# Speech Emotion Recognition Based on Power Normalized Cepstral Coefficients in Noisy Conditions

M. Bashirpour* and M. Geravanchizadeh*

**Abstract:** Automatic recognition of speech emotional states in noisy conditions has become an important research topic in the emotional speech recognition area, in recent years. This paper considers the recognition of emotional states via speech in real environments. For this task, we employ the power normalized cepstral coefficients (PNCC) in a speech emotion recognition system. We investigate its performance in emotion recognition using clean and noisy speech materials and compare it with the performances of the well-known MFCC, LPCC, RASTA-PLP, and also TEMFCC features. Speech samples are extracted from the Berlin emotional speech database (Emo DB) and Persian emotional speech database (Persian ESD) which are corrupted with 4 different noise types under various SNR levels. The experiments are conducted in clean train/noisy test scenarios to simulate practical conditions with noise sources. Simulation results show that higher recognition rates are achieved for PNCC as compared with the conventional features under noisy conditions.

**Keywords:** Noisy Acoustic Condition, Noise Robust Auditory Feature, Power Normalized Cepstral Coefficients (PNCC), Speech Emotion Recognition.

## 1 Introduction

The human speech conveys different information pertaining to the message, speaker, language, emotion, and so on. Among other information, the same textual message would be conveyed with different meanings by incorporating appropriate emotions. This implies the need to develop speech processing systems that can process emotions along with the message. Automatic recognition of human emotions from speech signal is a very active research topic which has attracted recently much attention in many fields such as speech processing, pattern recognition and artificial intelligence.

Most previous approaches in the speech emotion recognition (SER) area have focused on detecting emotions in clean speech which was recorded in a quiet acoustical conditions [1-3]. However, the human auditory system is able to perceive emotions even in adverse noisy environments. In recent years, robust emotion recognition in noisy conditions has become an important research topic in the emotional speech recognition area, because in these scenarios emotional speech signals are usually disturbed with different noise types, causing the performance of such recognizing systems to decrease.

The present methods of robust speech emotion recognition can be classified into three main categories. In the first category, some of the efforts in this field are constrained to simple feature selection methods. Schuller et al. [4] employed a fast information gain ratio–based feature selection approach to find the suitable feature subsets from a large acoustic feature set according to the noise situation. In order to find the most appropriate features for speech emotion recognition in the presence of babble noise with different signal-to-noise ratios, Karimi and Sedaaghi, extracted 286 features from speech utterances of two emotional speech datasets in German and Persian [5]. Then, the best features were selected using different filter and wrapper methods. To reduce the influence of noise, the authors in [6, 7] used a feature dimensionality reduction method, called enhanced Lipschitz embedding.

In the second category of the methods for robust SER, the focus is on speech emotion classifiers. In [8], by using a weighted sparse representation model based on the maximum likelihood estimation (MLE), an enhanced sparse representation classifier (enhanced-SRC) was proposed for robust emotion recognition in noisy speech. The proposed classifier was used to perform spoken emotion recognition, and its

Iranian Journal of Electrical & Electronic Engineering, 2016.
Paper first received 21 June 2016 and accepted 15 October 2016.

* The Author is with the Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran.
E-mail: meysam.bashirpour@tabrizu.ac.ir and geravanchizadeh@tabrizu.ac.ir
Corresponding Author: M. Geravanchizadeh.

performance was investigated on both clean and noisy emotional speeches.

In the third category, the efforts focus on feature extraction. In [9], the Teager-energy-based Mel-frequency cepstral coefficients (TEMFCC) was proposed for automatic speech emotion recognition (ASER) in noisy environments.

The above-mentioned studies reflect the fact that most studies in robust SER are constrained to simple feature selection methods. These studies often review the many features to obtain an optimal set of features with the highest recognition rate for a particular testing environment. These brute-force methods seem to work, but they only evaluate recognition performance under matched conditions, where the test data is under the same noisy condition as the training data. In addition to this, in most of these works, only one or two types of noises are considered. Undoubtedly, the performance of these methods decreases with changes in the test environment. In contrast to the feature selection approaches for robust SER, very little work has been done in the field of feature extraction to classify speech emotions in noisy conditions. Feature extraction is one of the most critical aspects of any successful speech emotion recognition system which deserves a detailed investigation.

Many of the systems developed for automatic speech, speaker, and emotion recognition, and related research fields are based on variants of the following types of features: Mel-frequency cepstral coefficients (MFCC) [10, 11], perceptual linear prediction (PLP) coefficients [12], and linear prediction cepstral coefficients (LPCC) [13, 14]. However, it is well established that their performance degrades severely when there is a mismatch between the training and testing conditions, typically due to background noise. Recently, a new feature extraction algorithm, called power normalized cepstral coefficients (PNCC) has been introduced [15, 16] that is based on auditory processing. As described in several papers, the PNCC has been shown to provide better speech recognition accuracy than the other algorithms such as MFCC, zero-crossing peak amplitude [17], RelAtive SpecTrAl perceptual linear prediction (RASTA-PLP) [18], and perceptual minimum variance distortionless response [19], particularly in mismatched training conditions [15, 20, 21].

In this paper, we evaluate and compare the performance of the PNCC against conventional acoustic features such as the MFCC, LPCC, RASTA-PLP, and also newly proposed TEMFCC feature [9] by artificially adding 4 types of noises at different levels to the speech signal and then computing their recognition accuracy. Since there is not any published or reported work for employing the PNCC in the field of speech emotion recognition, this work represents first attempt that develops a speech emotion recognition system based on the PNCC and performs a comprehensive comparative evaluation with other known features under different types and levels of noises.

To generalize the results and evaluate the robustness of different speech features according to the spoken language, we use 2 speech corpora in 2 different languages; Berlin emotional speech database (EMO DB) [22] and Persian Emotional Speech database (Persian ESD) [23].

The organization of the paper is as follows: Section 2 reviews emotion recognition system and its building blocks. The feature extraction algorithm as well as the classification procedure are presented in this section. An explanation of the datasets used in the experiments and the classification performance of these features under clean and noisy conditions are investigated in Section 3. Finally, Section 4 gives the concluding remarks.

## 2 Method

### 2.1 General structure of the SER

A basic emotion recognition system, as shown in Fig. **1**, consists of two steps: feature extraction and emotion classification. Feature extraction is concerned with extracting suitable features efficiently characterizing different emotions from speech signals and the second step aims to identify the underlying emotions of speech utterances.

For classification purposes, some pre-trained class models are required. Therefore, each emotion recognition system has a training phase in which the class models are trained. In a conventional emotion recognition system (i.e., clean train/clean test scenario), clean speech is used for both of training and testing. However, in real conditions (i.e., noisy conditions), clean speech is used for training, but non-clean (e.g., noise contaminated) speech is used for testing. In this paper, clean train/noisy test scenario is considered in the experiments to study the robustness of features in noise.

### 2.2 Feature extraction

In this step, we follow the standard procedure to extract the MFCC [10], LPCC [24], RASTA-PLP [18], and also TEMFCC [9]. Regarding our focus on the PNCC, its feature extraction algorithm is discussed in detail below.

#### 2.2.1 Power normalized cepstral coefficients (PNCC)

The structure of the PNCC extraction algorithm is illustrated in Fig. 2. The processing stages of this algorithm are as follows [15]. First, a pre-emphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ is applied to the input signal. Then, a short-time Fourier transform (STFT) of the signal is computed using the Hamming windows of duration 20 ms with 10 ms between frames, and a DFT size of 1024. The computed squared magnitude of STFT outputs are then weighted by a 40-channel Gammatone filterbank whose center frequencies are linearly spaced

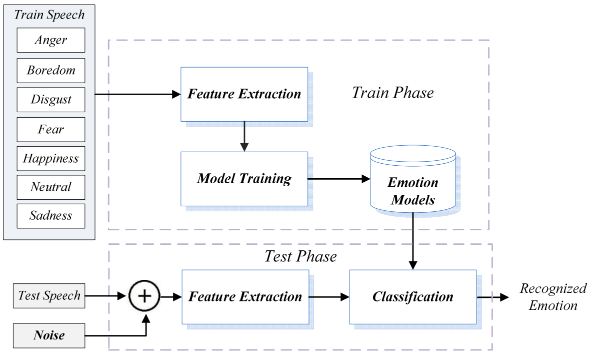

**Fig. 1** Block diagram of emotion recognition system.

in equivalent rectangular bandwidth (ERB) [25] between 200 Hz and 8000 Hz.

The next processing stage concerns the “Medium-Time Processing”. Here, a long-duration temporal analysis (using e.g., 5 frames) is performed to estimate the noise floor level and to subtract it from the instantaneous power of the input signal. The output of the “Medium-Time Processing” is a transfer function that modulates the original signal in the “Time-Frequency Normalization” stage. Then, the “Mean Power Normalization” stage normalizes the signal power by dividing the input by a running average of the overall power.

At the next stage, a power function nonlinearity with exponent 1/15 is applied to the input. Experiments demonstrated that incorporating the power function non linearity improves the recognition accuracy [26]. The final stages of the PNCC extraction algorithm are the computation of the DCT and the mean normalization, respectively.

In this paper, the PNCCs are employed as a new auditory feature in the emotion classification system.

### 2.3 Classification procedure

Here, the simple maximum likelihood (ML) estimation is used as the classification method. This method requires some pre-trained class models. In the training phase, for each emotional state $\varepsilon$, feature vectors extracted from the training utterances are used to obtain a Gaussian mixture model (GMM), $\Lambda_\varepsilon$. During the test phase, the models are employed to decide underlying emotion using simple likelihood estimation. Let $\mathbf{Y}$ represent the speech feature vector obtained from the input signal. Then, the recognized emotion for the input signal is the one that maximizes the likelihood function $p(\mathbf{Y}\,|\,\Lambda_\varepsilon)$:

$$\hat{\varepsilon} = \arg\max_{\varepsilon}\, p(\mathbf{Y}\,|\,\Lambda_\varepsilon) \tag{1}$$

## 3 Experiments and evaluations

### 3.1 Experimental setup

To verify the robustness and effectiveness of the PNCC feature, the performance of different features are evaluated in acoustic multi-style emotion identification experiments. Here, a 5-fold cross-validation scheme is used, and the average classification results are computed. To this aim, each classification model is trained on 80% of the total data and tested on the remaining 20%. This process is repeated 5 times (corresponding to 5 folds), each with a different partitioning seed to account for variance between the partitions.

The analysis and experimental results are presented under two different conditions. In one experimental condition, emotion recognitions are performed on clean speech utterances from the speech emotional databases. In the second experimental condition, the experiments are conducted on the noisy speech when 4 different types of noise, including Babble, White, Speech Shaped Noise (SSN), and Factory (Noisex-92 [27]), are added to each utterance at various signal-to noise ratios (SNRs). The effect of noise addition is investigated in 5 dB steps, starting from the -10 dB SNR and terminating at 20 dB SNR.

The block diagram shown in Fig. **1** is used as the emotion recognition system, in which the recognition performances of the MFCC, LPCC, RASTA-PLP, and TEMFCC are evaluated and compared with that of the PNCC Here, the speech frames with the duration of 20 ms at a frame rate of 10 ms are used to extract 13 dime-

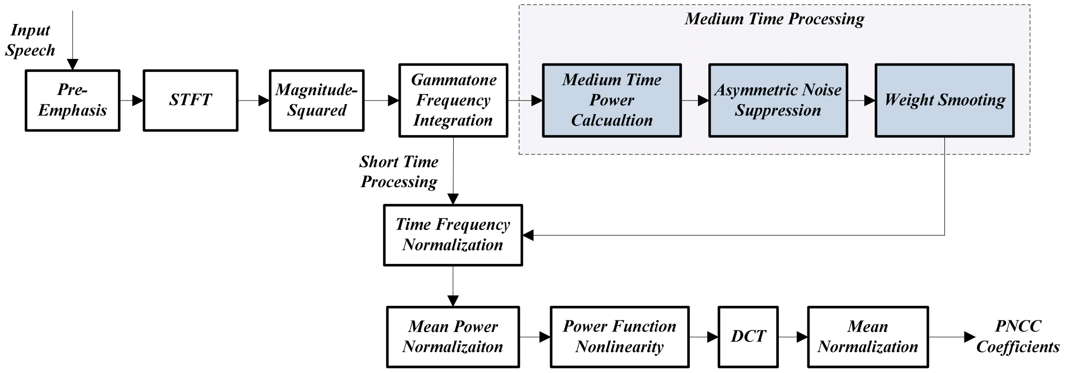

**Fig. 2** The structure of the PNCC feature extraction algorithm.

nsional features. In addition to this, Δ and ΔΔ features are used to construct a final 39-dimensional feature vector. For the extraction of the PNCC, its reference implementation and for the RASTA-PLP, MFCC and LPCC extraction, the most commonly used implementations are used [28-30].

The GMMs used for emotional models are composed of 32 components with diagonal covariance matrices.

### 3.2 Databases

To generalize the results, we use 2 speech corpora in 2 different languages: The Berlin database of German emotional speech (EMO-DB) which has been developed in the Institute for Speech and Communication, at the Berlin Technical University [22] and the Persian emotional speech database (Persian ESD) which has been recorded in a professional recording studio in Berlin, Germany, under the supervision of a linguist expert and an acoustician [23]. The detailed specifications for both of the databases is shown in Table 1.

### 3.3 Experimental results

#### 3.3.1 Emotion recognition under clean conditions

Fig. 3 represents the average emotion recognition accuracies achieved with various features over all emotional states in clean conditions for both of the EMO and Persian ESD databases. As the figure shows, the PNCC, LPCC, and MFCC features lead to almost the same rates, for each database.

All of cepstral-based features (i.e., all except the RASTA-PLP) were implemented and evaluated with and without cepstral mean normalization (CMN). The CMN (also called cepstral mean subtraction (CMS)) is the well-known approach employed to decrease the effects of channel variability in speaker and speech recognition systems. The principle behind cepstral mean normalization is based upon the behavior of the cepstrum under the convolution operation which generates a constant offset in the cepstral domain for linear channel distortion, and the assumption that the channel filter is approximately invariable over the duration of the utterance [31]. The CMN is the final stage in a cepstral-based feature extraction. In Fig. **3**, the results of emotion recognition are shown for features implemented with and without the CMN process. As shown, all features reach their maximum performances in recognition rates when implemented without the CMN. This means that using the CMN does not preserve accurately the emotional content of a speech. The difference between the results with and without the CMN, varies per feature. The maximum variation happens to be for the LPCC whereas the minimum is for the PNCC which means that the PNCC is less affected by using the CMN.

**Table 1:** Specifications of emotional speech databases used in the experiments.

| **Parameters** | | **Values** | |
|---|---|---|---|
| | | **EMO DB** | **Persian ESD** |
| **Number of Utterances** | Anger | 127 | 62 |
| | Boredom | 69 | - |
| | Disgust | 46 | 58 |
| | Fear | 69 | 58 |
| | Happiness | 71 | 58 |
| | Neutral | 79 | 180 |
| | Sadness | 62 | 56 |
| | Total | 535 | 472 |
| **Number of Speakers** | | 10 | 2 |
| **Number of Sentences** | | 10 | 90 |
| **Sample Rate** | | 16 kHz | 44.1 kHz |
| **Language** | | German | Persian |

In this paper, the acoustical features are implemented without incorporating the CMN process to achieve high recognition rates.

As the results show, in general, the recognition rates obtained by 5 features in the case of the EMO DB are lower than the rates obtained with the Persian ESD. This can be justified by considering the number of speakers used in the generation of databases. In contrast to the EMO DB with 10 speakers, in the Persian ESD, 2 speakers were used for the generation of emotional speech utterances. Therefore, a speech emotion recognition system based on the Persian ESD is more speaker-dependent than the system employing the EMO DB.

To explore the recognition accuracy per emotion in clean conditions, the confusion matrices for both of the databases are illustrated for 5 different features in , where the bold numbers represent the recognition accuracies per emotion. It is observed that, in general,

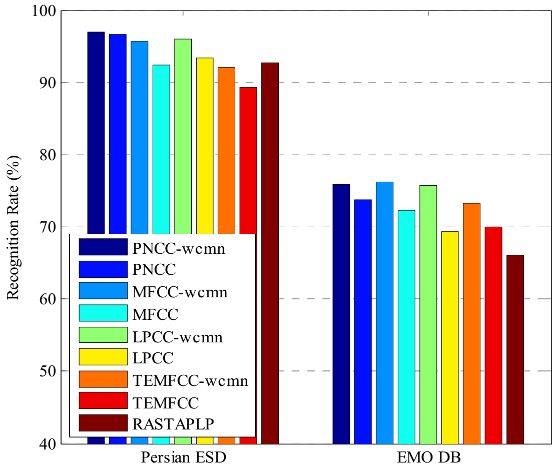

**Fig. 3** Recognition performances obtained by 5 different features in clean conditions for the EMO DB and Persian ESD databases (Note: “wcmn” stand for without cepstral mean normalization).

**Table 2:** Emotion recognition accuracies (%) obtained by different features for the EMO DB and Persian ESD databases in clean conditions.

| | Real Emotions | Classified Emotions | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Persian ESD | | | | | | EMO DB | | | | | | |
| | | An. | Di. | Fe. | Ha. | Ne. | Sa. | An. | Bo. | Di. | Fe. | Ha. | Ne. | Sa. |
| PNCC Feature | Anger | **98** | 0 | 2 | 0 | 0 | 0 | **86** | 0 | 6 | 2 | 6 | 0 | 0 |
| | Boredom | - | - | - | - | - | - | 1 | **84** | 4 | 0 | 0 | 5 | 6 |
| | Disgust | 0 | **100** | 0 | 0 | 0 | 0 | 0 | 0 | **84** | 3 | 0 | 6 | 7 |
| | Fear | 0 | 0 | **100** | 0 | 0 | 0 | 2 | 2 | 2 | **51** | 14 | 17 | 12 |
| | Happiness | 0 | 0 | 4 | **90** | 6 | 0 | 27 | 2 | 3 | 10 | **53** | 3 | 2 |
| | Neutral | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 16 | 2 | 0 | 0 | **78** | 4 |
| | Sadness | 0 | 0 | 0 | 4 | 2 | **94** | 0 | 2 | 0 | 0 | 0 | 3 | **95** |
| | Average Classification Accuracy: **97%** | | | | | | | Average Classification Accuracy: **75.85%** | | | | | | |
| MFCC Feature | Anger | **88** | 0 | 4 | 6 | 2 | 0 | **91** | 0 | 6 | 3 | 0 | 0 | 0 |
| | Boredom | - | - | - | - | - | - | 0 | **70** | 3 | 0 | 0 | 17 | 10 |
| | Disgust | 0 | **100** | 0 | 0 | 0 | 0 | 2 | 10 | **84** | 0 | 0 | 4 | 0 |
| | Fear | 0 | 0 | **100** | 0 | 0 | 0 | 5 | 1 | 2 | **60** | 8 | 15 | 9 |
| | Happiness | 0 | 0 | 4 | **96** | 0 | 0 | 35 | 2 | 3 | 7 | **46** | 5 | 2 |
| | Neutral | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 13 | 2 | 0 | 0 | **82** | 3 |
| | Sadness | 0 | 0 | 0 | 10 | 0 | **90** | 0 | 0 | 0 | 0 | 0 | 0 | **100** |
| | Average Classification Accuracy: **95.67%** | | | | | | | Average Classification Accuracy: **76.15%** | | | | | | |
| LPCC Feature | Anger | **94** | 0 | 6 | 0 | 0 | 0 | **94** | 0 | 0 | 1 | 5 | 0 | 0 |
| | Boredom | - | - | - | - | - | - | 0 | **67** | 2 | 0 | 0 | 18 | 13 |
| | Disgust | 0 | **98** | 2 | 0 | 0 | 0 | 4 | 0 | **82** | 0 | 0 | 8 | 6 |
| | Fear | 0 | 0 | **96** | 0 | 0 | 4 | 8 | 4 | 6 | **65** | 2 | 7 | 8 |
| | Happiness | 0 | 0 | 0 | **88** | 4 | 8 | 37 | 0 | 2 | 5 | **54** | 0 | 2 |
| | Neutral | 0 | 0 | 0 | 0 | **100** | 0 | 2 | 9 | 3 | 8 | 2 | **74** | 2 |
| | Sadness | 0 | 0 | 0 | 0 | 0 | **100** | 0 | 4 | 0 | 0 | 0 | 2 | **94** |
| | Average Classification Accuracy: 96% | | | | | | | Average Classification Accuracy: **75.71%** | | | | | | |
| RASTA-PLP Feature | Anger | **88** | 0 | 6 | 6 | 0 | 0 | **82** | 0 | 1 | 2 | 7 | 4 | 4 |
| | Boredom | - | - | - | - | - | - | 0 | **82** | 4 | 0 | 0 | 6 | 8 |
| | Disgust | 0 | **100** | 0 | 0 | 0 | 0 | 0 | 15 | **76** | 0 | 0 | 2 | 7 |
| | Fear | 2 | 0 | **98** | 0 | 0 | 0 | 9 | 4 | 7 | **37** | 13 | 9 | 21 |
| | Happiness | 4 | 0 | 2 | **90** | 4 | 0 | 46 | 0 | 2 | 10 | **33** | 7 | 2 |
| | Neutral | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 27 | 4 | 2 | 2 | **56** | 9 |
| | Sadness | 0 | 2 | 0 | 16 | 2 | **80** | 0 | 2 | 0 | 0 | 0 | 0 | **98** |
| | Average Classification Accuracy: **92.67%** | | | | | | | Average Classification Accuracy: **66.3%** | | | | | | |
| TEMFCC Feature | Anger | **74** | 0 | 4 | 0 | 22 | 0 | **86** | 2 | 5 | 0 | 5 | 2 | 0 |
| | Boredom | - | - | - | - | - | - | 0 | **77** | 5 | 2 | 0 | 11 | 5 |
| | Disgust | 0 | **100** | 0 | 0 | 0 | 0 | 0 | 2 | **86** | 4 | 0 | 6 | 2 |
| | Fear | 0 | 0 | **100** | 0 | 0 | 0 | 12 | 2 | 4 | **70** | 0 | 10 | 2 |
| | Happiness | 0 | 2 | 10 | **80** | 2 | 6 | 37 | 0 | 6 | 5 | **42** | 8 | 2 |
| | Neutral | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 25 | 6 | 7 | 0 | **62** | 0 |
| | Sadness | 0 | 0 | 0 | 0 | 2 | **98** | 0 | 5 | 0 | 0 | 0 | 5 | **90** |
| | Average Classification Accuracy: **92%** | | | | | | | Average Classification Accuracy: **73.3%** | | | | | | |

the PNCC, MFCC, and LPCC features lead to nearly the same recognition rates for both of corpora. In the case of the Persian ESD, the recognition results are almost the same for Disgust, Fear, and Neutral states, but they differ for Anger, Happiness, and Sadness. For Anger, the highest rate is obtained by the PNCC (98%), as compared with the MFCC attaining the lowest recognition rate (88%). Among other features, the MFCC and LPCC result the highest accuracy for Happiness and Sadness, respectively. For the EMO DB, all features attain the worst result for Fear and Happiness emotional states. The table shows that almost the most confusion occurs for Happiness in which about one-third of the utterances are recognized as anger.

The results of  show clearly that the performances of features for both corpora are different. For example, with the Persian ESD the highest rates for Anger are achieved by the PNCC, but in the case of the EMO DB, the highest recognition rate is obtained by the LPCC. This can be justified by the differences in both languages, in which emotions are conveyed and perceived differently.

On average, the accuracy obtained by the PNCC with the Persian ESD is 97% in clean conditions, which is 1.33% and 1% higher than that achieved with the MFCC and LPCC, respectively. For the EMO DB, the PNCC results in an average accuracy of 75.85% where the performance is 0.3% lower than the MFCC and 0.14% higher than the LPCC.

### 3.3.2 Emotion recognition under noisy conditions

To evaluate the performance of the PNCC for emotion recognition task in noise, 4 different types of noises, including the Babble, White, SSN, and Factory are added to each utterance from the 2 speech corpora at various input SNRs from -10 dB to 20 dB in 5 dB steps. and 5 represent the average recognition results obtained at different SNR levels for the Persian ESD and EMO DB, respectively. In the case of using the Persian ESD, it can generally be seen from  that at each selected SNR level, the PNCC performs the best among all features used. Specifically, at SNR values from 0 dB to 20 dB the PNCC outperforms significantly other features. At low SNRs, the accuracy of the PNCC is, generally speaking, close to other features, but again performing the best in the Babble and White noise conditions.

Fig. 5 shows the recognition results when the EMO DB is used in the experiments. Here, the PNCC still achieves superior recognition performance as compared to the MFCC, LPCC, RASTA-PLP, and TEMFCC.

In summary, as the results of Fig. 4 and Fig. 5 show, in general, the PNCC outperforms other features in the sense of recognition rates. Furthermore, it is observed that the LPCC acts close to the PNCC for the Babble and SSN noises. For other three features, namely, the MFCC, RASTA-PLP, and TEMFCC, the performance varies between the databases used. Based on the results for the Persian ESD, MFCC, RASTA-PLP, and

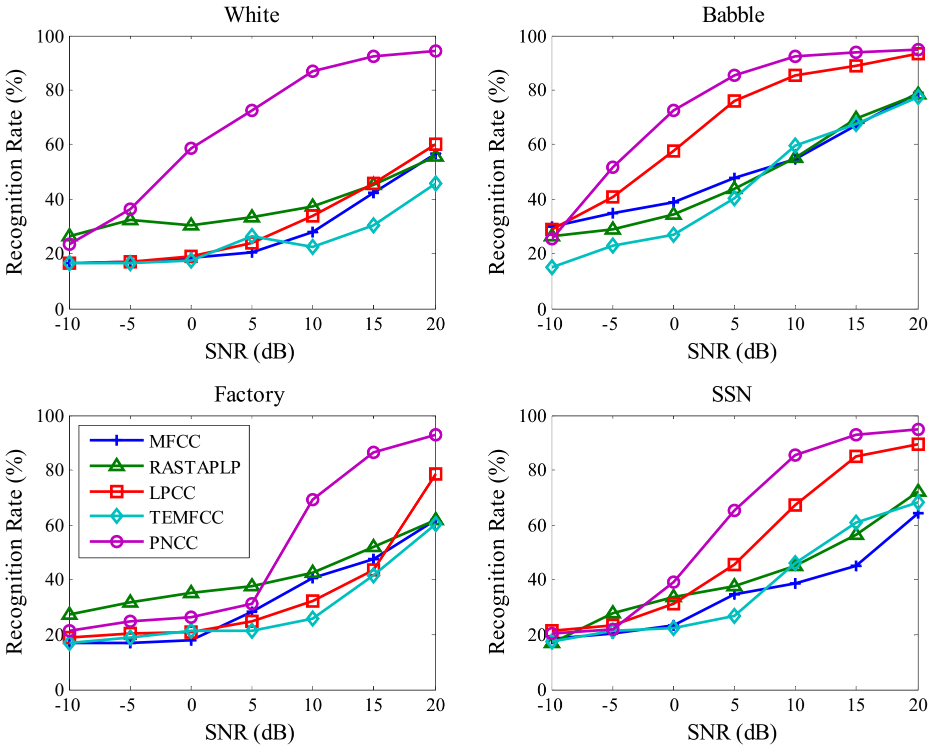

**Fig. 4** Recognition rates (R.R.) obtained by 5 different features under 4 different noise types as a function of input SNR for the Persian ESD.

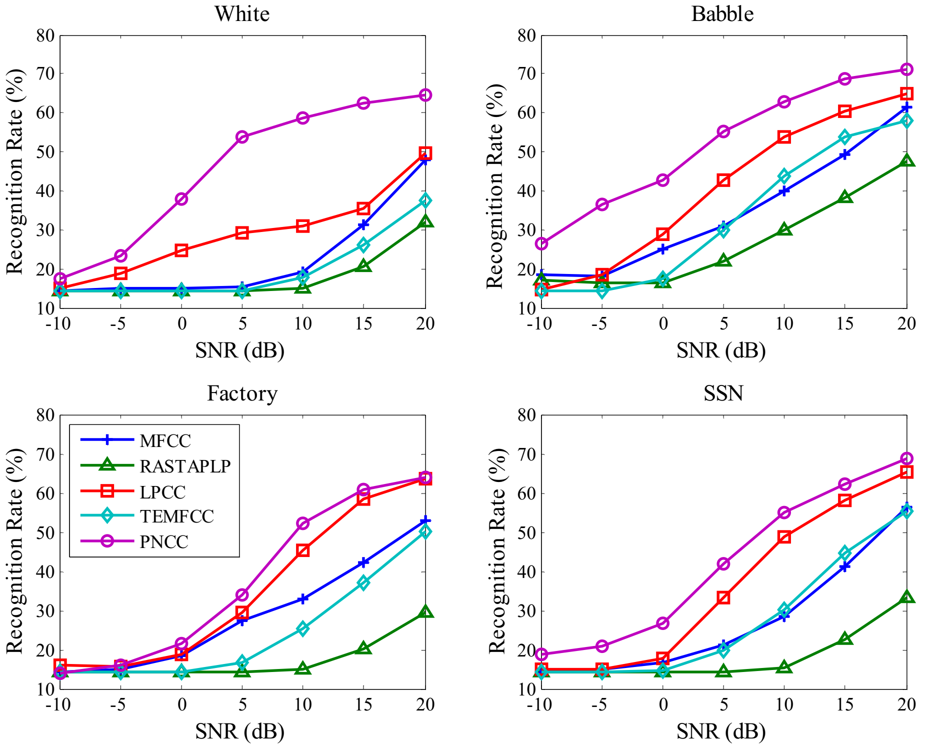

**Fig. 5** Recognition rates (R.R.) obtained by 5 different features under 4 different noise types as a function of input SNR for the EMO DB.

TEMFCC behave almost the same. However, in case of the EMO DB, the performance of these features are more distinguishable. It can be observed clearly from Fig. 5 that, the RASTA-PLP has the worst performance in the sense of noise robustness and the MFCC and TEMFCC achieve nearly the same rates.

Experimental results demonstrate the effectiveness of the PNCC for robust emotion recognition in noise. This is due to incorporating different processing stages in the implementation of the PNCC, including the use of a power-law nonlinearity, employing a noise-suppression algorithm based on asymmetric filtering, and using a module that accomplishes temporal masking.

## 4 Conclusion

In this paper, we addressed the implementation of an automatic emotional-state recognition system using the PNCC feature as a noise robust feature extracted from an audio signal. The emotion recognition performance of this feature was compared with those of the most commonly used features MFCC, LPCC, RASTA-PLP, and also TEMFCC. The experiments were carried out using the EMO DB and Persian ESD speech corpora. The evaluations were performed under two acoustic conditions: clean condition and noisy condition in the presence of 4 different types of noises, including the Babble, Factory, SSN, and White at various input SNR levels. Under the clean experimental condition, the simulations show that all of the features, (with the exception of RASTA-PLP) reach their maximum recognition performances when implemented without the CMN processing. In this acoustic condition, the performance of the PNCC (97% for Persian ESD, 75.85% for EMO DB) was approximately the same as for the MFCC (95.67% for Persian ESD, 76.15% for EMO DB) and a little bit better than the LPCC (96% for Persian ESD, 75.71% for EMO DB) but it was considerably higher than the recognition rates for the RASTA-PLP and TEMFCC. For the noisy condition, the PNCC provides substantial improvements (maximal improvement of 45 % for white noise and 20 % for babble, in the case of EMO DB) in the emotion recognition accuracy compared with the MFCC, RASTA-PLP, LPCC, and TEMFCC. The results of experiments conducted for EMO DB and Persian ESD databases approve the robustness and effectiveness of the PNCC for speech emotion recognition in both clean and noisy conditions. As future work, the authors plan to consider more auditory-based features for the representation of emotional states and more realistic scenarios, including reverberation for the task of emotion classification.

## References

[1] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE*

*Transactions on Speech and Audio Processing,* Vol. 13, pp. 293-303, 2005.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication,* Vol. 48, pp. 1162-1181, 2006.

[3] M. Shami and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication,* Vol. 49, pp. 201-212, 2007.

[4] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," *Speech Prosody, Dresden,* pp. 276-289, 2006.

[5] S. Karimi and M. H. Sedaaghi, "Robust emotional speech classification in the presence of babble noise," *International Journal of Speech Technology,* Vol. 16, pp. 215-227, 2013.

[6] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 1653-1656, 2006.

[7] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing,* Vol. 71, pp. 1913-1920, 2008.

[8] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Computing and Applications,* vol. 24, pp. 1539-1553, 2014.

[9] A. Georgogiannis and V. Digalakis, "Speech emotion recognition using non-linear teager energy based features in noisy environments," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pp. 2045-2049, 2012.

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 28, pp. 357-366, 1980.

[11] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, Y.-M. Cheng, and C. S. Chien, "Feature combination for better differentiating anger from neutral in mandarin emotional speech," in *Affective Computing and Intelligent Interaction*, ed: Springer, pp. 741-742, 2007.

[12] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* Vol. 87, pp. 1738-1752, 1990.

[13] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.

[14] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and W.-Y. Liao, "Combining acoustic features for improved emotion recognition in mandarin speech," in *Affective Computing and Intelligent Interaction*, ed: Springer, pp. 279-285, 2005.

[15] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4101-4104, 2012.

[16] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH*, pp. 28-31, 2009.

[17] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *Speech and Audio Processing, IEEE Transactions on,* Vol. 7, pp. 55-69, 1999.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing, ,* Vol. 2, pp. 578-589, 1994.

[19] U. H. Yapanel and J. H. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication,* Vol. 50, pp. 142-152, 2008.

[20] G. Sárosi, T. Mozsolics, B. Tarján, A. Balog, P. Mihajlik, and T. Fegyó, "Recognition of multiple language voice navigation queries in traffic situations," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, ed: Springer, pp. 199-213, 2011.

[21] F. Kelly and N. Harte, "Auditory Features Revisited for Robust Speech Recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 4456-4459, 2010.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, pp. 1517-1520, 2005.

[23] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klann-Delius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," *Behavior research methods,* pp. 1-20, 2014.

[24] S. Pazhanirajan and P. Dhanalakshmi, "EEG Signal Classification using Linear Predictive Cepstral Coefficient Features," *International Journal of Computer Applications,* Vol. 73, 2013.

[25] B. C. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acta Acustica united with Acustica,* Vol. 82, pp. 335-345, 1996.

[26] E. Principi, S. Squartini, and F. Piazza, "Power Normalized Cepstral Coefficients based supervectors and i-vectors for small vocabulary speech recognition," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, pp. 3562-3568, 2014.

[27] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A

database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication,* Vol. 12, pp. 247-251, 1993.

[28] C. Kim. (2012). *Open Source MATLAB Code for PNCC.* Available: http://www.cs.cmu.edu/~robust/archive/algorithms/PNCC_IEEETran.

[29] D. Ellis. (2006). *PLP and RASTA (and MFCC, and inversion) in MATLAB using melfcc.m and invmelfcc.m.* Available: http://labrosa.ee.columbia.edu/matlab/rastamat/

[30] J. Lyons. (2013). A set of speech feature extraction functions for ASR and speaker identification written in matlab. Available: https://github.com/jameslyons/matlab_speech_features

[31] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication,* Vol. 53, pp. 991-1001, 2011.

**Meysam Bashirpour** was born in Tabriz, Iran, in 1984. He received the B.Sc. degree in Electronics Engineering from the Sharif University of Technology in 2007 and M.Sc. degree in Communication Engineering from the University of Tehran in 2010. He is currently pursuing the Ph.D. degree in the Faculty of Electrical and Computer Eng. at the University of Tabriz, Tabriz, Iran. His current research interests are focused on Speech Recognition, Binaural Signal Processing, Emotion Recognition from Speech, and Pattern Classification.

**Masoud Geravanchizadeh** received the B.Sc. degree in Electronics Engineering from the University of Tabriz, Tabriz, Iran, in 1986, and the M.Sc. and Ph.D. degrees in Signal Processing from the Ruhr-University Bochum, Bochum, Germany, in 1995 and 2001, respectively. Since 2005, he has been with the Faculty of Electrical and Computer Eng. at the University of Tabriz, Tabriz, where he is currently an Associate Professor. His research interests include Binaural Signal Processing, Auditory-based Emotional Speech Recognition, Improvement of Speech Quality and Intelligibility for Normal Hearing and Hearing-Imparied Listeners, Sound Source Localization and Separation, Pattern Classification, and Stochastic Signal Processing.