Chirp Group Delay based Feature for Speech Applications

Malarvizhi Muthuramalingam[1], Anushiya Rachel Gladston[1], P Vijayalakshmi[2], T Nagarajan[1]
[1]Department of CSE, Shiv Nadar University Chennai, India
[2]Department of ECE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
malarvizhim@snuchennai.edu.in, anushiyarachelg@snuchennai.edu.in,
vijayalakshmip@ssn.edu.in, nagarajant@snuchennai.edu.in

Abstract

Conventional Fast Fourier Transform (FFT), computed on the unit circle, gives an accurate representation of the spectrum if the signal under consideration is because of the sustained oscillations. However, practical signals are not sustained oscillations. For the signals that are either decaying/growing along time, the phase spectrum computed using conventional FFT is not accurate, and in turn, the magnitude spectrum too. Hence a feature, based on a variant of the group delay spectrum, namely the chirp group delay (CGD) spectrum, is proposed. The efficacy of the proposed feature is evaluated in Gaussian Mixture Model (GMM) and Convolutional Neural Network (CNN)-based speaker identification systems. Analysis reveals a significant increase in performance when using the CGD-based feature over the magnitude spectrum.

1 Introduction

The characteristics and behaviour of speech signals are often inferred from their features. Various well-known features include Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Discrete Wavelet Coefficient (Tüfekci and Gowdy , 2000), and modified group delay (Hegde et al. , 2007). However, recent neural approaches to speech processing predominantly involve the use of linear and Mel spectrograms. These are derived from the FFT magnitude spectrum, which has a multiplicative property and so closely spaced poles cannot be easily differentiated/resolved. To address this frequency resolution issue, the group delay spectrum, derived from the negative derivative of the phase spectrum, could be used instead to achieve better resolution.

The properties of both the phase and group delay spectra are discussed in (Murthy and Yegnanarayana , 2011). Since the group delay spectrum is additive in nature, poles and zeros are well-resolved as peaks and valleys (Nagarajan et al. , 2003). It has also been observed that a perfect minimum phase signal, derived from a one-sided inverse Fourier magnitude spectrum, as described in (Berkhout , 1973), is energy bounded, and only when the causal region is considered, this function becomes uneven, resulting in a net phase change of zero ($\theta[\pi]-\theta[0]$). Consequently, the group delay spectrum derived from this minimum phase signal behaves similar to the magnitude spectrum (Nagarajan et al. , 2003) and can be used in spectrum estimation (Yegnanarayana and Murthy , 1992). In this minimum phase group delay function, poles and zeros can be distinguished easily, where peaks correspond to poles while valleys correspond to zeros (Nagarajan et al. , 2004).

The phase spectrum, which is naturally wrapped between $-\pi$ and π, can be used in various speech applications, including speech recognition systems and speech enhancement systems (Paliwal et al. , 2011). However, the location of the poles on the unit circle (Paraskevas and Rangoussi , 2012) leads to discontinuity in phase wrapping, which poses challenges in feature extraction. Conventionally, the FT is computed on the unit circle. In such a case, if some singularities, especially zeros due to windowing, lie on the unit circle, the resultant group delay spectrum becomes spiky (Sripriya and Nagarajan , 2015). To address these limitations, the chirp group delay spectrum may be used, which is computed at a radius that is not equal to 1, and this case, at a radius greater than 1 as shown in (Bozkurt , 2007). This method is employed in (Gladston et al., 2015), where the speech signal is assumed to be the FT spectrum of an arbitrary signal, and the group delay spectrum of the arbitrary signal measured at a radius greater than 1 is used to estimate the location of zeros lying outside the unit circle. Previous studies by (Joysingh et al. , 2025) have also shown that chirp magnitude spectrum-based MFCC outperforms the conventional MFCC. In addition to retaining the error in phase spectrum (and magnitude spectrum) because of using chirp MFCC, if a feature derived from gd is also considered, the frequency resolution issue may also be

reolved. Therefore, building on this, the current work proposes a chirp group delay-based feature.

The paper is organized as follows: Section 2 describes the issues with group delay(GD) and the significance of chirp group delay spectrum(CGD). Section 3 presents the experimental setup, the corpus used, and the performance analysis and comparison with conventional magnitude and group delay spectra, and Section 4 summarizes the conclusions drawn.

2 Chirp Group Delay Spectrum

The group delay is defined as the negative derivative of the phase spectrum. As it is derived from the phase, it is additive in nature. As a result, it exhibits the higher resolution than magnitude spectrum, as discussed in the previous section. However, the group delay spectrum appears spiky when there are zeros on the unit circle. Further, since speech signals are not sustained oscillations, the phase spectrum derived from the conventional FT, measured at a radius of 1, that is on the unit circle, may not be accurate, as described in (Joysingh et al. , 2025). Therefore, to address these issues, the chirp FT, measured at a radius not equal to 1, and hence the chirp group delay spectrum may be derived.

The chirp spectrum is defined as

$$X(\omega) = \sum_{n=0}^{N-1} \left[r_c^{-n} \cdot x(n)\right] e^{-j\omega n} \tag{1}$$

where r_c is the chirp radius, $x(n)$ is the input signal, N is the length of the input sequence, and ω is the angular frequency.

2.1 Mathematical Interpretation of Chirp Group Delay

If some of the singularities, particularly the zeros due to windowing, lie on the unit circle, the resulting group delay spectrum becomes spiky, with values around $-\pi$. Similarly, for zeros located outside the unit circle, the group delay spectrum exhibits a valley, with values around -2π, as evident from (Sripriya and Nagarajan , 2015), which is discussed as follows:

Consider a system, $H(z)$, with a single zero at an angular location, ω_0, as given below:

$$H(z) = 1 - az^{-1} \tag{2}$$

where $a = re^{j\omega_0}$, and r is the radius of the zero in the z-plane. The corresponding Fourier transform can be expressed as:

$$H(e^{j\omega}) = 1 - r\cos(\omega - \omega_0) + jr\sin(\omega - \omega_0) \tag{3}$$

At the angular location of the zero, $\omega = \omega_0$:

$$H(e^{j\omega_0}) = 1 - r \tag{4}$$

When $r > 1$, $H(e^{j\omega_0})$ becomes negative, implying that when a zero lies outside the unit circle, the Fourier transform at the frequency bin $\omega = \omega_0$ may have a negative value.

Considering again a system with a single zero at the angular location ω_0 with radius r. The frequency bins above and below ω_0 are ω_1 and ω_2, where $\omega_1 = \omega_0 - \delta$ and $\omega_2 = \omega_0 + \delta$, with δ being very small. Using equation (3), the Fourier transform of the system at ω_1 is given by:

$$H(e^{j\omega_1}) = 1 - r\cos\delta + jr\sin\delta \tag{5}$$

From the above equation, the phase at ω_1 can be expressed as:

$$\theta(e^{j\omega_1}) = \tan_4^{-1}\left(\frac{-r\sin\delta}{1 - r\cos\delta}\right) \tag{6}$$

Similarly, the phase at ω_2 is:

$$\theta(e^{j\omega_2}) = \tan_4^{-1}\left(\frac{-r\sin\delta}{1 - r\cos\delta}\right) \tag{7}$$

where the phase is a four-quadrant inverse tangent function. The group delay function at the location of the zero can be expressed as:

$$\tau = \theta(e^{j\omega_1}) - \theta(e^{j\omega_2}) \tag{8}$$

When the order of the Fourier transform is high, the difference between adjacent frequency bins δ is very small. Therefore, the group delay function is given by:

$$\tau = -2\tan_4^{-1}\left(\frac{-r\delta}{1 - r\delta}\right) = \tau_c \tag{9}$$

The above group delay function is referred to as the conditional group delay function.

When a zero lies outside the unit circle, $1 - r$ is negative. This implies the denominator of the fourth quadrant inverse tangent function in equation (9) is negative, while the numerator is positive. In this case, equation (9) becomes:

$$\tau_c = -2\left[\tan_2^{-1}\left(\frac{-r\delta}{1 - r\delta}\right) + \pi\right] \tag{10}$$

For δ being very small and $r > 1$, $\tan_2^{-1}\left(\frac{-r\delta}{1-r\delta}\right)$ is very small and negative. Therefore, $\tau_c \approx -2\pi$ at the angular location of the zero. This is due to phase wrapping. Similarly, for a zero on the unit circle, i.e., if $r = 1$, the denominator is zero with the numerator positive, leading to $\tau_c \approx -\pi$. To overcome these issues, we measure the chirp group delay with $r > 1$. The equations are cited from (Gladston et al., 2015).

2.2 Effects of the Location of Poles in Chirp Group Delay

In order to use the group delay spectrum that yields the information provided by the magnitude spectrum, but with a better resolution, the signal should be a minimum phase signal. Therefore, instead of directly computing the group delay spectrum from the speech signal, it is derived from the inverse FT of the causal portion of the magnitude spectrum. As discussed earlier, it would be desirable to measure the FT and the group delay at a radius that is not equal to 1. In order to understand the impact of measuring the FT of a minimum phase signal, which is inherently decaying, consider three scenarios: (i) $r_c = 1$, (ii) $r_c > 1$ and (iii) $r_c < 1$.

- When $r_c = 1$ (conventional FT), $r_c^{-n}x(n)$ in equation 1 would result in a decaying signal, since $x(n)$ is decaying and hence result in an inaccurate phase spectrum as discussed earlier.
- When $r_c > 1$, since r_c^{-n} and $x(n)$ are both decaying, the phase spectrum would be more inaccurate than that derived when $r_c = 1$.
- When $r_c < 1$, r_c^{-n} is growing and will therefore compensate the decay in $x(n)$. Therefore, the chirp phase/group-delay spectrum measured will be more accurate.

In the group delay spectrum of a minimum phase signal, both the peaks and valleys represent poles and zeros, respectively. However, in the case of non-minimum phase signals, the zeroes outside the unit circle, instead of showing up as valleys, appear as peaks at the corresponding angular frequencies. So if the chirp radius is less than or equal to one ($r_c \leq 1$), the chirp group delay will match the conventional group delay spectrum. However, if the radius is greater than one ($r_c > 1$), the signal becomes maximum phase, and adding a negative sign lead to an inverted group delay spectrum.

Based on these observations, the proposed chirp group delay spectrum, derived from the minimum phase signal, measures poles with a radius greater than 1 and applies a negative sign to the group delay. Therefore, $x(n)$ is derived to be a minimum phase signal and r_c^{-n} is modified as r_c^n.

The chirp group delay spectrum can be defined in two ways: either (i) by multiplying the basis function with the exponential component r_c^n and then computing the conventional group delay from the minimum phase signal, or (ii) by multiplying the original speech signal by the exponential component r_c^n, to convert it to a minimum phase signal, and then computing the group delay. The second approach modifies the signal instead of computing the FFT at a radius, $r_c \neq 1$, in such a case, existing FFT algorithm can still be used.

2.3 Steps involved in deriving the chirp group delay spectrum

The steps involved in computing chirp group delay spectrum, for an input speech signal, $x(n)$ are as follows:

- Compute the short-time Fourier transform (STFT) based magnitude spectrum of $x(n)$ using overlapping windows.
- Compute the inverse discrete Fourier transform (IDFT) of the magnitude spectrum.
- Consider the causal portion of the signal derived in the previous step to obtain a minimum phase signal.
- Compute the chirp FT for this signal, with a radius, $r_c > 1$ (here $r_c = 1.00005$).
- Compute the chirp group delay spectrum by taking the negative derivative of the phase spectrum.
- Convert the chirp group delay spectrum to the Mel scale.
- Compute the Discrete Cosine Transform (DCT) of the Mel scaled chirp group delay spectrum directly and use it as a feature for the system.

3 Performance Analysis

In order to assess the efficacy of the proposed feature, speaker identification systems are trained with the existing and proposed features and their performances are analyzed. The speech corpus used for training the speaker identification systems, the

models and features used, and the results of the experiments are described below.

3.1 Speech corpus

The VCTK dataset (Yamagishi et al. , 2019) contains recordings from 101 English speakers with various accents, each lasting 2-6 seconds. From the dataset, 86 speakers have been selected to assess the performance of the proposed feature. Each speaker has 325 utterances. Out of which 300 utterances from each speaker are used for training and 25 utterances from each speaker are used for testing. The sentences are drawn from the Rainbow Passage and an elicitation paragraph. All speech data was recorded using the same recording setup, using an omni-directional microphone.

3.2 Experimental Setup

For the experiment, the proposed feature is evaluated in a speaker identification system using both statistical and neural network-based methods. In the GMM-based speaker recognition system, the proposed chirp group delay feature is trained using 64 mixture components and a chirp radius of 1.00005. Convolutional Neural Network (CNN) is also trained using the proposed feature. The model consist of 1D convolutional layer (Conv1D) that applies 64 filters, each with a kernel size of 3, and ReLU activation. This is followed by a MaxPooling1D layer to reduce the dimensionality of the feature maps. The second convolutional layer uses 128 filters and a kernel size of 3, followed by another max-pooling layer. The resulting feature maps are flattened into a single vector, which is then passed to fully connected (dense) layers. A dense layer with 128 units and ReLU activation is applied, followed by an output layer using SoftMax activation to predict class probabilities, with the number of classes corresponding to the number of speakers. The Adam optimiser and sparse categorical cross-entropy loss are used, with accuracy as the primary evaluation metric.

3.3 Comparitive Analysis

The speaker recognition system trained with the proposed chirp group delay spectrum as a feature is compared with two other systems: one trained using a conventional magnitude spectrum and another trained with a group delay spectrum derived from a minimum-phase signal. From Table 1, it is observed that in both the systems, the chirp group delay spectrum performs marginally better than the group delay spectrum, but significantly better than the conventional magnitude spectrum, due to the additive property of group delay and the significance of the chirp FT. With the CNN model, the proposed feature yields a performance that is significantly better than the magnitude spectrum and marginally better than the group delay spectrum.

Table 1: GMM and CNN-based speaker identification accuracy

Feature /Model	**Chirp GD-based MFCC**	**GD-based MFCC**	**Magnitude spectrum-based MFCC**
GMM	91%	90%	81%
CNN	89.9%	89.03%	83.50%

To further explore the impact of the chirp radius on system performance, a detailed analysis was carried out by varying the chirp radius values. Using 1.001 as the reference point, additional experiments were conducted with chirp radii of 1.002, 1.0001, and 1.00005. Fig. 1 reveals the performance of the GMM and CNN-based speaker identification systems at different chirp radii. The graphs clearly demonstrate that as the chirp radius approaches the unit circle, the resolution of the features improves, resulting in better speaker recognition accuracy for both models.

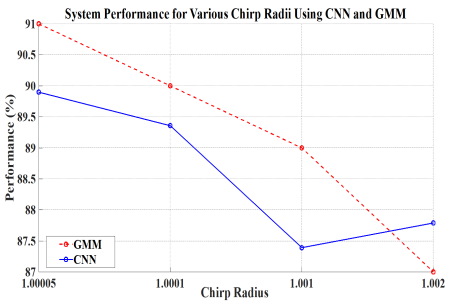

Figure 1: System performance different chirp radii using GMM and CNN

4 Conclusion

In summary, these experiments confirm that the suggested chirp group delay spectrum is an extremely useful feature for speaker identification. The property of high frequency resolution and the effect of measuring the FT of a decaying signal outside the unit circle, make this feature highly effective compared to conventional features. The

variation in chirp radius has a notable influence on the system's accuracy, with radii closer to the unit circle yielding the best results which is marginally better than the group delay spectrum, but significantly better than the conventional magnitude spectrum. Thus, the chirp group delay spectrum offers a promising approach for future advancements in speech technology.

5 Limitations

While computing the chirp group delay based-MFCC for a large dataset, the term r^{-n} increases the computational complexity compared to conventional MFCC approaches. This added complexity may impact the processing time and resource requirements, making it less efficient for real-time applications or large-scale implementations. Furthermore, optimizing this aspect of the algorithm is necessary to balance the trade-off between improved accuracy and computational cost.

6 Ethical Considerations

The current work complies with the ACL ethics policy. All data used in this study was sourced from publicly available datasets with appropriate permissions for research use.

References

Aarabi, P., Shi, G., Shanechi, M., and Rabi, S., *Phase-Based Speech Processing*, World Scientific Publishing Co., 2005.

Ahmed, N., Natarajan, T., and Rao, K. R., *Discrete Cosine Transform*, IEEE Transactions on Computers, vol. C-23, no. 1, pp. 90–93, 1974.

Berkhout, A. J., *On the Minimum-Length Property of One-Sided Signals*, Geophysics, vol. 38, no. 4, pp. 701–709, 1973.

Bozkurt, B., *Chirp Group Delay Analysis of Speech Signals*, Speech Communication, vol. 49, no. 3, pp. 159–176, 2007.

Gladston, A. R., Vijayalakshmi, P., and Nagarajan, T., *Estimation of glottal closure instants from telephone speech using a group delay-based approach that considers speech signal as a spectrum*, Interspeech 2015, 1181–1185, 2015.

Hegde, R. M., Murthy, H. A., and Gadde, V. R. R., *Significance of the Modified Group Delay Feature in Speech Recognition*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 190–202, 2007.

Joysingh, S. J., Vijayalakshmi, P., and Nagarajan, T., *Significance of chirp MFCC as a feature in speech and audio applications*, Computer Speech & Language, vol. 89, p. 101713, 2025.

Murthy, H. A., and Gadde, V., *The modified group delay function and its application to phoneme recognition*, 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), vol. 1, pp. I-68, 2003.

Murthy, H. A. and Yegnanarayana, B., "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, 2011.

Nagarajan, T., Murthy, H. A., and Hegde, R. M., *Group delay based segmentation of spontaneous speech into syllable-like units*, ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003, pp. MAP20.

Nagarajan, T., Prasad, V., and Murthy, H. A., *Minimum phase signal derived from root cepstrum*, Electronics Letters, vol. 39, no. 15, pp. 941–942, 2003.

Nagarajan, T., Prasad, V., and Murthy, H. A., *Automatic segmentation of continuous speech using minimum phase group delay functions*, Speech Communication, vol. 42, no. 4, pp. 429–446, 2004.

Sripriya, N., and Nagarajan, T., *Estimation of glottal closure instants by considering speech signal as a spectrum*, Electronics Letters, vol. 51, pp. 649–651, Apr. 2015.

Paliwal, K, Wójcicki, K and Shannon, B., "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

Paraskevas, I., and Rangoussi, M., *Feature Extraction for Audio Classification of Gunshots Using the Hartley Transform*, Open Journal of Acoustics, vol. 2, no. 3, pp. 145–156, 2012.

Tüfekci, Z., and Gowdy, J. N., *Feature extraction using discrete wavelet transform for speech recognition*, Conference Proceedings - IEEE SOUTHEASTCON, pp. 116–123, 2000.

Yamagishi, J., Veaux, C., and MacDonald, K., *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)* [sound], University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2019.

Yegnanarayana, B., and Murthy, H. A., *Significance of group delay functions in spectrum estimation*, IEEE Transactions on Signal Processing, vol. 40, no. 9, pp. 2281–2289, 1992.