Modified Group Delay Features for Emotion Recognition

Chapter · December 2023

DOI: 10.1007/978-3-031-45170-6_33

Aditya Pusuluri
Dhirubhai Ambani Institute of Information and Communication Technology
3 PUBLICATIONS 2 CITATIONS

Hemant Patil
Dhirubhai Ambani Institute of Information and Communication Technology
289 PUBLICATIONS 2,867 CITATIONS

Modified Group Delay Features for Emotion Recognition

S. Uthiraa, Aditya Pusuluri, and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar Gujarat
{uthiraa_s, aditya_pss, hemant_patil}@daiict.ac.in

Abstract. As technological advancements progress, dependence on machines is inevitable. Therefore, to facilitate effective interaction between humans and machines, it has become crucial to develop proficient techniques for Speech Emotion Recognition (SER). This paper uses phase-based features, namely Modified Group Delay Cepstral Coefficients for SER. To the best of our knowledge, this paper is the first attempt to use the MGDCC feature on emotions. Experiments were performed using the EmoDB database on emotions, anger, happy, neutral, and sad. The proposed feature outperformed the baseline Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) by **7.7 %** and **5.14 %**, respectively. The noise robustness characteristics of MGDCC were tested on stationary and non-stationary noise and the results were promising. The latency period was also analysed and MGDCC proved to be the most practically suitable feature.

Keywords: Speech Emotion Recognition · Narrowband Spectrogram · Group Delay Function · Modified Group Delay · EmoDB · Vocal Tract Features.

1 Introduction

The easiest and most effective way of communication is through speech and the emotional aspect of speech is what leads to effective interpersonal communication. One can infer someone's emotions using facial expressions [1], speech [2], body language [3], etc. To that effect, this work focuses on emotion detection using *only* speech signals. As important as it is, it comes with its own challenges as well [4]. After examining emotions, their causes, and their effects, researchers have categorised them using a 4-D model, where each dimension—duration, quality, intensity, and pleasure—is distinct from the others [5].

The increasing technological advancements have led to a commensurate growth in human reliance on machines. Notably, the distinction between human-human interaction and human-machine interaction lies in the absence of emotional elements in the latter. This aspect has prompted the emergence of a novel research domain, namely, Speech Emotion Recognition (SER). Its applications include monitoring patients, call centre services [6], analysing driver's behaviour, etc.

Characteristics related to prosody, such as pitch, fundamental frequency (F_0), pitch frequency, duration, energy, and others, are widely employed for SER in the literature [7]. Nonetheless, these features are limited to characterizing only the vocal folds state. Therefore, incorporating a feature that characterizes both the vocal tract and vocal fold state would enhance the emotion classification performance.

This paper explores the applicability of phase spectrum in SER due to its demonstrated effectiveness in speech recognition [8] and source/system information extraction [9]. To extract intricate details from the spectral envelope, the study employs group delay and modified group delay functions, which have been found to capture system information more effectively than the magnitude or Linear Prediction (LP) spectrum. The robustness of these proposed features is also tested with state-of-the-art features used for SER.

2 Phase Based Features

Extracting phase-based features is a challenging task since the phase spectrum is discontinuous in the frequency domain. For the phase to be used, it has to be unwrapped to make it a continuous function, however, the phase unwrapping technique is computationally complex due to the non-uniqueness associated with it. On the other hand, the group delay and modified group delay techniques have similar properties to the unwrapped phase and are known to be extracted directly from the signal [10].

2.1 Group Delay and Modified Group Delay Functions

The group delay function is characterized as the negative derivative of the unwrapped Fourier transform phase. It is also possible to compute the group delay of the signal p(n) from the signal itself using the following method-

$$T_m(\omega) = -Im\frac{d(P(\omega))}{d\omega}, \tag{1}$$

upon solving the Eq (1) as stated in [10], we arrive at:

$$T_m(\omega) = \frac{P_R(\omega)Q_R(\omega) + P_I(\omega)Q_I(\omega)}{|P(\omega)|^2}. \tag{2}$$

where $P(\omega)$ and $Q(\omega)$ are Fourier transforms of $p(n)$ and $np(n)$, respectively. The $P_R(\omega)$ and $P_I(\omega)$ indicates the real and imaginary parts of $p(\omega)$, respectively. The representation of the group delay function using cepstral coefficients is given by [11]:

$$T_m(\omega) = \sum_{n=1}^{+\infty} nc(n)cos(n\omega), \tag{3}$$

where $c(n)$ indicates the n-dimensional cepstral coefficients. This operation is replicated by applying Discrete Cosine Transform (DCT). The two most important properties of group delay feature that gives them an edge compared to magnitude-based features are **additivity** and **high resolution**. Nevertheless, despite the numerous advantages it offers, the group delay function can be effectively applied in speech processing tasks only when the signal satisfies the *minimum phase* condition. If the signal is a non-minimum phase, the presence of the roots of the Z-transformed signal outside (or) close to the unit circle gives rise to the spikes in the group delay spectrum causing distortion of the fine structure of the envelope contributed by the vocal tract system and masking the formant location [11]. These spikes are due to a smaller denominator term in Eq (2) indicating that the distance between the corresponding zero location and the frequency bin on the unit circle is small.

Any meaningful use of the phase-based features comes with the reduction of inadvertent spikes due to the smaller denominator value in Eq (2). One such representation is the **modified group delay function (MODGF)** introduced to maintain the dynamic range of the group delay spectrum. The MODGF is given by [11]:

$$T_m(\omega) = \frac{T(\omega)}{|T(\omega)|}|T(\omega)|^{\alpha}, \tag{4}$$

where

$$T(\omega) = \frac{P_R(\omega)Q_R(\omega) + P_I(\omega)Q_I(\omega)}{|S(\omega)|^{2^{\gamma}}}, \tag{5}$$

where $S(\omega)$ represents the cepstrally-smoothed version of $P(\omega)$. It was seen that introducing $S(\omega)$, very low values can be avoided. The parameters α and γ are introduced to reduce the spikes and restore the dynamic of the speech spectrum, respectively. Both parameters α and γ vary from 0 to 1. In order to obtain the cepstral coefficients, DCT is applied to convert the spectrum to cepstral features. The first coefficient of the cepstral coefficients is ignored as this value corresponds to the average value in the GDF. Including the effect of linear phase due to window and location of pitch peaks w.r.t window, the importance of the value is yet to be explored [11].

2.2 Robustness of Modified Group Delay Function

In this section, we demonstrate the resilience of the modified group delay function to additive noise through analytical means. Let $u(n)$ represent a clean speech signal, which has been degraded by the addition of uncorrelated, additive noise $v(n)$ with zero mean and variance σ^2. The resulting noisy speech $c(n)$ can be represented as follows:

$$c(n) = u(n) + v(n). \tag{6}$$

Taking the Fourier Transform and obtaining the power spectrum, we have,

$$P_c(\omega) = P_u(\omega) + P_v(\omega). \tag{7}$$

Considering low SNR situation, we have:

$$P_c(\omega) = \sigma^2(\omega)(1 + \frac{P_u(\omega)}{\sigma^2(\omega)}). \tag{8}$$

Taking the logarithm on both sides and using the Taylor series expansion results in:

$$ln(P_c(\omega)) \approx ln(\sigma^2(\omega)) + \frac{P_u(\omega)}{\sigma^2(\omega)}. \tag{9}$$

Since $P_u(\omega)$ is a continuous, periodic function of ω, it can be expanded using the Fourier series, we get:

$$ln(P_c(\omega)) \approx ln(\sigma^2(\omega)) + \frac{1}{\sigma^2(\omega)}\left[\frac{d_0}{2} + \sum_{k=1}^{+\infty} d_k cos(\frac{2\pi}{\omega_0}\omega k)\right]. \tag{10}$$

where d_k's are the Fourier series coefficients. Since $P_u(\omega)$ is an even function of ω, the coefficients of sine terms are zero. Assuming the additive noise as a minimum phase signal [11], we can obtain the cepstral coefficients as [12]:

$$T_c(\omega) \approx \frac{1}{\sigma^2(\omega)}\sum_{k=1}^{+\infty} kd_k cos(\omega k). \tag{11}$$

Eq. (11) reveals that the group delay function exhibits an inverse relationship with the noise power in regions where the noise power surpasses the signal power. For the high SNR case, upon repeating the Eq. (7) - Eq.(11), we arrive at the conclusion indicating that the group delay function is directly proportional to the signal power [12]. This conveys that the group delay spectrum follows the envelope of the signal rather than that of noise. Hence, it preserves the formant peaks well in presence of additive noise [12].

3 Experimental Setup

3.1 Dataset Details

The present study employed the widely-used EmoDB dataset, developed in 2005, to assess the performance of phase based features on SER. EmoDB is a German speech corpus comprising ten actors (5 Male and 5 Female), who uttered ten German phrases under favorable recording conditions, expressing seven emotions, namely anger, joy, neutral, sadness, disgust, boredom, and fear [13]. The current investigation focused on four emotions, namely anger, happiness, neutrality, and sadness, with one male speaker reserved for test.

3.2 Classifier Used

The utilization of deep learning models, specifically Convolutional Neural Networks (CNN), has become prominent in SER, leading us to apply the same for MGDCC. Our model comprises of 2 convolution layers with filter sizes of *8* and *16*, respectively. To mitigate the issue of vanishing gradients and reduce computational complexity, we employ the Rectified Linear Unit (ReLU) activation function. The kernel size considered is (3x3). A dropout layer of *0.2*, strides of *2*, and learning rate at *0.001* is employed. *5* fold cross-validation split of *80%* and *20%* for train and test, *adam* optimizer and *categorical cross entropy* as loss function and *accuracy* as evaluation metrics is used.

3.3 Baseline Considered

The state-of-the-art features, Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) are used for comparison. To maintain uniformity among features, *20*-D feature vectors with a window length of *25*ms and a hop length of *10*ms is used for all.

4 Experimental Results

To restore the dynamic range of phase based features, MGDCC has two additional constraint parameters, alpha (α), and gamma (γ). The CNN classifier is used to fine-tune these parameters by varying them from 0 to 1 with a step size of *0.1.* The optimal parameters thus found by classification accuracy is α=**0.1,** γ=**0.1**.

4.1 Spectrographic Analysis

Panel-A, Panel-B, and Panel-C of Fig. 1 represent the Spectrogram, Mel Spectrogram, and MGDCC-gram analysis of various emotions, respectively. Fig 1(a), Fig. 1(b), Fig. 1(c), and Fig. 1(d) show the analysis for anger, happy, sad, and neutral, respectively. Mel spectrograms give broaden and dull representation of utterance and thus have obstructed the ability to identify the fine formant structures and energy distribution. It can be observed from the plots that the fine structure of the formants that can be observed in the magnitude spectrum (Panel-A) can also be seen in the spectrogram obtained by the modified group delay spectrum. Hence, there is no information loss while using phase-based cepstral coefficients. Additionally, the resolution between the formants is high in the phase-based, i.e., modified group delay spectrum resulting in a better distinction among the formants. This is due to the fact that the denominator term at the formant frequencies becomes 0 (as the pole radius approaches to unit circle) resulting in peaks that give a higher resolution formants. Additionally, phase features are able to capture irregularities in the speech signal. The presence of turbulence in a speech signal changes with emotion and these irregularities are captured better through phase signal rather than the magnitude spectrum.

4.2 Comparision with Baseline Features

From Table 1, it can be observed that LFCC is the best-performing baseline feature. The MGDCC feature outperforms the magnitude-based features i.e., MFCC and LFCC by a margin of **7.7**% and **5.14** %, respectively. This might be because of the high-resolution property of the modified group delay function which can be noticed in Fig 1. They capture the fine structures of spectral envelope and thus formant structures are emphasized well. However, GDCC fails to achieve similar performance. This is because of the noisy structure resulting from the GDCC occurring from the presence of zeros close to or outside the unit circle. These spikes cause formant masking, making it difficult to obtain

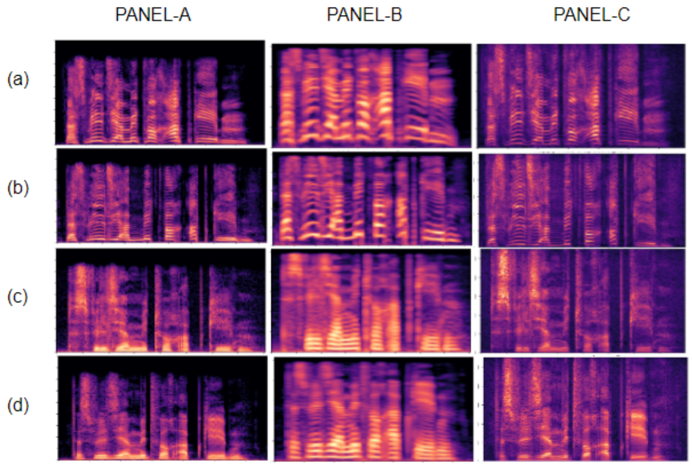

Fig. 1. Panel-A, Panel-B, and Panel-C represent the Spectrograms *vs.* Mel Spectrograms *vs.* MGD spectrogram of a male speaker uttering the same sentence in emotions- (a) anger, (b) happy, (c) sad, and (d) neutral, respectively

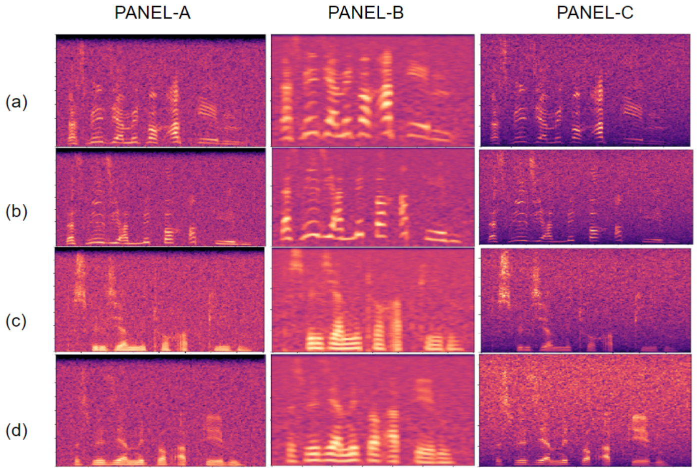

Fig. 2. Panel-A, Panel-B, and Panel-C represent the Spectrograms *vs.* Mel Spectrograms *vs.* MGD spectrogram of white noise added speech of a male speaker uttering the same sentence in emotions- (a) anger, (b) happy, (c) sad, and (d) neutal, respectively

valuable features for the classification task. It is also observed LFCC captures emotion information well in higher frequency regions as compared to MFCCs as in MFCC, the the width of the triangular filters increases with frequency and thus, ignoring fine details (Section 4.1). The emotions, in particular, anger and happy operate in higher frequency regions, which is captured better by LFCC due to constant difference between the width of filterbanks throughout.

Table 1. Classification Accuracy on CNN

Feature Set	MFCC	LFCC	GDCC	MGDCC
Test Acccuracy	71.79	74.35	56.41	**79.49**

4.3 Robustness under Signal Degradation

The robustness of the proposed features is tested using various noise types, such as white, pink, babble, and street noise with SNR levels of -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. When we consider additive white noise for evaluation, due to the nature of AWGN, the noise is distributed across all the bands of frequency. From Table 2, at the low SNR values, MGDCC clearly outperforms both magnitude-based features, MFCC and LFCC by a significant margin of **3.41 %**, **10.25 %**, respectively. Similarly, at higher SNR values, MGDCC outperforms baseline features MFCC and LFCC by **17.95 %**, **7.79 %**, respectively. Considering that the signal is degraded by the pink noise, which has higher noise power in lower frequencies rather than the higher frequencies, the MGDCC feature set outperforms both MFCC and LFCC features. Additionally, when considered non-stationary noises (noises which vary w.r.t time) such as street noise or traffic noise and babble noise are considered. The MGDCC noise robustness is evident in any kind of noise. Based on these findings, it can be inferred that the performance of the baseline features is degraded in the presence of stationary and non-stationary noise, whereas the performance of MGDCC remains intact across various noise types. These results prove the additive noise robustness property, and also that the emphasis of the group delay spectrum on the signal spectrum, rather than the noise spectrum, is a well-known characteristic. This can also be attributed to the nature of the MGDCC feature set pushes the zeros into the unit circle in an attempt of making the signal a minimum phase, which may also help in the suppression of noise. Additionally, it can be noted that LFCC and MFCC are not equally robust in white noise as the energy in higher frequency speech regions is weak making it more susceptible to noise corruption. The LFCC contains more subband filters at higher frequencies than MFCC, making it less robust to white noise. As the noise power decreases, the LFCC feature set still outperforms MFCC due to its linearly-spaced subband filters instead of the Mel filterbank. This reasoning also explains the comparable performance of MFCC to LFCC, when the signal is corrupted with pink noise.

4.4 Analysis of Latency Period

In this study, we explored the latency period of MGDCC feature set in comparison to the baseline features, i.e., MFCC and LFCC. To evaluate the performance of CNN based on different feature sets, we measured the accuracy % with respect to latency period, as depicted in Fig.3. The latency period denotes the time elapsed between the utterance of speech and the system's response, expressed as a percentage fold accuracy that represents the number of frames utilized for utterance classification. Therefore, if the system demonstrates superior performance at lower latency periods, it implies that it can classify the speech utterance effectively without requiring a prolonged duration of speech.

Table 2. Classification Accuracy on CNN with different noise types on EmoDB

NOISE	FEATURE	-10dB	-5dB	0dB	5dB	10dB	15dB
Babble	MGDCC	79.48	81.29	82.05	81.66	81.66	82.66
	MFCC	74.35	76	79.48	79.48	79.48	79.48
	LFCC	61.53	66.66	79.48	76.92	79.48	79.48
Street	MGDCC	75.35	80	81.66	81.66	71.79	86.66
	MFCC	74.35	76	76.92	79.48	88.48	82.05
	LFCC	74.35	70.23	79.48	71.79	76.92	79.48
White	MGDCC	76.92	79.48	74.35	71.79	76.92	74.35
	MFCC	69.23	76.92	74.35	43.58	82.05	43.58
	LFCC	71.79	64.10	64.10	58.97	71.79	69.23
Pink	MGDCC	74.35	69.23	71.79	71.79	71.79	74.35
	MFCC	41.79	38.46	41.02	43.58	70.35	71.79
	LFCC	71.79	66.66	66.66	61.53	71.79	71.79

The duration of utterance is upto *3* sec and is plotted at an interval of *0.5* sec. It is observed that MGDCC features give significant classification performance throughout, the highest accuracy being *79.48* % at 1.5 sec (Fig.3). On the contrary, the baseline features constantly down-perform and take longer duration to achieve comparable performance. This encourages the practical suitability of proposed MGDCC feature set.

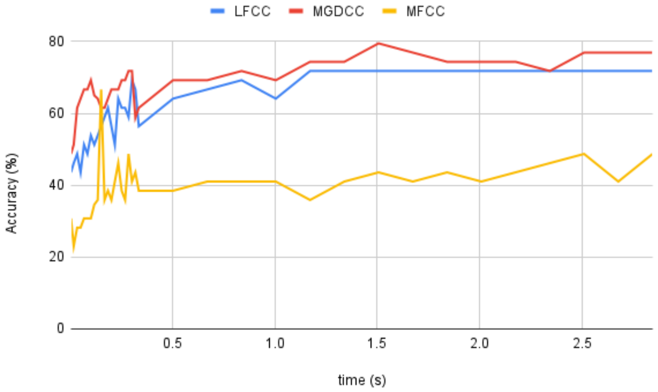

Fig. 3. Latency Period of MFCC, LFCC, and MGDCC.

5 Summary and Conclusions

In this study, phase-based vocal tract features were proposed for emotion recognition. Other spectral features MFCC and LFCC were used for comparison. The objective was to capture the irregularities in speech signal and the formant structure better for efficient SER. MGDCC also proved to perform well for stationary and non-stationary noise-added dataset due to its additive noise robustness property. The significance of linear filterbanks over Mel filterbanks was observed for emotion classification. The practical suitability of MGDCC was also calculated and promising results were seen. Further, this work can be extended by testing the robustness of convolution-type noise and its performance in SER for the mentally challenged.

References

1. P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, “Emotion recognition using facial expressions,” *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.
2. S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, “A review on emotion recognition using speech,” in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT).* IEEE, 2017, pp. 109–114.
3. L. Abramson, R. Petranker, I. Marom, and H. Aviezer, “Social interaction context shapes emotion recognition through body language, not facial expressions,” *Emotion*, vol. 21, no. 3, p. 557, 2021.
4. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.
5. M. Cabanac, “What is emotion?” *Behavioural processes*, vol. 60, no. 2, pp. 69–83, 2002.
6. V. Sethu, E. Ambikairajah, and J. Epps, “Group delay features for emotion detection,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.
7. M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.
8. R. Schluter and H. Ney, “Using phase spectrum information for improved speech recognition performance,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 1. IEEE, 2001, pp. 133–136.
9. H. A. Murthy, “Algorithms for processing fourier transform phase of signals,” Ph.D. dissertation, PhD dissertation, Indian Institute of Technology, Department of Computer . . . , 1992.
10. H. A. Murthy and V. Gadde, “The modified group delay function and its application to phoneme recognition,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03).*, vol. 1. IEEE, 2003, pp. I–68.
11. R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, “Significance of the modified group delay feature in speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2006.
12. S. H. K. Parthasarathi, P. Rajan, and H. A. Murthy, “Robustness of group delay representations for noisy speech signals,” Idiap, Tech. Rep., 2011.
13. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, “A database of german emotional speech.” in *Interspeech*, vol. 5, 2005, pp. 1517–1520.