Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification

Conference Paper · November 2017

DOI: 10.1007/978-3-319-69900-4_40

3 authors, including:

Dharmesh Agrawal
Dhirubhai Ambani Institute of Information an…
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Hemant Patil
Dhirubhai Ambani Institute of Information an…
157 PUBLICATIONS **462** CITATIONS
SEE PROFILE

Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification

Rishabh N. Tak (0000-0002-1924-8482), Dharmesh M. Agrawal (0000-0003-0047-4903), and Hemant A. Patil (0000-0002-4068-2005)

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology, (DA-IICT), Gandhinagar

Abstract. In Environment Sound Classification (ESC) task, only the magnitude spectrum is processed and the phase spectrum is ignored, which leads to degradation in the performance. In this paper, we propose to use phase encoded filterbank energies (PEFBEs) for ESC task. In proposed feature set, we have used Mel-filterbank, since it represents characteristics of human auditory processing. Here, we have used Convolutional Neural Network (CNN) as a pattern classifier. The experiments were performed on ESC-50 database. We found that our proposed PEFBEs feature set gives better results compared to the state-of-the-art Filterbank Energies (FBEs). In addition, score-level fusion of FBEs and proposed PEFBEs have been carried out, which leads to further relatively better performance than the individual feature set. Hence, the proposed PEFBEs captures the complementary information than FBEs alone.

Keywords: Sound Classification, Phase Encoded spectrogram, Score-level fusion, CNN.

1 Introduction

Environmental Sound Classification (ESC) is an important research problem due to its application in various field, such as, hearing aids, road surveillance system, security and safety purpose, etc. ESC task was earlier attempted using mel frequency cepstral coefficients (MFCCs) feature set and GMM classifier [3]. Recently, deep learning - based approaches are used for ESC task, such as, Convolutional neural network (CNN)-based classification built for end-to-end system for ESC on CNN framework [9].

In this paper, we propose the new phase-based approach for ESC task. In particular, we propose the phase encoded Mel Filterbank energies with CNN as a back-end for ESC task. In this paper, we explore importance of phase in audio processing task. To the best of the authors knowledge, this is the first approach in the literature that used phase encoded feature sets for ESC task. Results shows that the phase encoded based feature set perform better than the state-of-the-art feature namely, Mel-filterbank energies (FBEs). The score-level fusion of PEFBEs and FBEs gives the significant performance jump in classification accuracy.

2 Phase Encoded Feature set

2.1 Motivation

In speech processing, the phase spectrum of a speech signal has gained lesser attention than the magnitude spectrum. There are mainly two issues due to which phase information is discarded. First, the computationally complex phase unwrapping task during processing of phase spectrum [8]. Second, perceptually, magnitude spectrum is more relevant than the phase spectrum [8]. In addition, the very often used features, such as, mel cepstral frequency coefficients (MFCC), linear prediction cepstral coefficients (LPCC), frequency domain linear predicted (FDLP) coefficients etc., are derived from the magnitude spectrum of speech [8]. Recent studies have reported using FT phase-based features, such as, Modified Group Delay (MGD) [15], Relative Phase Shift (RPS) [10], Cosine-Phase [14], etc. Motivated by these studies, we propose a novel phase-based features. These features are derived from very recent findings of phase encoding in the magnitude spectrum of speech signal. It results into magnitude spectrum that contains both magnitude as well as phase information in it. The algorithm of phase encoding is developed for new class of signals known as Causal Delta Dominant (CDD) signal. By making a signal as a CDD signal, we can reconstruct back original signal from its magnitude spectrum alone [11, 12]. An interesting aspect of this work is that, there are no constraints on the signal, i.e., it is not necessary for signal to be minimum-phase or need not to have rational system function $(H(\mathcal{Z}))$ or corresponding frequency response $H(e^{j\omega})$ Fig. 1. The block diagram of phase encoding scheme for signal reconstruction is shown below.

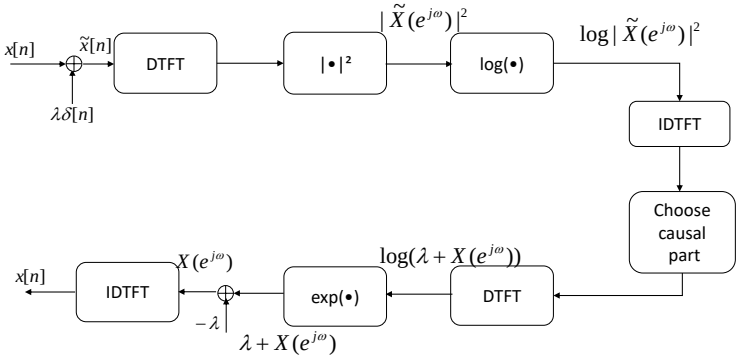

Fig. 1. *Block diagram of phase encoded spectrogram and signal reconstruction. After [11].*

2.2 Mel Filternbank Energies (FBEs)

Mel frequency analysis of speech is based on human perception experiments. It is observed that human ear acts as a bank of subband filters (i.e., filterbank). It concentrates on only certain frequency components (primarily due to the place theory of hearing). These filters are overlapped and non-uniformly spaced on the frequency-axis. In audio processing, it is shown that within 10-30 ms duration, the signal is considered to be the stationary and hence, smaller duration window is selected [4].

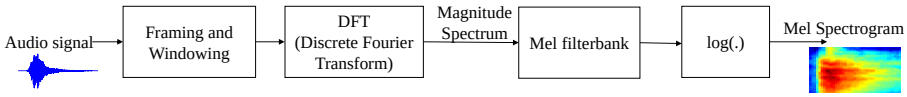

Fig. 2. Block diagram of Mel spectrogram of an audio signal. After [2].

2.3 Phase Encoded Filterbank Energies (PEFBEs)

To use the phase-encoded approach for speech-related applications, it is necessary to derive a set of features. As shown in Fig. 3, a Kronecker delta impulse of λ amplitude is origin at each frame of a signal. Next, we take DFT of every frame and apply the normalization on each FFT-bins. Then, calculate the power spectrum of individual frames. This identifies frequencies that are present in a given the frame. Mel filterbank is applied to the power spectra, which gives the total energy present in each subband filter. Then, we apply log-operation on subband energies. We refer these subband energies as *phase encoded filterbank energies (PEFBEs)*. We set number of FFT bins as total

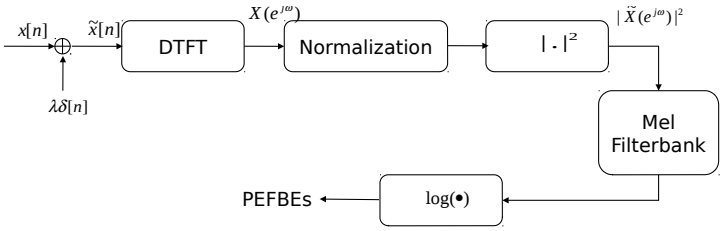

Fig. 3. *Block diagram of proposed PEFBEs feature extraction scheme.*

number of samples per frame. The proposed algorithm to extract PEFBEs features from the speech signal is given in Algorithm 1.

2.4 Importance of λ

To justify the importance of λ, an experiment was conducted on 1000 utterances of natural, VC and SS randomly selected from ASV spoof 2015 challenge database [16]. For

each of the utterance, its corresponding reconstructed signal back (using the approach shown in Fig. 1) for $\lambda = 0$ and $\lambda \neq 0$ is estimated. The log-spectral distortion (LSD) is calculated for $\lambda = 0$ and $\lambda \neq 0$, and compared with the LSD values for natural, VC and SS speech signals.

Table 1. *Mean log-spectral distortion (LSD) values of 1000 utterances for various λ values from ASVspoof 2015 database*

Speakers	$\lambda = 0$	$\lambda \neq 0$	Relative difference (%)
Natural	2.02	0.381	81.13
VC	2.053	0.360	82.46
SS	2.115	0.381	81.90

From Table 1, it is observed that result of relative difference between LSD values for $\lambda = 0$ and $\lambda \neq 0$ is found to be approximately 81-82 %. Thus, it indicates encoding of phase in the magnitude spectrum captures better signal reconstruction capability (i.e., synthesis) of the speech pattern. The key difference between Fig. 1 and Fig. 3 is the normalization block. It is observed that, with normalization, formants and harmonics are more visible as compared to without normalization. Hence, normalization increases the energy variations which is useful for ESC.

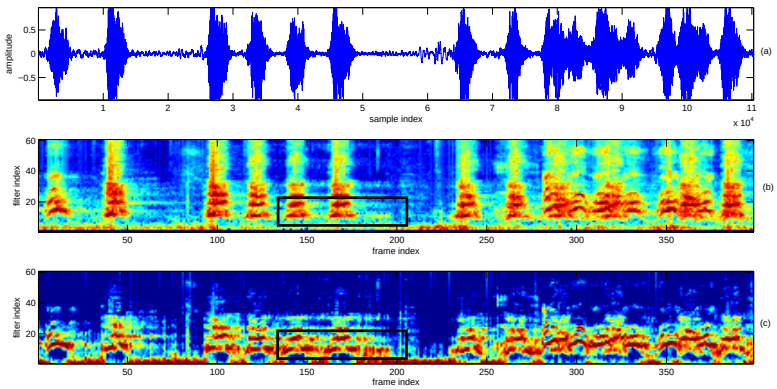

Fig. 4. Spectrographic analysis: (a) raw audio signal of dog sound, (b) Mel filterbank spectrogram, (c) phase encoded spectrogram. The regions indicated by blackboxes shows the differences between spectrum representation in (b) and (c).

As shown in Fig. 4(b) and Fig. 4(c), the proposed PEFBEs (Fig. 4(c)) has better representation in lower frequency region than FBEs (Fig. 4(b)). However, PEFBEs has slightly lower resolution in higher frequency regions as compared to the FBEs. Such

representation observed improvement in classification accuracy of classes, such as, harmonic sounds, transient sounds, etc .

Algorithm 1 Proposed PEFBEs Feature Extraction Algorithm

1: Take a speech signal $x[n]$.
2: Apply framing on the signal, let $(x_t)_{t\epsilon[1,P]}$ is the t^{th} frame with 20 ms window size and 10 ms window shift.
3: Add Kronecker impulse delta of λ amplitude to each speech frame at the origin, $\tilde{x_t}[n] = x_t[n] + \lambda\delta[n]$.
4: Take DFT of each frame, such as, $\tilde{X_t^i}(e^{j\omega})$ = λ + $X_t^i(e^{j\omega})$, where $X_t^i(e^{j\omega})$ indicates i^{th} FFT-bin, $\forall\ t\epsilon[1, P]$.
5: Perform the normalization on each FFT-bin.

$$S_t^i(e^{j\omega}) = \frac{\tilde{X_t^i}(e^{j\omega}) - mean(\tilde{X_t^i}(e^{j\omega}))}{std(\tilde{X_t^i}(e^{j\omega}))}$$

6: Perform absolute squaring that results in power spectra.
7: Apply Mel filterbank on power spectra.
8: Apply $\log(.)$ on Mel spectrum energies.

3 Experimental Setup

3.1 Dataset

In this paper, we have used the publicly available database ESC-50 [7] for the ESC task. The ESC-50 dataset consists of 2000 short (5 seconds) environmental recordings. These recordings are divided into 50 equally balanced classes. These 50 classes are divided into five major groups, namely, animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds and exterior/urban noises. The files are pre-arranged in 5-folds for comparable cross-validation. Due to this reason, the results of the experiments can be directly compared to the baseline results and with the previous approaches.

3.2 Convolutional Neural Network (CNN) classifier

We have used the CNN classifier with the architecture as proposed in [6] for the ESC task. However, we have not used data augmentation technique. Since the objective of this paper is to compare the performance of the front-end feature representation, we have not used the augmentation to analyze as to how these features perform for all the classes. Before feature extraction for CNN classifier, we first pre-process the audio signal. All the audio files were downsampled to 22.05 kHz. To extract features, the audio files were divided into frames by using 25 ms Hamming window with 50 % overlap. Then, we applied silence removal algorithm. For silence removal, we first check for more than three consecutive silence frames (approximately, 50 ms duration). If silence is present in more than three frames, then we remove the silence frames, else we keep

those frames. Simple energy thresholding algorithm was used to remove the silence regions. Mel Filterbank Energies (FBEs) are used as the baseline features. 60-D FBEs, and PEFBEs were extracted from files of audio frames. The short segments of 41 frames were used as the input to the CNN. The segments were extracted with 50 % overlap from the audio files.

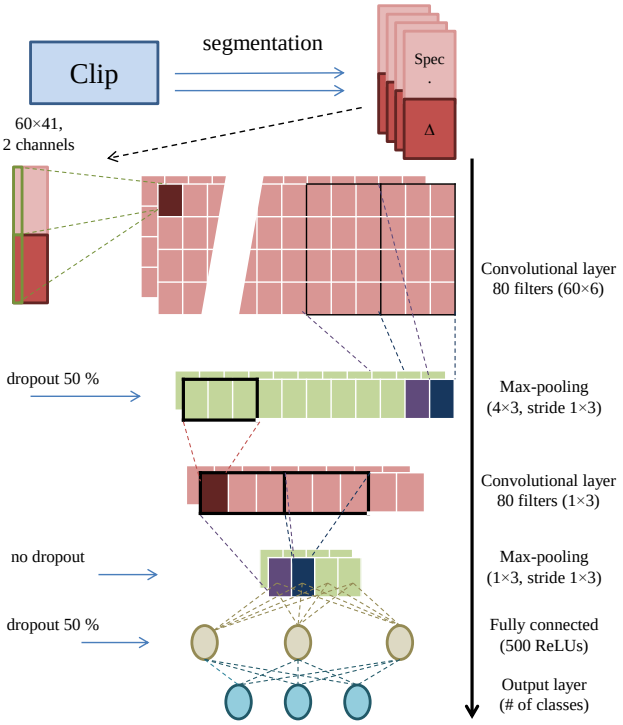

Fig. 5. *CNN architecture for ESC task. After [6].*

Figure 5 shows the details of each layer in the CNN architecture that we have used in ESC task. The network was implemented using Keras [1] with theano back-end on NVIDIA Titan-X GPU. A mini-batch implementation with 200 batch size was used to train the network. Network parameters were similar to as used in [6]. The learning rate of 0.002, L^2 regularization with the coefficient 0.001 and network was trained for 300 epochs. At the testing time, the class of the test audio files were using the probability prediction scheme [6]. We performed score-level fusion of different feature sets as used in [5].

4 Experimental Results

To evaluate the performance of various feature sets, 5-fold cross-validation was performed on ESC-50 dataset. We compare the performance of PEFBEs with FBEs. The overall results of the proposed method and baseline feature sets are summarized in Table 2 with CNN as classifier. It can be observed that PEFBEs perform significantly better than FBEs with an absolute improvement of 5.45 % in classification accuracy. Moreover, to investigate the possibility of any complementary information captured by different feature sets, we have done their score-level fusion. The score-level fusion of PEFBEs with FBEs improves the performance. However, the score-level fusion of FBEs (73.25 %) and PEFBEs (67.80 %) achieved the best accuracy of 84.15 % in this paper. This shows that the proposed PEFBEs contains highly complementary information over the FBEs, which is helpful in the ESC task. Our proposed work is also

Table 2. *% Classification accuracy of ESC-50 dataset with different feature sets and its score-level fusion. The $\oplus$ sign and α indicate score-level fusion and fusion factor, respectively.*

Feature Sets	α	Accuracy (%)
FBEs	-	67.80
PEFBEs	-	**73.25**
FBEs $\oplus$ PEFBEs	0.5	**84.15**

compared with the other studies reported in the literature in (as shown Table 3). Again, it can be observed from Table 3 that, PEFBEs performs significantly better than CNN with FBEs [6], [13]. In [13], filterbank is learned from the raw audio signal using CNN as an end-to-end system. The EnvNET [13] performs better when combining with log Mel CNN. However, our proposed PEFBEs outperform EnvNET [13] even without the system combination indicating the significance of phase for the ESC task.

Table 3. *Comparison of classification accuracy of ESC-50 dataset in the literature. The $\otimes$ sign indicated system combination before soft-max.*

Feature Sets	Accuracy (%)
PEFBEs (proposed)	**73.25**
FBEs $\oplus$ PEFBEs (proposed)	**84.15**
Piczak FBEs-CNN [6]	64.50
Human [7]	81.30
EnvNET [13]	64.00
logmel-CNN [13]	66.5
logmel-CNN $\otimes$ EnvNet [13]	71.00

5 Summary and Conclusions

In this study, we use the state-of-the-art feature set FBEs, and proposed PEFBEs for ESC task. Performance of ESC system was compared with FBEs on publicly available dataset, ESC-50. The proposed PEFBEs feature set gave better results for this application with the same parametrization as that of state-of-the-art ESC system. Moreover, the results suggested that using score-level fusion of FBEs and proposed PEFBEs gave

better accuracy than the individual feature set. This indicates that the proposed PEFBEs contains complementary information than FBEs alone. Our future work plan includes the use of proposed PEFBEs feature set for different datasets, such as, UrbanSound8K and RWCP datasets.

References

1. Chollet, F.: Keras. https://github.com/fchollet/keras { Last accessed on 26^{th} February, 2017}
2. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on acoustics, speech, and signal processing 28(4), 357–366 (1980)
3. Elizalde, B., Lei, H., Friedland, G., Peters, N.: An i-vector based approach for audio scene detection. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (2013)
4. Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based context recognition. IEEE Trans. on Audio, Speech, and Language Processing 14(1), 321–329 (2006)
5. Li, J., Dai, W., Metze, F., Qu, S., Das, S.: A comparision of deep learning methods for environmental sound detection. In: IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP). pp. 126–130. New Orleans, USA (2017)
6. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 25^{th} Int. Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. Boston, MA, USA (2015)
7. Piczak, K.J.: ESC: Dataset for environmental sound classification. In: Proc. of the 23^{rd} Int. Conf. on Multimedia. pp. 1015–1018. Brisbane, Australia (2015)
8. Raitio, T., Juvela, L., Suni, A., Vainio, M., Alku, P.: Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis. Speech Communication 81, 104–119 (2016)
9. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. 24(3), 279–283 (March 2017)
10. Saratxaga, I., Sanchez, J., Wu, Z., Hernaez, I., Navas, E.: Synthetic speech detection using phase information. Speech Communication 81, 30–41 (2016)
11. Seelamantula, C.S.: Phase-encoded speech spectrograms. In: INTERSPEECH. pp. 1775–1779. San Francisco, USA (2016)
12. Shenoy, B.A., Mulleti, S., Seelamantula, C.S.: Exact phase retrieval in principal shift-invariant spaces. IEEE Transactions on Signal Processing 64(2), 406–416 (2016)
13. Tokozume, Y., Harada, T.: Learning environmental sound with end-to-end convolutional neural network. In: IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP). pp. 2721–2725. New Orleans, USA (2017)
14. Wu, Z., Siong, C.E., Li, H.: Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,. In: INTERSPEECH, Portland, Oregon, USA. pp. 1700–1703 (2012)
15. Yegnanarayana, B., Saikia, D., Krishnan, T.: Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Transactions on Acoustics, Speech, and Signal Processing 32(3), 610–623 (1984)
16. Zhizheng, Kinnunen, T., Evans, N.W.D., Yamagishi, J., Hanilçi, C., Sahidullah, M., Sizov, A.: ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,. In: INTERSPEECH. pp. 2037–2041. Dresden, Germany (2015)