF_{H}(\omega) \cdot |N(\omega)|^2 \]

\[ = \left( \frac{\alpha_{x}(\omega)}{|N(\omega)|^2} - \frac{\alpha_{D}(\omega)}{|D(\omega)|^2} \right) \cdot |N(\omega)|^2 \]

\[ \text{MODGDF}(\omega) = \frac{\alpha_{x}(\omega)}{|D(\omega)|^2} \cdot |N(\omega)|^2 \]  \hspace{1cm} (18)

Therefore we solve the problem because \(|N(\omega)|\) is not in denominator of the first term of the above equation and GDF\(_{N}(\omega)\) doesn’t tend to be \(\infty\) if \(|N(\omega)|\) tends to be 0 therefore GDF\(_{H}(\omega)\) doesn’t tend to be \(\infty\).

But we have a new problem. We don’t know what is \(|N(\omega)|\) and we have to approximate the \(|N(\omega)|\) from \(|X(\omega)|\).

\[ 2.5.2 \text{ Approximating} |N(\omega)|: \]

For approximated \(|N(\omega)|\) from main signal we use from cepstrum technique.

\[ 2.5.2.1. \text{Cepstrum Technique:} \]

We have a system that is contained from tow subsystems namely H1 and H2 that convolved with them. We compute Fourier transform from this system, therefore the convolution convert to multiplication:

\[ \mathcal{H}(e^{j\omega}) = \mathcal{H}_1(e^{j\omega}) \cdot \mathcal{H}_2(e^{j\omega}) \]  \hspace{1cm} (19)

And we given logarithm from the system therefore the multiplication convert to addition:

\[ \log(H(\omega)) = \log(H_1(\omega)) + \log(H_2(\omega)) \]  \hspace{1cm} (20)
This transform is the cepstrum technique that converts the convolution to addition and its axis called quefrency.

Now if H1 and H2 have various frequencies the location of H1 and H2 in quefrency axis will be various and we can approximate one of the tow subsystems with a window.

2.5.2.2 the Speech Signal in Quefrency Axis:

According to many researches, the speech magnitude spectrum is combined from slow and quickly varying parts.

1) “Slowly varying”:
This part contains the effective information of system such as poles of vocal tract.
This part location in quefrency axis is in low frequency.

2) “Quickly varying”:
This part doesn’t contain the effective information of system such as zeros of system that are very close to unit circle or the source signal of vocal model.
This part location in quefrency axis is in high frequency.

![Fig. 8: The speech signal in cepstrum axis](image)

For approximating $|N(\omega)|$ from $|X(\omega)|$ first we convert the frequency axis to quefrency axis and in quefrency axis we approximated the "slowly varying” with using a window called "lifterw” and called this part $|Sc(\omega)|$.
This window begins from 0 HZ and its length is 4 to 9.

![Fig. 9: the speech signal in cepstrum axis](image)

And finally for approximated the “quickly varying” we divided the $|X(\omega)|$ by $|Sc(\omega)|$:

$$|N(\omega)| = E(\omega) = \frac{X(\omega)}{Sc(\omega)}$$  \hspace{1cm} (20)

2.5.3 Primary Modifying the GDF:

Remember that for modifying the Group delay function multiplying it with $|N(\omega)|$, we replace $|N(\omega)|$ with $|E(\omega)|$:

$$MODGDF(\omega) = GDF_H(\omega) \cdot |E(\omega)|^2$$

$$MODGDF(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \cdot |E(\omega)|^2$$

$$MODGDF(\omega) = \left( \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^2} \right)$$  \hspace{1cm} (22)
An experiment is given in following to highlight the first modifying of GDF:
Remember the system that its zeros are very close to unit circle and its GDF doesn’t contain the system information.

![Simulated system](image1)

**Fig. 10:** Simulated system

![GDF of above system](image2)

**Fig. 11:** GDF of above system

Know we illustrate the MODGDF of same system:

![MODGDF of same system](image3)

**Fig. 12:** MODGDF of same system

In this figure we brightly see the MODGDF of system that is contained information of system.

### 2.5.4 Finally Modifying of Group Delay Function:

The used approximation for modifying of GDF generated an error in MODGDF.

This error causes GDF and MODGDF of the system without zero isn’t same and this isn’t desired because we only want to emit extra zeros that are very close to unit circles.

For solve this problem create tow new parameter namely $\alpha$ and $\gamma$:
\[ MODGDF(\omega) = \left( \frac{\tau(\omega)}{|\tau(\omega)|} \right) \left( |\tau(\omega)| \right)^\alpha \]

\[ \tau(\omega) = \left( \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \right) \]

(23)

\( a \) and \( \gamma \) is variable between 0 to 1.

Now, in order to fix the values of \( a \) and \( \gamma \), we consider a system characterized by four formants (four complex conjugate pole pairs) and it system has no zero. The system is excited with an impulse and get the corresponding signal. Now we vary \( a \) and \( \gamma \) from 0 to 1 and with step of 0.1 and computed Mean square error between GDF and MODGD by following equation:

\[ \text{MSE} = \frac{1}{N} \sum_{1}^{N} \epsilon(k)^2 \]

\( \epsilon(k) = (\tau_{\text{MOD}}(\omega) - \tau_{\text{MIN}}(\omega)) \)

(24)

then find the point of minimum mean square error and find \( a \) and \( \gamma \) in this point this is the Optimal Values for \( a \) and \( \gamma \). (\( a=0.4 \), \( \gamma=0.9 \))

3. Introduction of Speech Production:
3.1 Vocal Model:

Modeling process is usually divided into two parts: the excitation source of larynx filter and the larynx filter (or vocal tract).

3.1.1 the Excitation Source of Larynx Filter:

The airstream from the lungs passes through the glottis to the vocal tract (larynx filter). 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 airstream 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 airstream, generating the characteristic “hissing sound of whispering” [85]. In voiceless phonation, the area of the glottis will be larger and the airstream is only slightly turbulent when it enters the vocal tract.

Voicing is 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 airstream bursts into pharynx with a speed of 2-5 m/s.

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

3.1.2 the Larynx Filter (or 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.
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:

$$F_n = \frac{(2n - 1)c}{4l},$$  \hspace{1cm} (25)

Where $F_n$ is the $n$th formant frequency [Hz], $c$ is the speed of sound in air [m/s], and $l$ is the total length of the tube [m].

In result the speech signal formed with multiplying the excitation function in the larynx filter (or vocal tract) and with refer to this signal we can extracted information from the larynx filter (or vocal tract).

All required "Styles" are predefined, and so there is no need to define a new one. Just select the appropriate style with respect to different sections of a paper.

4. Decrease Dimensions:

We know that the length of GDF of each frame is equal to half length of its frame and it is very long for modeling and matching. Therefore we must reduce this length that so-called decreasing dimensions.

4.1 Decrease Dimensions with Using Discrete Cosine Transform:
This way usually used in many context for decreasing dimensions.

In this way first we get discrete cosine transform from analyzed frame and co-called \( c[n] \):

\[
c(n) = \sum_{k=0}^{k-N} t_m(k) \cos(n(2k+1)\pi / N_f)
\]

We choose the first \( n \) coefficient (in this text 20 coefficient) of \( c[n] \) and then apply to neural network.

4.2 Decrease Dimensions with Using Correlation Function:

We explain in section 2 that the speech voicing signal is generated from multiplying the impulses train by the larynx filter therefore the voicing speech signal is the impulse response of the larynx filter. And for speaker recognition we often use from voicing signal.

It is our idea that two type of information is in various frames of speech signal, first the information about the phoneme that is speaked in the frame and seconds the information about the speaker that is speaking in the frame.

The first type of information in various frames is different because the phoneme that is spoke 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 with extraction of similarity between frames we only extract the same information between frames and remove the various information between frames and because the same information between frames is only about speaker therefore the way may be very good.

For extracting the similarity between frames we use from correlation function.

4.2.1 Definition of Correlation Function:

We know that the correlation coefficients computed the correlation of one signal and another signal and the autocorrelation coefficients are computed the correlation of signal with it.

The following equation is the one dimensional time continuous correlation function:

\[
R_i(a) = \int f(x).f(x+a)dx
\]

This equation in discrete time given by:

\[
R_i(a) = \sum_{-\infty}^{+\infty} f[x].f[x+a]
\]

The "a" variable determines the direction of correlation.

If \( a=4 \) the signal will be multiplication with its four time shifted version.

The following equation is the 2 dimensional correlation functions.

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

In this equation the matrix will be multiplication with it's a rows and b Columns shifted version.

4.2.2 Computing the 20 Correlation Coefficients from One Frame and next Frame (Auto1):

In this way first we get Discrete cosine transform from analyzed frame and so-called \( f[n] \)(n=1 to N/2 and N is the FFT order of signal):

We extract the \( Rf[0] \) to \( Rf[19] \) from following equation:

\[
Rf(a) = \sum_{-\infty}^{+\infty} f[x].f[x+a]
\]
First we compute the correlation coefficients of first frame and second frame, then compute the correlation coefficients of second frame and 3rd frame and ... Finally the number of vectors in the Auto1 is equal to the number of frames. For computed the final vector we use from the final frame and the first frames.

4.2.2 Computing the 20 Autocorrelation Coefficients from One Frame and next 16 Frame (Auto2):

First create a matrix that is contained one frame and next 16 frames:

![Matrix for computing the first vector of Auto1](image1)

Fig. 15: The matrix for computing the first vector of Auto1

Then with using of two dimensional autocorrelation function computed $R[0, 0]$ to $R[0, 19]$. Then we create another matrix that is contained second frame and next 16 frames:

![Matrix for computing the second vector of Auto1](image2)

Fig. 16: The matrix for computing the second vector of Auto1

3891
And extract the autocorrelation coefficients and do this again and again.
In this way for computing the final vectors we use the first 17 frames.

5. Experimental Setup:
5.1 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.

5.2 the Neural Network:
The used neural network in this project is three layers back propagation network that in first layer has 6 neurons, in hidden layer has 15 neurons and in output layer has 20 neurons that is equal to number of speakers.
This network has 20 inputs.
The used function in first layer and hidden layer is tansig that is between -1 and 1.
The used function in output layer is logsig that is between 0 and 1.
Applying 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. The speaker with higher probability is winner in the frame.

5.3 Training:
First apply feature vector of each frames at input of network and the network trained using the desired output.
The selected algorithm for training this network is LM.

5.4 Test:
First, we divide one sentence of one speaker into many frames then extract a feature vector from each frame then, apply each feature vector 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.
Now we have one matrix for one sentence that contains the probability of each speaker and with this matrix we can know whom speaker is speaking in the sentence.

5.5 Results:
The simulation results illustrate in following table:

<table>
<thead>
<tr>
<th>Number of speakers</th>
<th>feature</th>
<th>Recognition%</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>GDF</td>
<td>85</td>
</tr>
<tr>
<td>20</td>
<td>GDFAuto1</td>
<td>78.3333</td>
</tr>
<tr>
<td>20</td>
<td>GDFAuto2</td>
<td>91.6667</td>
</tr>
</tbody>
</table>

The recognition percent in this table is low because the simulation is text independent.
In text independent, the network training and testing data base is various.
Usually, in the text independent simulation the recognition percent is lower than text dependent simulation.
Finally, we can see in the table that the recognition percent of the GDFAuto2 is better than other analysis.

6. Summary and Concluding Remarks
In this text we describe a method for decreasing the length of a analyzed frame using of extracting similarity between one frame and next frames.
We use from correlation function and extract the correlation coefficient in various direction.
Note that this method can be used for every analysis and only extracts the information of speaker and remove 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.
Fig. 16: The testing process

REFERENCES

Adjoudj Réda and Boukelif Aoued, 2005. "Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition".

Adjoudj Réda & Boukelif Aoued, 2005. "Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition".

Antanas LIPEIKA, Joana LIPEIKIEN ꞉ E, Laimutis TELKSNYS. "Development of Isolated Word Speech Recognition System". September 2001


Hat Yai, Songkhla, " MODIFIED MEL-FREQUENCY CEPSTRUM COEFFICIENT".


Muzhir Shaban Al-Ani and Thabit Sultan Mohammed and M. Aljebory, 2007 "Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform".


Yiyin Zhou . "Speaker Identification ".