Linear Frequency Residual Cepstral Coefficients for Speech Emotion Recognition

Baveet Singh Hora(✉), S. Uthiraa, and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar, Gujarat, India
{201901256,uthiraa_s,hemant_patil}@daiict.ac.in

Abstract. As technology advances, our reliance on machines grows, necessitating the development of effective approaches for Speech Emotion Recognition (SER) to enhance human-machine interaction. This paper introduces a novel feature extraction technique called Linear Frequency Residual Cepstral Coefficients (LFRCC) for the SER task. To the best of our knowledge and belief, this is the first attempt to employ LFRCC for SER. Experimental evaluations were conducted on the widely used EmoDB dataset, focusing on four emotions: anger, happiness, neutrality, and sadness. Results demonstrated that the proposed LFRCC features outperform state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) relatively by a significant margin: **25.64 %** and **10.26 %**, respectively, when using a residual neural network (ResNet); and **12.37 %** and **4.67 %**, respectively when combined with the Time-Delay Neural Network (TDNN) as classifier. Furthermore, the proposed LFRCC features exhibit a better Equal Error Rate (EER) than the other two baseline methods. Additionally, classifier-level and score-level fusion techniques were employed, and the combination of MFCC and LFRCC at the score-level achieved the highest accuracy of **94.87 %** and the lowest EER of **3.625%**. The better performance of the proposed feature set may be due to its capability to capture excitation source information via linearly-spaced subbands in the cepstral domain.

Keywords: Speech emotion recognition · Narrowband spectrogram · Linear frequency residual cepstral coefficients (LFRCC) · EmoDB · LP residual · Excitation source

1 Introduction

It is well known that our emotional state has a significant influence on our speech patterns, such that identical utterances may be perceived differently based on the associated emotion. With the development of technology and man's reliance on machines, reliable emotion detection is crucial for successful human-computer interaction. This has led to the development of a new research field, namely,

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14338, pp. 116–129, 2023.
https://doi.org/10.1007/978-3-031-48309-7_10

Speech Emotion Recognition (SER). Its applications include driver's behavior monitoring during autonomous driving, call center services, monitoring patients, mental health issues detection, and better human-machine interaction.

A speech signal comprises two essential components with respect to its production: the source, known as the excitation source, and the system, which represents the vocal tract system. In particular, the vocal tract is a natural physical system with the attribute of inertia for being unable to change its state unless an external source is applied. In practice, its primary excitation source is a glottal activity & glottal vibration, but no vibration can also act as the driving force to excite the vocal tract system to produce intelligible speech.

Various features, namely prosodic, source, and system features, have been employed for SER in the existing literature. Notably, the prosodic and system-based features have received the most attention in research studies. However, exploring excitation source information in speech for SER is relatively limited, as highlighted in [8]. This limitation serves as a motivation for the adoption of Linear Frequency Residual Cepstral Coefficients (LFRCC) in this study.

In earlier studies, traditional machine learning algorithms, such as Gaussian Mixture Models (GMMs) or Support Vector Machines (SVMs), were commonly employed in SER [15]. GMMs effectively modeled speech feature distributions and captured variations among different emotion classes, while SVMs were utilized for their capacity to learn decision boundaries that can distinguish between emotions [2].

As research progressed, more advanced classifiers, particularly neural network-based models, gained popularity and became prominent due to their ability to learn hierarchical and temporal representations from speech data automatically [1,3]. In this study, we leverage the advantages of Residual Neural Networks (ResNet) and Time Delay Neural Networks (TDNN) for SER tasks. ResNet's skip connections address the vanishing gradient problem, enabling deeper networks, feature reuse, and efficient training with a reduced risk of overfitting [5]. TDNN, on the other hand, excels at modeling temporal dependencies, handling variable-length inputs, offering efficiency and parallelization [12], and exhibiting shift-invariance. These qualities motivated their selection for this work.

In this work, we investigate the LFRCC features, which already had proved effective for anti-spoofing [16] for speaker verification, i.e., to capture the characteristics of natural *vs.* spoofed speech. Given the success of LFRCC in capturing the acoustic characteristics of natural speech signals, this study investigates its possible potential for SER tasks. We compare the proposed features with the state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) features using deep learning models, namely, ResNet and TDNN with Attention Statistics Pooling.

The rest of the paper is organized as follows: Sect. 2 presents details of Linear Prediction. Section 3 provides the information for the extraction of the features. Section 4 shows the experimental setup used for the study. Section 5 presents the experimental results and analysis. Finally, Sect. 6 concludes the paper along with the potential future of research directions.

2 Linear Prediction

The initial application of the Linear Prediction (LP) method can be traced back to speech coding applications, drawing inspiration from the fields of system identification and control [10]. In LP analysis, the representation of each speech sample involves a linear combination of past 'p' speech samples. Here, 'p' denotes the order of linear prediction, and the weights associated with the combination are referred to as Linear Prediction Coefficients (LPCs) [10]. Given the current speech sample as s(n), the predicted sample can be expressed as follows:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \tag{1}$$

a_k is the Linear Prediction Coefficient (LPC) coefficient. The LPCs (i.e., a_k) are utilized to predict the speech samples $s(n)$ by $\hat{s}(n)$, and the discrepancy between the actual speech sample $s(n)$ and the predicted sample $\hat{s}(n)$ is termed as the LP residual $r(n)$, which can be expressed as follows:

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{2}$$

Specifically, our method applies all-pole inverse filtering to the speech signal with the LP analysis. We have

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \tag{3}$$

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{4}$$

In this context, the variable $A(z)$ denotes an inverse filter associated with an all-pole Linear Prediction (LP) filter $H(z)$. This filter represents the system information related to the vocal tract, while the term G is referred to as the gain term in the LP model. It is worth noting that the system information is combined with the excitation source information, and the LP residual effectively captures this source information. In particular, it is known in the speech literature that LP residual attains a peak at every Glottal Closure Instant (GCI). The LP residual is subjected to cepstral-domain processing to extract the desired information. This process enables the spectral envelope representation of the excitation source signal, which is discussed in the next section.

3 Linear Frequency Residual Cepstral Coefficients

Figure 1 illustrates the functional block diagram of the proposed feature set. The input speech signal undergoes pre-emphasis filtering to balance the lower and higher frequency components [16]. Subsequently, the signal is processed with the steps mentioned in the LP block, resulting in the LP residual waveform, denoted as $r(n)$. The LP residual waveform is then divided into frames and subjected to windowing, with a duration of 25 ms and a frame shift of 15 ms. In the next step, the power spectrum is estimated for each frame of the LP residual and passed through a filterbank consisting of 40 linearly-spaced triangular subband filters. To obtain the desired LFRCC features with minimal distortion, the Discrete Cosine Transform (DCT) and Cepstral Mean Normalization (CMN) techniques are applied to the power spectrum. This sequence of operations yields the final LFRCC features that capture relevant speech signal characteristics.

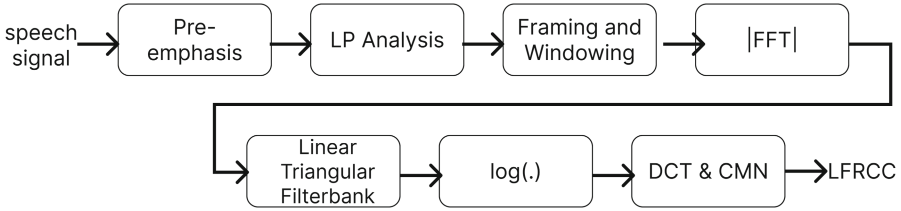

Fig. 1. Schematic block diagram of LFRCC feature extraction. After [16].

4 Experimental Setup

4.1 Dataset Used

To evaluate the effectiveness of excitation source-based features for SER, the present research utilized the EmoDB dataset developed in 2005. EmoDB is a German speech corpus comprising recordings from ten actors, comprising an equal distribution of five male and five female speakers. These speakers were asked to utter ten German phrases while being recorded under favorable conditions. The dataset encompasses seven distinct emotions: anger, joy, neutral, sadness, disgust, boredom, and fear [4] as shown in Table 1. The current study narrowed the focus to four specific emotions: anger, happiness, neutrality, and sadness. Additionally, one male speaker was reserved for the purpose of testing in this investigation as per Leave One Speaker Out (LOSO) method.

Table 1. Main Properties of the Speech Corpus in the Experiments (F - Female, M - Male).

Database	Speakers	Emotions
German Emotional Database (EmoDB)	10 (5M, 5F)	7 (anger, joy, disgust, neutral, sadness, boredom, and fear)

4.2 Classifiers Used

In SER, neural network (NN) models are used extensively for their ability to capture the non-linear relation present among speech signal samples [9]. This work implements two deep learning models to classify emotions, whose details are below.

Time Delay Neural Network (TDNN): The state-of-the-art TDNN architecture [12] has been chosen for experimentation. Table 2 provides a comprehensive overview of this architecture's layers. A dropout rate of *0.2* is applied at all TDNN layers during training to enhance the model's performance and prevent overfitting. The *categorical cross-entropy* loss function is utilized for classification tasks for training. Additionally, we used the *attentive statistics pooling* mechanism, which incorporates both higher-order statistics and attention mechanisms and has proven effective in capturing speaker-related information [11]. However, its potential for capturing emotional information is also noteworthy. Considering the standard deviation vector, this pooling method can capture long-term variability and the dynamic nature of emotions expressed over an utterance.

Furthermore, the attention mechanism enables the identification of speech frames that are more important and informative for discriminating emotions. This selective frame weighting and non-linear activation functions allow the pooling method to capture the complex patterns and variations inherent in emotional expressions. Ultimately, by aggregating frame-level features with higher weights, the attentive statistics pooling method can create an utterance-level representation that emphasizes emotional cues, making it a promising approach for extracting and understanding emotional information from speech signals.

Residual Neural Network (ResNet): The ResNet model has already been adopted for emotion recognition tasks [17]. The ResNet architecture used for this experiment is shown in Table 3. The loss function we used during training was *categorical cross-entropy*. The Resblock consists of two Conv1D layers; Batch Normalization and ReLU activation follow each layer. The residual connections in ResNet alleviate the vanishing gradient problem and improve the flow of gradient information during training [5]. This enables the network to learn emotional cues more effectively and capture subtle patterns associated with different emotions, thereby enhancing the network's ability to discriminate between emotional states.

Table 2. TDNN Architecture.

Layer	Kernel size	Filter size	Stride	Output shapes
TDNN	5	64	1	(batch_size, 64, 1)
TDNN	5	128	1	(batch_size, 128, 1)
TDNN	7	128	1	(batch_size, 128, 1)
TDNN	1	64	1	(batch_size, 64, 1)
TDNN	1	64	1	(batch_size, 64, 1)
Classic Attention	–	64	–	(batch_size, 64, 1)
Linear	128	128	1	(batch_size, 128)
Statistics pooling here				
Linear	64	64	1	(batch_size, 64)
Output	–	4	–	(batch_size, number of classes = 4)

Table 3. ResNet Architecture.

Layer	Details
Conv1	Input Channels: *in_channels* Output Channels: *out_channels* Kernel Size: 3 Stride: *1* Padding: 1
BN1	Type: BatchNorm1d
Res_Block1	Input Channels: *64* Output Channels: *32*
Res_Block2	Input Channels: *32* Output Channels: *16*
Res_Block3	Input Channels: *16* Output Channels: *16*
Avg_Pool	AdaptiveAvgPool1d
FC	Input Channels: *16* Output: num_of_classes = *4*

4.3 Baseline Considered

The state-of-the-art MFCC and LFCC features are used for comparing the proposed features. Then *39*-D feature vector was formulated containing static, delta, and double-delta parameters. The window length used was 25 ms, the number of subband filters was 40, and the number of points used in the Fast Fourier Transform (NFFT) was 512.

5 Experimental Results

5.1 Spectrographic Analysis

Figures 3 and 4 present the spectrograms of the original signal and the LP residual signal for both a female and a male speaker uttering the same sentence. In comparison to the spectrogram of the original signal, the LP residual spectrogram preserves certain details, such as the explicit representation of formants, pitch harmonics (horizontal pitch striations in a narrowband spectrogram), and overall spectral energy distribution. Further, since the LP residual is known to be intelligible, we also expect to see formant structures in the LP residual spectrogram. A notable observation shared across all emotions is the prominent energy presence at higher frequencies in the LP residual spectrograms, as highlighted by the black boxes in Fig. 3 and Fig. 4. Moreover, the width between the horizontal striations is found to be more significant for females due to their higher fundamental frequency (F_0) compared to males. Notably, anger exhibits short pauses caused by irregular breathing (puffs) in contrast to neutral and sad emotions, which feature longer pauses and low formant fluctuation due to deeper breathing. A significant distinction between happy and angry emotions lies in the distribution of high-energy content. In the case of happy emotions, the large energy density is evenly spread throughout the utterance, gradually diminishing towards the end. Conversely, anger emotions concentrate the large energy density at higher frequencies, maintaining it consistently throughout the utterance, as clearly depicted in the LP residual spectrogram. These observations were made by analyzing multiple sentences, one of which is represented in Fig. 3 and Fig. 4.

5.2 Impact of LP Order

In the proposed method, the LFRCC feature set is obtained by varying the prediction order (p) from 4 to 25 for a sampling frequency of 16 KHz, as depicted in Fig. 5. The results indicate that the highest classification accuracy is achieved using an LP order of *20*. Remarkably, this optimal LP order remains consistent regardless of the classifier employed. For both TDNN and ResNet, an LP order of *20* yields relatively highest accuracy rates of **89.29 %** and **87.17 %**, respectively (as illustrated in Fig. 5). Additionally, it is worth noting that the accuracy of emotion classification tends to be higher for higher LP orders (16–25) than for lower LP orders (4-15). This observation can be attributed to the fact that higher LP orders allow for better capture of contextual information of emotional aspects, especially concerning speech prosody, which is predominantly *suprasegmental* in nature and, thus, requires a longer duration of the speech signal.

5.3 Significance of Pitch Contour

Speech is a powerful form for expressing emotions, and the pitch or fundamental frequency (F_0) contour plays a crucial role in conveying emotional information [14]. It encompasses variations and patterns in pitch that provide valuable

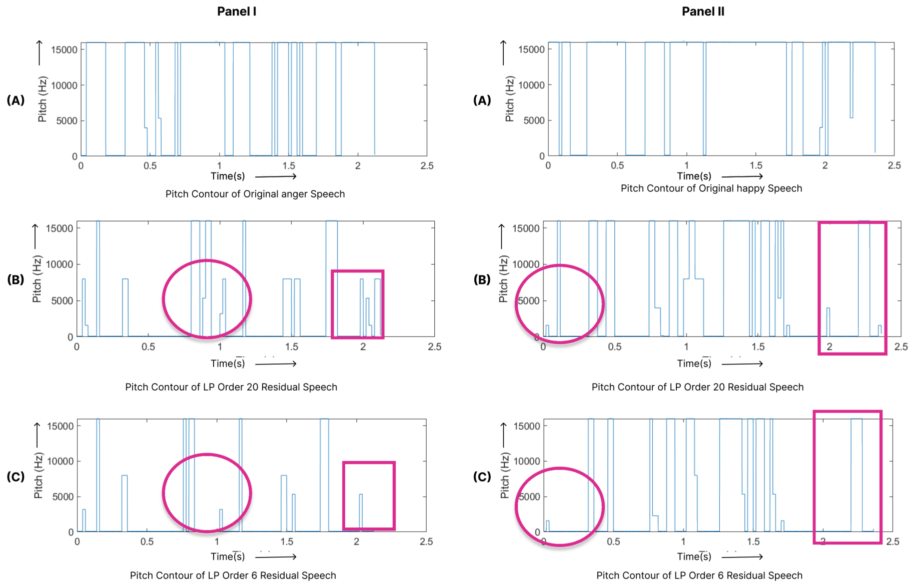

Fig. 2. Panel-I (a) pitch contour of original anger speech sample; Panel-II (a) pitch contour of original happy speech sample; Panel-I (b) and Panel-II (b) shows pitch contour of the LP residual for p=20; Panel-I (c) and Panel-II (c) shows pitch contour of the LP residual for p=6.

cues about the speaker's emotional state. We can gain initial insights into the speaker's emotional expression by analyzing the pitch level. Higher pitch levels may indicate excitement or anger, while lower levels can suggest sadness or calmness. It's crucial to acknowledge that speaker-dependent variations exist, owing to individual vocal characteristics and cultural influences that can give rise to distinct emotional pitch patterns. Understanding the pitch contour allows us to discern the emotional nuances present in speech signals. Our experiment employed the *YIN* algorithm for pitch contour plotting. Notably, we opted not to utilize overlapping frames in our analysis, considering that overlapping might introduce complexity and potentially obscure the specific effects of LP orders on pitch information. Figure 5 depicts the impact of LP order on speech and compares the pitch contour plots of LP order 6 (lowest accuracy in Fig. 5) and LP order 20 (highest accuracy in Fig. 5) for *anger and happy emotion* samples. The encircled areas in Fig. 2 depict the differences in pitch information for the two LP orders.

The pitch contour of the original emotional signals suggests significant pitch variations, possibly due to changes in speech style and prosody. In contrast, the pitch contour in the residual signal suggests that the LPC modeling successfully captures and decreases the pitch-related variations compared to the original speech signal, resulting in a more consistent and uniform plot. The increased

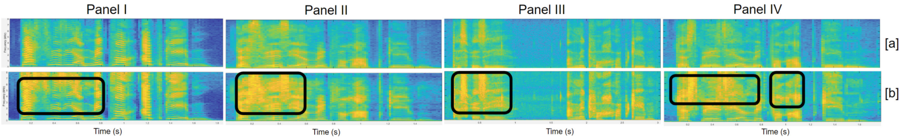

Fig. 3. Spectrographic analysis for (a) original speech signal, and (b) the corresponding LP residual. Panel I, Panel II, Panel III, and Panel IV represent the spectrograms of a female speaker for the emotions anger, happy, sad, and neutral, respectively, for the sentence *"Das will sie am Mittwoch abgeben (She will hand it in on Wednesday)".*

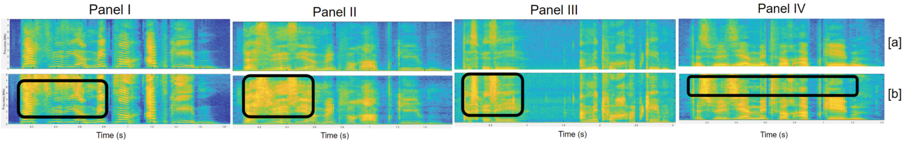

Fig. 4. Spectrographic analysis for (a) original speech signal, and (b) the corresponding LP residual. Panel I, Panel II, Panel III, and Panel IV represent the spectrograms of a male speaker for the emotions anger, happy, sad, and neutral, respectively, for the sentence *"Das will sie am Mittwoch abgeben (She will hand it in on Wednesday)".*

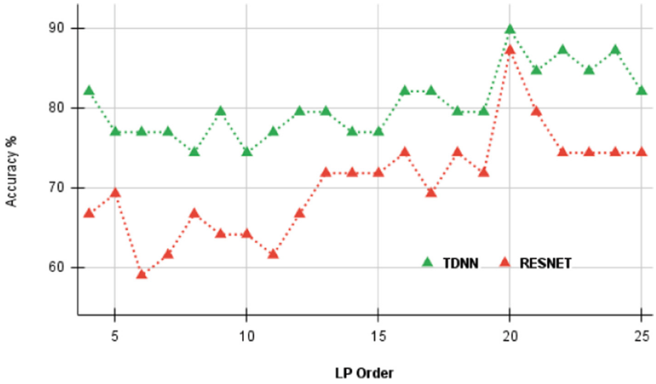

Fig. 5. Effect of LP orders for LFRCC using TDNN and ResNet Classifiers.

number of peaks (encircled areas) in the pitch contour plot for LP order 20, compared to LP order 6, indicates that the higher order LP model more accurately captures finer pitch details. This finding enhances the credibility of the results in Fig. 5. LP order 20 outperforms LP order 6 in accuracy, suggesting the presence of more relevant and steadfast pitch information that contributes to its superior performance.

5.4 Results with Score-Level Fusion

To understand the complementary information captured by different features, score-level fusion is performed using the following data fusion strategy. We will try different α values between 0.0 and 1.0 with a step size of 0.1

$$L_{Score\_fused} = \alpha L(classifier)_{feature1} + (1-\alpha)L(classifier)_{feature2}, \quad (5)$$

In the equation, $L_{feature1}$ is the raw score from the classifier of either MFCC or LFCC given as input feature, while $L_{feature2}$ represents the classifier raw score for LFRCC. Figure 6 and Fig. 7 illustrate the accuracy, and Fig. 9 and Fig. 10 show the Equal Error Rate (EER) of the score-level fusion on ResNet and TDNN, respectively. It is observed that MFCC and LFRCC give the best classification accuracy of **94.87%** and *87.18%*, and the feature pair gives the lowest EER for TDNN and ResNet classifiers, respectively, as observed in Fig. 9 and 10. MFCC and LFRCC achieve the highest classification results because, in our opinion, MFCC effectively captures spectral information in the lower frequency regions of speech, while LFRCC captures excitation information in the higher frequency regions. By combining these two feature sets, the major emotional content in speech signals can be captured comprehensively.

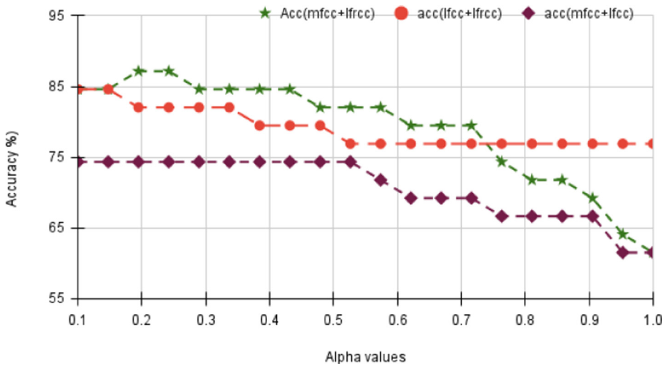

Fig. 6. Score-level fusion of features using the ResNet classifier.

5.5 Results with Classifier-Level Fusion

$$L_{Classifier\_fused} = \alpha L(feature)_{Classifier1} + (1-\alpha)L(feature)_{Classifier2}, \quad (6)$$

Figure 8 depicts the classifier-level fusion. In this, raw scores from each classifier (TDNN and ResNet) are multiplied by the weight α, and then all the weighted outputs are added together to obtain the final output, as shown in the

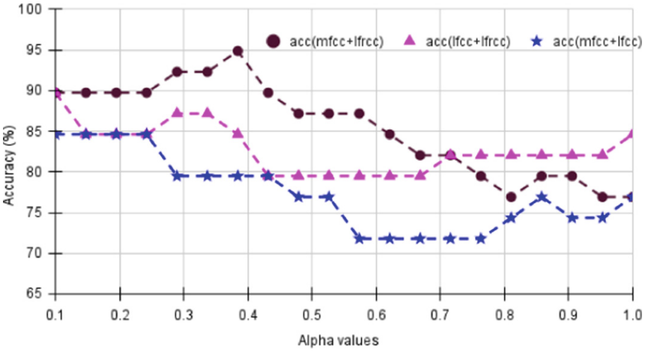

Fig. 7. Score-level fusion of features using the TDNN classifier.

above equation (6). The final score is then used to determine the accuracy of the classification. In short, we use the same feature as input to both classifiers and combine their outputs to obtain a final score in classifier-level fusion, whereas, in score-level fusion, classifiers remain the same while raw scores from different feature sets are used. It is observed that the highest classification accuracy obtained for MFCC is **76.92 %**, LFCC is **84.62 %**, and LFRCC is **87.18 %**, at the α values shown in Fig. 8, these are close to the outcomes of the single TDNN classifier as visible in Table 4. This suggests that classifier characteristics did not significantly enhance accuracy compared to score-level fusion, where a substantial improvement was observed from the combination of MFCC and LFRCC feature sets. This proves that the results obtained are not significantly dependent on the classifier, thereby proving that the proposed LFRCC feature set captures emotional information, at least not worse than MFCC or LFCC.

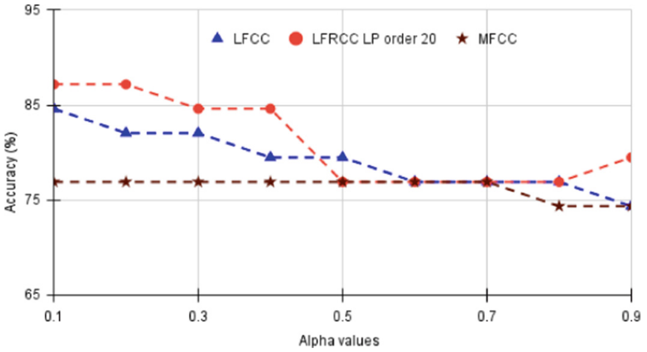

Fig. 8. Classifier-level fusion for single feature sets using the TDNN and the ResNet classifiers.

5.6 Performance of LFRCC on SER

Table 4 shows that the cepstral coefficients with linear filterbanks (i.e., LFCC) result in better classification than the Mel filterbanks (i.e., MFCC), irrespective of the classifiers used for SER. This suggests that LFCC can potentially capture emotional information well in higher frequency regions compared to the MFCC as the width of the triangular filters in MFCC increases with frequency, thus ignoring fine spectral details. In particular, anger and happiness operate in higher frequency regions (Sect. 5.1), which is captured better by LFCC due to the constant difference between the width of subband filters in the filterbank.

The LFRCC feature set demonstrates superior performance compared to the baseline MFCC and LFCC features, achieving a **25.64%** and **10.26%** absolute improvement, respectively, with the ResNet and a **12.37 %** and **4.67 %** improvement, in the case of TDNN (as shown in Table 4). Also, this improved performance can be seen from lower EER for the case of LFRCC, obtaining **10.56%**, **7.19%** for ResNet and TDNN, respectively, as can be observed from Table 4.

It is widely acknowledged that the lungs are crucial in providing the necessary airflow, which acts as a power supply for speech production. Consequently, changes in respiratory patterns directly impact the timing, duration, and overall rhythm of speech. As a result, respiratory patterns can influence emotional expression in speech signals [6,7]. The LFRCC feature set effectively captures this excitation source information (as discussed in Sect. 5.1). Specifically, the section around the Glottal Closure Instants (GCI) in the LP residual signal exhibits a high signal-to-noise ratio (SNR) due to periodic excitation. During the Glottal Closure (GC) phase, the excitation source is completely isolated from the vocal tract [9]. This region contains valuable information that cannot be adequately captured by MFCC and LFCC, making LFRCC a compelling choice for representing these features.

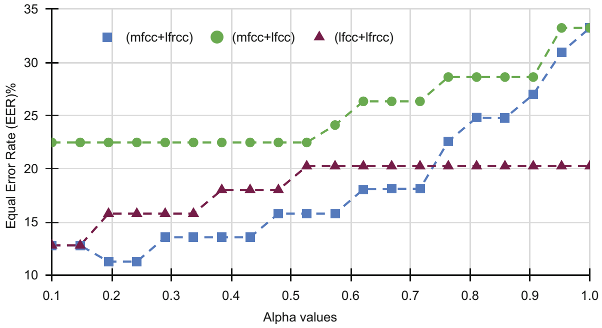

Fig. 9. EER values for score level fusion on Resnet.

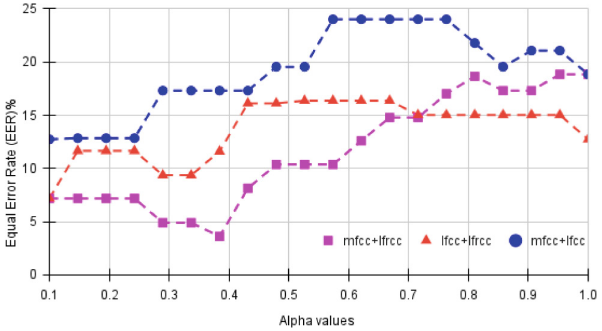

Fig. 10. EER values for score-level fusion on TDNN.

Table 4. Classification Accuracy and EER.

Classifier	Feature	Accuracy (%)	EER (%)
ResNet	**MFCC**	61.54	33.32
	LFCC	76.92	20.25
	LFRCC	**87.18**	***10.56***
TDNN	**MFCC**	76.92	18.84
	LFCC	84.62	12.74
	LFRCC	**89.29**	**7.19**

6 Summary and Conclusion

This study proposed an excitation source-based LFRCC feature set for emotion recognition. The vocal tract system features, such as MFCC and LFCC, were used for performance comparison. The objective was to exploit the complementary information in LP residual-based features, and they proved to perform much better than the state-of-the-art spectral features, indicating that the proposed features contain specific discriminative power to classify emotions. The significance of using a linear filterbank over Mel filterbank was also observed. Pitch contour and spectrogram analysis provide cues regarding optimal LP order selection and relevant emotion-based information. Moreover, the results from classifier-level fusion proved the efficiency of the proposed feature regardless of the classifier used. It is also known that breathing patterns are different for normal *vs.* mentally disturbed patients and indicates the possible potential of emotional recognition of mentally challenged individuals. As part of future research directions, exploring the incorporation of an average pitch level as a reference point holds the potential to mitigate speaker-dependent variations and can improve inferences from contour plots. Additionally, investigating the integration of this feature set with advanced neural network architectures, such as DCNNs [3] and autoencoders [13], presents a promising avenue for further research.

References

1. Akçay, M.B., Oğuz, K.: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. **116**, 56–76 (2020)
2. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. **43**, 155–177 (2015)
3. Bhangale, K., Kothandaraman, M.: Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics **12**(4), 839 (2023)
4. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al.: A database of german emotional speech. In: Interspeech, Lisbon, Portugal, vol. 5, pp. 1517–1520 (2005)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770–778 (2016)
6. Homma, I., Masaoka, Y.: Breathing rhythms and emotions. Exp. Physiol. **93**(9), 1011–1021 (2008)
7. Jerath, R., Beveridge, C.: Respiratory rhythm, autonomic modulation, and the spectrum of emotions: the future of emotion recognition and modulation. Front. Psychol. **11**, 1980 (2020)
8. Koolagudi, S.G., Reddy, R., Rao, K.S.: Emotion recognition from speech signal using epoch parameters. In: International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IISc Bangalore, India (2010)
9. Krothapalli, S.R., Koolagudi, S.G.: Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. **16**, 181–201 (2013)
10. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
11. Okabe, K., Koshinaka, T., Shinoda, K.: Attentive statistics pooling for deep speaker embedding. In: Interspeech, Hyderabad, India, pp. 2252–2256 (2018)
12. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Interspeech, Dresden, Germany, pp. 3214–3218 (2015)
13. Sadok, S., Leglaive, S., Séguier, R.: A vector quantized masked autoencoder for speech emotion recognition. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 1–5 (2023)
14. Scherer, K.R.: Vocal communication of emotion: a review of research paradigms. Speech Commun. **40**(1–2), 227–256 (2003)
15. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. **21**(1), 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z
16. Tak, H., Patil, H.A.: Novel linear frequency residual cepstral features for replay attack detection. In: Interspeech, Hyderabad, India, pp. 726–730 (2018)
17. Tripathi, S., Kumar, A., Ramesh, A., Singh, C., Yenigalla, P.: Focal loss based residual convolutional neural network for speech emotion recognition. arXiv preprint arXiv:1906.05682 (2019)