# Robustness of Phase based Features for Speaker Recognition

*R.Padmanabhan*[1], *Sree Hari Krishnan Parthasarathi*[2,3], *Hema A. Murthy*[1]

[1]Department of Computer Science and Engineering,
Indian Institute of Technology Madras,
Chennai, India
[2] Idiap Research Institute, Martigny, Switzerland
[3]École Polytechnique Fédérale de Lausanne, Switzerland

padmanabhan@lantana.tenet.res.in, hema@lantana.tenet.res.in
Hari.Parthasarathi@idiap.ch

## Abstract

This paper demonstrates the robustness of group-delay based features for speech processing. An analysis of group delay functions is presented which show that these features retain formant structure even in noise. Furthermore, a speaker verification task performed on the NIST 2003 database show lesser error rates, when compared with the traditional MFCC features. We also mention about using feature diversity to dynamically choose the feature for every claimed speaker.

**Index Terms**: group delay functions, speaker verification

## 1. Introduction

A crucial task in the development of automatic speaker recognition systems is choosing a parametric representation for speech signals which is robust to mismatches in the training and testing conditions. In real world data, there are high levels of variation in the speech signals the system typically encounters. The sources of variability include intra-speaker variations (due to health, emotional state etc. of the speaker) and also channel variability (due to the use of a different telephone handset, microphone or environment the speaker is speaking from.)

Popular parametric representations of speech are based on cepstral representations of the magnitude spectrum. In recent speaker recognition evaluations conducted by NIST, the Mel-frequency cepstral coefficients (MFCC) are a commonly used feature. The MFCCs are derived from the short time magnitude spectrum of speech. The spectral representation of speech is complete only when the magnitude and phase spectra are specified. Study of phase-based parametrisation of speech has resulted in several representations including the modified group delay feature (MODGDF) [3].

The objective of this paper is to demonstrate, both analytically and experimentally, that group delay based features are robust to additive noise.

## 2. Group delay processing of speech

The group delay function $\tau(\omega)$ of a signal $x(n)$ can be computed directly from the signal as follows [7]:

$$\tau(\omega) = -\text{Im}\left(\frac{d}{d\omega}\log(X(\omega))\right) \quad (1)$$

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

where Im denotes the imaginary part, $x(n) \leftrightarrow X(\omega)$ and $y(n) \leftrightarrow Y(\omega)$ are Fourier transform pairs and $y(n) = nx(n)$.

Several studies have been done on the effectiveness of group delay representations of speech signals [2] [3] [4]. The group delay function is ill behaved when there are zeros near the unit circle. To mitigate the effect of these zeros, the modified group delay function is defined as [3]:

$$\tau_m(\omega) = \left(\frac{\tilde{\tau}_m(\omega)}{|\tilde{\tau}_m(\omega)|}\right)(|\tilde{\tau}_m(\omega)|)^{\alpha} \quad (3)$$

where

$$\tilde{\tau}_m(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X_c(\omega)|^{2\gamma}} \quad (4)$$

The MODGDF is decorrelated by using the discrete cosine transform (DCT).

## 3. Analysis of group delay in noise

In this section, we analytically show why group delay functions are robust to noise.

Let $x[n]$ denote a clean speech signal degraded by uncorrelated, zero-mean, additive noise $v[n]$. Then, the noisy speech, $y[n]$, can be expressed as,

$$y[n] = x[n] + v[n] \quad (5)$$

Taking the Fourier transform, we have

$$Y(\omega) = X(\omega) + V(\omega) \quad (6)$$

Multiplying by corresponding complex conjugates and taking the expectation, we have the power spectrum

$$P_Y(\omega) = P_X(\omega) + \sigma^2(\omega) \quad (7)$$

where we have used the assumption that the expectation of noise is zero. The power spectra of the resulting noisy speech signal can be related to noise power and (clean) speech power in one of three mutually exclusive frequency regions: (i) the high noise power case where $P_X(\omega) \ll \sigma^2(\omega)$ (ii) the high signal power case where $P_X(\omega) \gg \sigma^2(\omega)$ and (iii) the equal power case where $P_X(\omega) \approx \sigma^2(\omega)$. The power spectra of the noisy speech signal in each case are denoted respectively as $P_Y^n(\omega)$, $P_Y^s(\omega)$ and $P_Y^e(\omega)$. We analyse the group delay representation of noisy speech in the three cases mentioned above.

### 3.1. High noise power spectral regions ($P_Y^n(\omega)$)

In this subsection, we consider frequencies $\omega$ such that $P_X(\omega) \ll \sigma^2(\omega)$, i.e., regions where the noise power is higher than signal power. From Equation 7 we have

$$\begin{aligned} P_Y^n(\omega) &= P_Y(\omega) \quad \forall\,\omega \quad \text{s.t.} \quad P_X(\omega) \ll \sigma^2(\omega) \\ &= P_X(\omega) + \sigma^2(\omega) \\ &= \sigma^2(\omega)\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right) \end{aligned}$$

Taking logarithms on both sides, using the Taylor series expansion[1] of $\ln(1 + \frac{P_X(\omega)}{\sigma^2(\omega)})$, and ignoring the higher order terms,

$$\begin{aligned} \ln\left(P_Y^n(\omega)\right) &= \ln\left[\sigma^2(\omega)\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right)\right] \\ &= \ln\left(\sigma^2(\omega)\right) + \ln\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right) \\ &\approx \ln\left(\sigma^2(\omega)\right) + \frac{P_X(\omega)}{\sigma^2(\omega)} \qquad (8) \end{aligned}$$

Expanding $P_X(\omega)$ as a Fourier series ($P_X(\omega)$ is a periodic, continuous, function of $\omega$ with a period $\omega_0 = 2\pi$),

$$\ln\left(P_Y^n(\omega)\right) \approx \ln\left(\sigma^2(\omega)\right) + \frac{1}{\sigma^2(\omega)}\left[\frac{d_0}{2} + \sum_{k=1}^{\infty} d_k \cos\left(\frac{2\pi}{\omega_0}\,\omega\, k\right)\right] \qquad (9)$$

where, $d_k$ are the Fourier series coefficients in the expansion of $P_X(\omega)$. Since $P_X(\omega)$ is an even function, coefficients of the sine terms are zero.

For a minimum phase signal, the group delay function can be computed in terms of the cepstral coefficients of the log-magnitude spectrum, as given in [4],

$$\begin{aligned} \log|X(\omega)| &= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(\omega\, k) \\ \tau(\omega) &= \sum_{k=1}^{\infty} k\, a_k \cos(\omega\, k) \qquad (10) \end{aligned}$$

where, $\tau$ is the group delay function and $a_k$ are the cepstral coefficients. From (10), it can be observed that the group delay function can be obtained from the log-magnitude response by ignoring the dc term, and by multiplying each coefficient with $k$. Applying this observation to Equation (9), we get the group delay function as:

$$\tau_{Y^n}(\omega) \approx \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \qquad (11)$$

This expression shows that the group delay function is inversely proportional to the noise power ($\sigma^2(\omega)$) in regions where noise power is greater than the signal power.

### 3.2. High signal power spectral regions ($P_Y^s(\omega)$)

Now consider frequencies $\omega$ such that $P_X(\omega) \gg \sigma^2(\omega)$. Starting with Equation (7), and following the steps similar to those in previous subsection:

$$\ln\left(P_Y^s(\omega)\right) \approx \ln\left(P_X(\omega)\right) + \frac{\sigma^2(\omega)}{P_X(\omega)} \qquad (12)$$

Since $P_X(\omega)$ is non-zero, continuous, and periodic in $\omega$, $\frac{1}{P_X(\omega)}$ is also periodic and continuous. Consequently, $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ can be expanded using Fourier series, giving

$$\ln\left(P_Y^s(\omega)\right) \approx \frac{d_0 + \sigma^2(\omega)\, e_0}{2} + \sum_{k=1}^{\infty}\left(d_k + \sigma^2(\omega)\, e_k\right)\cos(\omega\, k)$$

Using the properties of group delay function listed in Equation (10), and following the steps in the previous case[2], we obtain the expression for the group delay function as,

$$\tau_{Y^s}(\omega) \approx \sum_{k=1}^{\infty} k\left(d_k + \sigma^2(\omega)\, e_k\right)\cos(\omega\, k) \qquad (13)$$

where $d_k$ and $e_k$ are the Fourier series coefficients of $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ respectively. It is satisfying to observe that if $\sigma^2(\omega)$ is negligible, the group delay function can be expressed solely in terms of log-magnitude spectrum.

### 3.3. Signal power $\approx$ noise power regions ($P_Y^e(\omega)$)

For frequencies $\omega$ such that $P_X(\omega) \approx \sigma^2(\omega)$, we again start with Equation (7), and follow the steps similar to those in previous subsections, except in this case we do not need the Taylor series expansion:

$$\begin{aligned} P_Y^e(\omega) &\approx 2P_X(\omega) \\ \ln\left(P_Y^e(\omega)\right) &\approx \ln 2 + \ln\left(P_X(\omega)\right) \qquad (14) \end{aligned}$$

Expanding $\ln\left(P_X(\omega)\right)$ as a Fourier series, since it is a periodic, continuous, function of $\omega$ with a period $2\pi$, the group delay function can be computed as,

$$\tau_{Y^e}(\omega) \approx \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \qquad (15)$$

where $d_k$ are the Fourier series coefficients of $\ln(P_X(\omega))$.

### 3.4. Behaviour of minimum phase group delay functions in noise

From Equations 11, 13, and 15, the estimated group delay functions are summarised respectively for the three cases:

$$\tau(\omega) \approx \begin{cases} \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \\ \sum_{k=1}^{\infty} k\left(d_k + \sigma^2(\omega)\, e_k\right)\cos(\omega\, k) \\ \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \end{cases} \qquad (16)$$

where the first case is for $\forall\omega$ such that $P_X(\omega) \ll \sigma^2(\omega)$, the second for $\forall\omega$ such that $P_X(\omega) \gg \sigma^2(\omega)$, and the third for $\forall\omega$ such that $P_X(\omega) \approx \sigma^2(\omega)$. From Equation 16, we note that the group delay function of a minimum phase signal is *inversely* proportional to the noise power for frequencies corresponding to high noise regions in the power spectrum. Similarly, for low noise regions, from Equation 13, the group delay function becomes *directly* proportional to the signal power. In other words, its behaviour is similar to that of the magnitude spectrum. This shows that the group delay function of a minimum phase signal preserves the peaks and valleys in the magnitude spectrum well even in the presence of additive noise.

[1] Taylor series expansion of $\ln(1 + x)$ is: $\ln(1 + x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} x^{n+1} \qquad |x| < 1$

[2] Ignoring the dc term, and multiplying each coefficient with $k$

### 3.5. The modified group delay function

Practically, a frame of speech is typically non-minimum phase, due to the zeros introduced by nasals, pitch and the analysis window. Thus, the above analysis is directly applicable only to the minimum phase components derived from speech signals. To overcome this, we use the modified group delay (MODGD), which is an approximation to the minimum phase group delay. Using the modified group delay enables computation of the group delay even when the signal is not minimum phase [3].

### 3.6. The modified group delay feature

The modified group delay feature or MODGDF (also called modified group delay cepstra) is formed by converting the modified group delay (MODGD) into cepstral features using the discrete cosine transform [3]. This results in features that are linearly decorrelated. When compared to MODGD features, MODGDF features can be of considerably lower dimension.

## 4. Experimental verification of robustness

### 4.1. Speaker recognition system

For experimental evaluation of MODGDF features, a speaker detection task is performed on the NIST 2003 SRE dataset. Gaussian mixture models [5] were used to model the target speaker models and the background models.

For each frame of speech, the MODGDF feature is extracted as given in [3].

### 4.2. Performance analysis

To compare results, the MODGDF-based speaker recognition system is evaluated against a conventional MFCC-based system. The DET curves for this are shown in Figure 1.

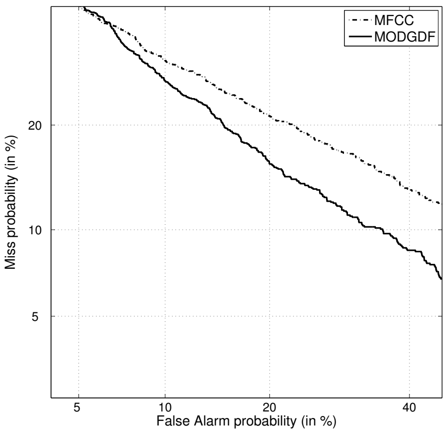

Figure 1: *DET plots for MFCC and MODGDF based systems. MODGDF shows better performance in most operating points.*

The target and imposter score distributions for MFCC and MODGDF features are shown in Figures 2 and 3 respectively. The likelihood ratio scores from every test are pooled for all target speakers as is done in the NIST SREs [6]. The MODGDF scores show narrower variances than the MFCC scores, resulting in better separability between target and imposter scores.

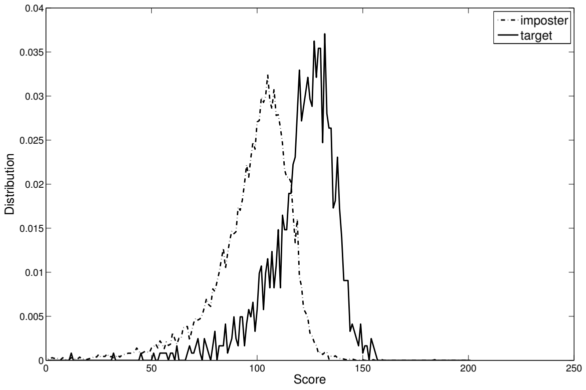

Figure 2: *Target and imposter score distributions for MFCC.*

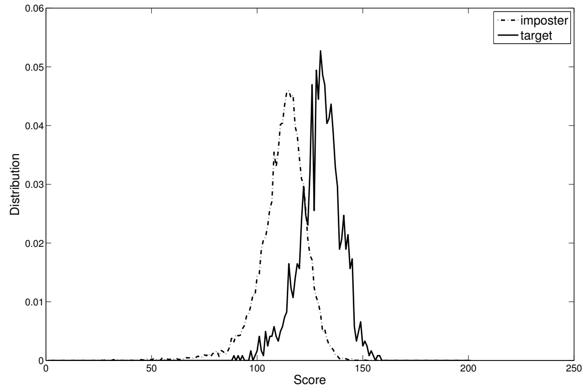

Figure 3: *Target and imposter score distributions for MODGDF.*

From these results, we conclude that phase-based features like MODGDF more accurately model speakers in noisy conditions, and are more robust to channel effects. The DET curves indicate that at most operating points (except at low flase alarm probablities), MODGDF features give better performance. Furthermore, score normalization techniques like ZNorm and HNorm can be applied to further reduce verification errors, but these experiments are not done in this paper.

## 5. Dynamic feature switching for speaker verification

After analysing the results of a closed-set speaker identification task on the NIST 2003 database, it was observed that some speakers are consistently accurately identified by MFCC, whereas others were identified by MODGDF. The feature (in this case, MFCC or MODGDF) which more accurately identifies a speaker is known as the *optimal feature* for that speaker. Similarly, the other feature is known as the non-optimal feature for that speaker. This can be made use of advantageously for a speaker verification task by *dynamically* choosing the better feature based on the speaker claim. For instance, if we know *a priori* that the claimed speaker has MFCC as the optimal feature, we perform verification with MFCC features. On the other hand, if MODGDF is the optimal feature, we perform verification with MODGDF features.

Based on this principle, the verification task was repeated incorporating feature switching. To compare results, the verification task was also done using the non-optimal feature. The score distributions for both optimal and non-optimal features are shown in Figures 4 - 7.

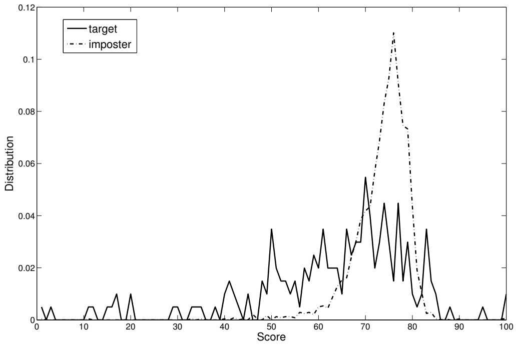

Figure 4: *Target and imposter score distributions for speakers with optimal MFCC feature.*

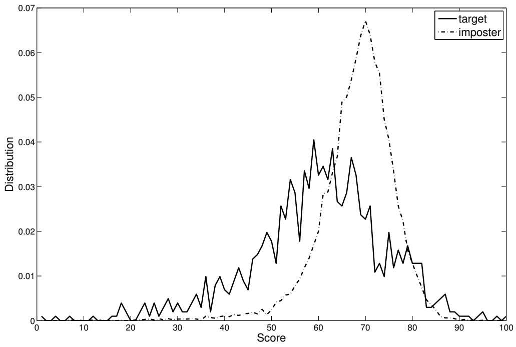

Figure 5: *Target and imposter score distributions for speakers with optimal MODGDF feature.*

The score distributions clearly show more separability with optimal features than with non-optimal features. This *feature diversity* can be used to improve performance of speaker verification systems. Methods of arriving at optimal features for each speaker is being investigated with measures like mutual information and KL-divergence.

## 6. Conclusion

In this paper, we demonstrated analytically that the phase-based modified group feature is robust to additive noise. A speaker verification task on the NIST 2003 dataset resulted in better performance for MODGDF features when compared to conventional MFCC features. Also, we looked into the concept of feature switching to always use a claimed speaker's optimal feature while performing recognition, resulting in better performance.

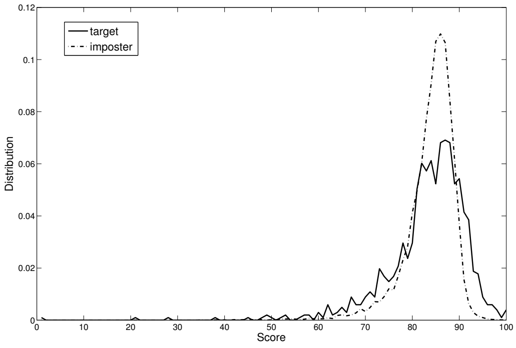

Figure 6: *Target and imposter score distributions for speakers with non-optimal MFCC feature.*

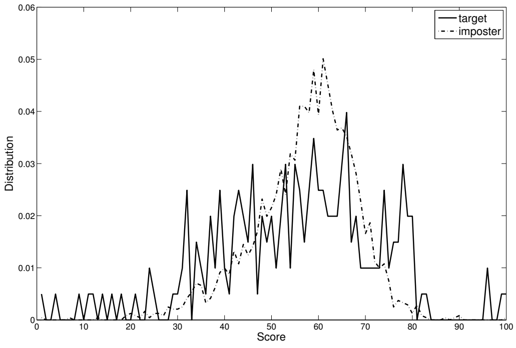

Figure 7: *Target and imposter score distributions for speakers with non-optimal MODGDF feature.*

## 7. References

[1] P. Mermelstein and S. B. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Proc., 28:357-366, 1980.

[2] H. A. Murthy and V. R. R. Gadde, "The modified group delay and its application to phoneme recognition", Proceedings of ICASSP, 1:68-71, 2003.

[3] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition", IEEE Trans. Audio, Speech and Language Proc., 15:190–202, 2007.

[4] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase", IEEE Trans. on Acoustics, Speech and Signal Proc., 32:610–622, June 1984.

[5] D. Reynolds and R. Rose, "Robust text independent speaker identification using Gaussian mixture speaker models", IEEE Trans. on Speech and Audio Proc., 3:72–82, 1995.

[6] "The NIST Year 2003 Speaker Recognition Evaluation Plan", http://www.itl.nist.gov/iad/mig/tests/sre/2003/index.html, 2002

[7] A. V. Oppenheim and R. W. Schafer, "Discrete-time Signal Processing", Prentice-Hall, 2000.