FEATURE EXTRACTION ALGORITHM USING NEW CEPSTRAL TECHNIQUES FOR ROBUST SPEECH RECOGNITION

Mohamed Cherif Amara korba[1*], Houcine Bourouba[2], Rafik Djemili[3]

[1]Department of Electrical Engineering, Institute of Science and Technology, Souk Ahras University, Algeria

[2]Department of Telecommunication, Institute of Science and Technology, Guelma University, Algeria

[3]Department of Electrical Engineering, Institute of Technology, Skikda University, Algeria

[1,2,3]PI : MIS Laboratory, Guelma, Algeria

Email: amara.korba.cherif@gmail.com[1*](corresponding author), bourouba2004@yahoo.fr [2], djemili_rafik@yahoo.fr [3]

DOI: https://doi.org/10.22452/mjcs.vol33no2.1

ABSTRACT

In this work, we propose a novel feature extraction algorithm that improves the robustness of automatic speech recognition (ASR) systems in the presence of various types of noise. The proposed algorithm uses a new cepstral technique based on the differential power spectrum (DPS) instead of the power spectrum (PS), the algorithm replaces the logarithmic non linearity by the power function. In order to reduce cepstral coefficients mismatches between training and testing conditions, we used the mean and variance normalization, then we apply auto-regression moving-average filtering (MVA) in the cepstral domain. The ASR experiments were conducted using two databases, the first is LASA digit database designed for recognition the isolated Arabic digits in the presence of different types of noise. The second is Aurora 2 noisy speech database designed to recognize connected English digits in various operating environments. The experimental results show a substantial improvement from the proposed algorithm over the baseline Mel Frequency Cepstral Coefficients (MFCC), the relative improvement is the 28.92% for LASA database and is the 44.43% for aurora 2 database. The performance of our proposed algorithm was tested and verified by extensive comparisons with the state-of-the-art noise-robust features in aurora 2.

Keywords: Noise-robust feature, Automatic speech recognition, Differential power spectrum, MVA, Aurora2

1.0 INTRODUCTION

Robustness with respect to noise is an important issue in automatic speech recognition (ASR) systems. Most of them are sensitive to environmental conditions in which they are used. The degradation of the ASR performances is quite significant when the mismatch between the training and testing conditions is large. Under realistic noise conditions, noise corrupts the speech samples and causes a mismatch due to the distortion in speech features. Most of ASR systems use the common MFCC [1] features as a baseline. However, it is established that the use of MFCC is not appropriate in noisy speech environments. For that reason, several variants of MFCC have been proposed to improve the ASR system's robustness [2, 3]. There are other methods based on Perceptual Prediction which are better under noisy conditions, like perceptual linear prediction (PLP) or relative spectra linear prediction (RASTA-PLP) [4]. The latter method applies a RASTA filter, which is mainly attended for convolutive noise.

There are several compensation methods developed to improve the environmental robustness of ASR systems such as cepstral mean normalization (CMN) [5], and variance mean normalization (VMN) [6, 7]. Both compensation methods are efficient in the presence of quasi-stationary noise. These methods assume that the mean or the mean and variance of the cepstral vectors should be the same for all utterances. There is a more effective compensation method called (MVA) [8, 9, 10] which combines the two methods namely; CMN and VMN plus ARMA filtring. The MVA technique is directly applied as post processing in the cepstral domain. In our approach, we used the MVA technique as an integral part of our algorithm and not as post processing phase. Recent works in the literature have shown significant progress in robust ASR by applying deep neural network (DNN) [11] and very deep convolutional neural networks (CNNs) [12, 13], CNN have shown to provide better performance than traditional DNN.

In this paper, we propose a novel noise-robust speech features extraction algorithm that improves the robustness of ASR systems. The main contributions of this paper are as follows:

- We use the cepstrum derived from differential power spectrum (DPS) instead of power spectrum for better resilience to acoustical distortions.
- To provide superior robustness, we use the power-law nonlinearity (PLN) functions instead of log nonlinearity functions used in MFCC, by suppressing small signals and their variability.
- We use the MVA technique to alleviate distortion in speech features, it has been shown in [8, 9, 10, 14] that it improves significantly the robustness of small-vocabulary ASR tasks.
- We compare the proposed feature extraction algorithm with a set of robust acoustic features that demonstrate appreciable robustness to different type of noise such as: minimum mean-square error (MMSE) spectral amplitude estimator [15], Power Normalized Cepstral Coefficient (PNCC) features [16], and Normalized modulation cepstral coefficients (NMCC) [17] features.The rest of this paper is organized as follows. In section 2, we present a detailed structure of the extraction procedure of the Proposed Noise Robust Features (PNRF). In section 3, we present a graphical comparison under different noise levels between our method extraction and chosen front-ends MFCC, RASTA-PLP and PNCC. In section 4, we present the experimental results and we conclude with illustrative accuracy speech recognition results for state-of-the-art ASR algorithms. In section 5, the work is concluded.

2.0 STRUCTURE OF THE PROPOSED FEATURE EXTRACTION TECHNIQUE

In this section we present the different processing stages leading to the PNRF coefficients. The structure of PNRF can be decomposed in three main phases of processing.

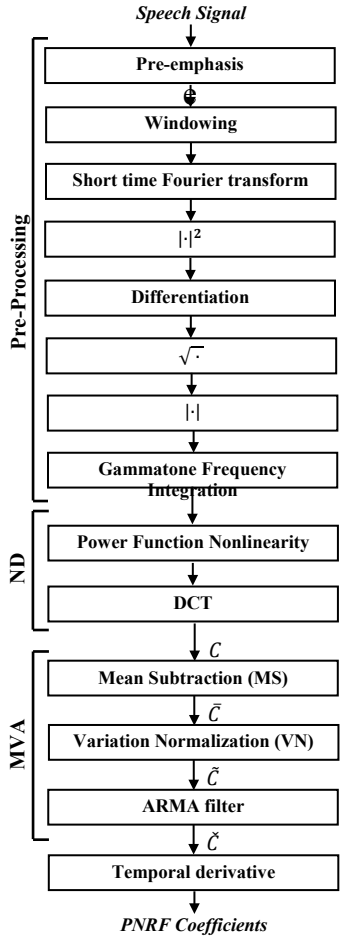


Fig. 1: Structure of the proposed feature extraction algorithm

2.1 Pre-processing Phase

Fig. 1 shows the structure of The PNRF approach. As in the case of MFCC processing, we apply a pre-emphasis filter of the form $H(Z) = 1 - 0.97Z^{-1}$ to input speech signal in the time domain to increase the high frequencies. A short time Fourier transform (STFT) is performed using a Hamming window with frame duration of 25.6ms and a frame shift of 10ms. DPS of speech signal $x(t)$ is obtained by differentiating the power spectrum with respect to frequency, it is defined by

$$D(\omega) = \frac{dX(\omega)}{d\omega} \qquad (1)$$

Where ω denotes frequency. There are several discrete counterparts for approximating of the formula above; we chose ….the difference equation given by

$$d(k) = x(k) - x(k+1) \qquad (2)$$

It was shown in [18] that this form of approximation gives better results. Then, an absolute operation is applied to DPS to make its negative parts positive. Finally, a normalized filter-bank based on 40 channel gammatone-shaped is applied to DPS; the filter-bank is applied between 130Hz and 6800Hz, whose center frequencies are linearly spaced in Equivalent Rectangular Bandwidth (ERB) [19]. ERB scale is defined by

$$ERBS(f) = 21.4 \cdot log(0.00437f + 1) \qquad (3)$$

The gammatone modeling widely used is a physiologically motivated technique that may be considered as an approximation of human cochlear filter-bank. We use Snaley's auditory toolbox [20] to get the impulse response of gammatone filter. In each channel the area under the squared transfer function is normalized to unity.

$$\int_{130}^{6800} |H_m(f)|^2 df = 1 \qquad (4)$$

Where $H_m(f)$ is the frequency response of the m^{th} gammatone channel. We obtain the short-time spectral power $P(l,m)$ using the squared gammatone summation as below:

$$P(l,m) = \sum_{k=0}^{\left(\frac{K}{2}\right)-1} \left[\left|d_l(e^{j\omega_k})\right| \cdot H_m(e^{j\omega_k})\right]^2 \qquad (5)$$

l frame indices, $d_l(e^{j\omega k})$ is the short time spectrum of the l^{th} frame of the signal, m gammatone channel indices, K is the discrete Fourier transform size, it is equal to 1024.

2.2 Non-Linerarity and DCT

Spectral power is enhanced by the expression

$$P_N(l,m) = P(l,m) \cdot 10^{\gamma} \qquad (6)$$

Where γ is empirically defined, it is equal to 4. We apply nonlinearity by exploiting the power-law nonlinearity instead of log nonlinearity. This technique is supposed to eliminate the weak signals and their variability, it provides more robustness to ASR [16]. The nonlinearity is obtained by the following formula

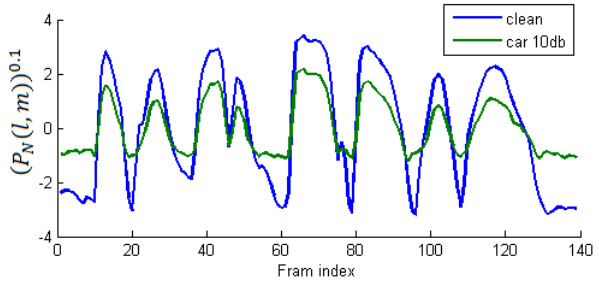
$$P'(l,m) = (P_N(l,m))^{0.1} \qquad (7)$$

Fig. 2 illustrates the effect of the power-law non linearity. We observe that the curve of the DSP $P'(l,1)$ after power-law non linearity of noisy signal follows same variation of the DSP of clean signal. However, the difference between the DSP $P'(l,1)$ after applying the logarithmic nonlinearity is pronounced in noisy environments.

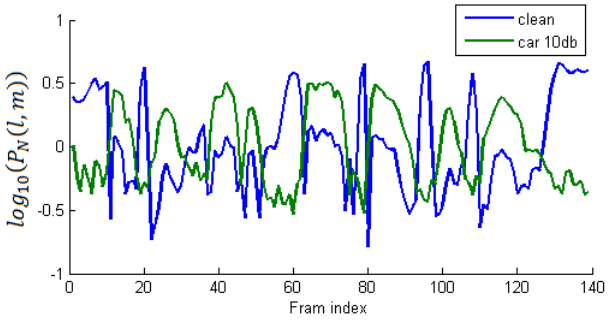
Cepstral parameters are obtained from the spectral power $P'(l,m)$ using the Discrete Cosine Transform (DCT)

$$C(l,k) = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} P'(l,m) cos\left(\frac{\pi k}{M}\left(m - \frac{1}{2}\right)\right), \quad m = 1 \cdots M, \; k = 1 \cdots K \qquad (8)$$

Where M is the number of gammatone filterbank channels, M and K are identical.



(a)



(b)

Fig. 2: Graphical time sequences of the first differential spectral power for the digit string "8375O33" for clean speech and for speech corrupted by car noise, a) the first DSP $P'(l,1)$ after power-law nonlinearity, b) the first DSP after applying the logarithmic nonlinearity.

2.3 Mean subtraction, Variance normalization, and ARMA filtering Phase

The MVA post-processing technique involves three steps, consisting of mean subtraction, variance normalization and ARMA filtering. For a given utterance, it can be represented by a $N \times T$ matrix of frames, say C, where each colomn represents the feature vector for a given frame and each row represents the time sequences of a given coefficient.

$$\begin{bmatrix} C_1(1) & \cdots & C_1(T) \\ \vdots & \ddots & \vdots \\ C_N(1) & \cdots & C_N(T) \end{bmatrix} \quad (9)$$

The first step is mean subtraction (MS) defined by

$$\bar{C}_n(t) = C_n(t) - \mu_n \quad (10)$$

Where $C_n(t)$ is the n^{th} component of the feature vector at time t, $\bar{C}_n(t)$ is the mean subtracted feature and μ_n is the mean vector estimated from data as

$$\mu_n = \frac{1}{T}\sum_{t=1}^{T} C_n(t) \quad (11)$$

The second step is variance normalization (VN) defined by

$$\tilde{C}_n(t) = \frac{\bar{C}_n(t)}{\sqrt{\sigma_n}} \quad (12)$$

where σ_d is the variance estimated from data as

$$\sigma_n = \frac{1}{T}\sum_{t=1}^{T}(C_n(t) - \mu_n)^2 \quad (13)$$

Where $\tilde{C}_n(t)$ is the mean-subtracted and variance normalized feature at time t. The third step is the ARMA filtering process . It is used as an effective technique for smoothing the speech features in the time domain, with the objective of making the noisy frame features much similar to clean features.

In this study we have used non causal ARMA filter, which is defined as

$$\check{C}_n(t) = \begin{cases} \frac{\sum_{i=1}^{Q} \tilde{C}_n(t-i) + \sum_{j=0}^{Q} \tilde{C}_n(t+j)}{2Q+1} & if Q < t \leq T - Q \\ \tilde{C}_n(t) & Otherwise \end{cases} \quad (14)$$

Where Q is the order of ARMA filter.The filter structure is based on a study developed in [9], where it has been shown that non causal ARMA filter provides better performance to ASR system with respect to other types of filters. We apply the MVA to static features which are concatenated with their first and second order time derivatives.

3.0 GRAPHICAL COMPARISON BETWEEN DIFFERENT FEATURES

The Fig. 3 shows a comparison between time sequences of MFCC, Rasta-PLP, PNCC and PNRF features of the utterance of digit string "7936596" corrupted by additive noise at different Signal-to-Noise Ratio (SNR) levels. Table 1 contains the configuration of the features used in the graphical illustration. We noticed a sharp degradation of baseline MFCC features in the presence of noise; RASTA-PLP and PNCC are more resistant to noise compared to MFCC features, whereas PNRF prevails at elevated noise levels. For SNR ≥ 0dB we can clearly see that PNRF is better in terms of noise robustness than MFCC, RASTA-PLP and PNCC features. We used the RASTA-PLP and PNCC implementation that are available respectively in [21, 22].

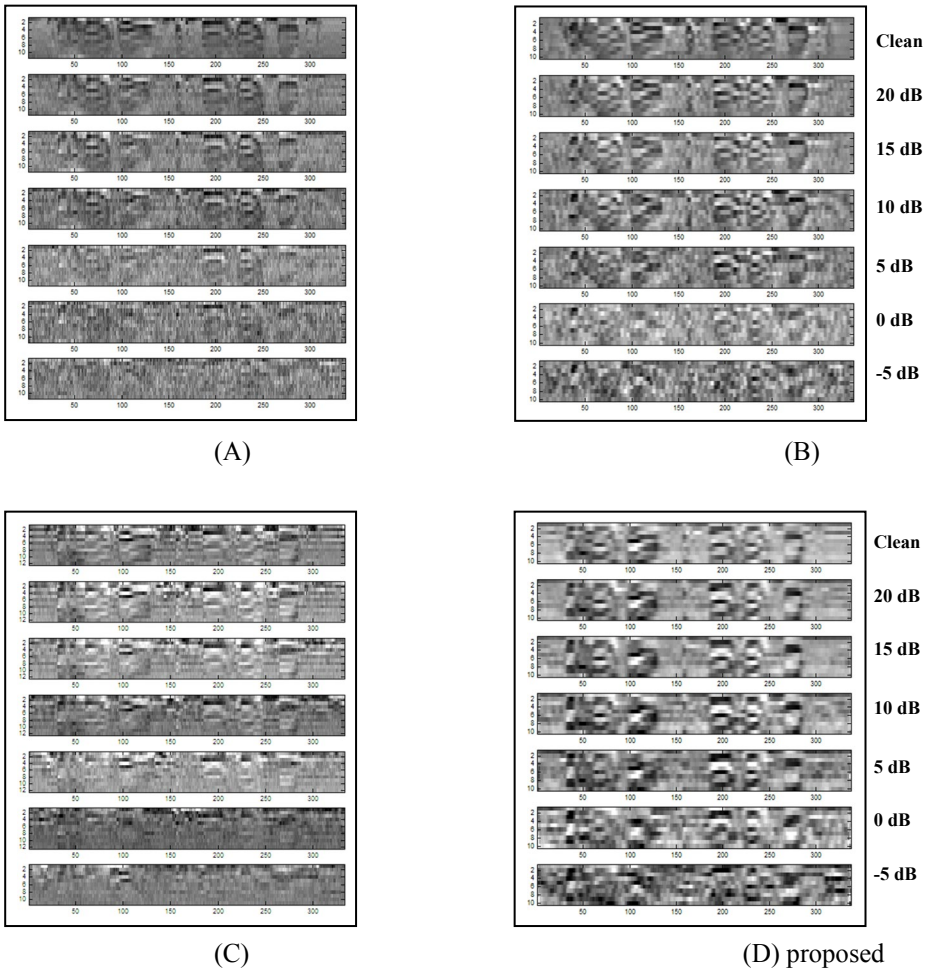


Fig. 3: Graphical time sequences of speech features for the digit string "7936596" , A) MFCC features with MS, B) RASTA-PLP features with MS, C) PNCC features with MS, D) PNRF features. The first coefficient has been omited for all the representations (table 1 shows configuration features).

Table 1: Features parameters used for experimental analysis

Configuration features	MFCC	RASTA-PLP	PNCC	PNRF
Frame length (ms)	25	25	25.6	25.6
Frame shift (ms)	10	10	10	10
FFT size	200	256	1024	1024
Nbr of filter-banks	23	/	40	40
Nbr of coefficients	12	12	13	13
Appended log frame energy	do	do	/	/
Size Δ/ΔΔ window (frames)	3/2	3/2	3/2	3/2
Feature dimension	39	39	39	39

4.0 EXPERIMENTAL RESULTS

In this section, we describe the speech databases used for ASR experiments, and then we present the experimental results to evaluate speech recognition performances of the PNRF features under a variety of acoustical environments. We conducted the experiments using two speech databases, each of them covers a different ASR topology: LASA database is isolated Arabic digit database which was designed for the training and evaluation of ASR algorithms [23]; Aurora 2 database is English connected digit recognition task [24].

4.1 Evaluation on LASA Database

The speech database contains a set of isolated digit utterances spoken in Arabic, developed by the LASA laboratory. It contains 9,000 utterances produced by 90 adult speakers, the vocabulary consists of Arabic digits from 0 to 9. Each speaker repeats each digit 10 times. The test set speech corrupted by four background noises taken from the NOISEX-92 [25] database. Noises are artificially added to the LASA database at different SNR levels, the SNR is defined as the ratio of signal to noise energy. Speech database includes two test mode defined as below:

- **Test mode A**, There are 10 utterances of each digit from each speaker: 6 of them are used for training and 4 remaining are used for testing, what gives 5,400 utterances for clean training, and 3,600 utterances affected by four additive noises (white, pink, factory1 and F16) in range of 7 SNRs values (20, 15, 10, 5, 0, -5dB and Clean) giving a total of 100800 ($3600 \times 4 \times 7$) test utterances.

- **Test mode B**, The training set includes 6,000 speech utterances produced by 60 speakers, the test set includes 3,000 utterances produced by 30 other speakers who have not contributed in the training phase. The same noises as in the test A are added to the test set utterances with the 7 level SNRs giving a total of 84000 ($3000 \times 4 \times 7$) test utterances.

A recognition system was developed using Hidden Markov Model (HMM) toolkit (HTK) [26], each digit HMM has thirteen emitting states with three Gaussian mixtures per state. In all our experiments, acoustic model training is performed on clean speech utterances.

Table 2 shows the results obtained with the PNRF features with different orders of the ARMA filter, the optimal order of the filter is determined empirically by varying the order until 9. We can see clearly that the optimal value of Q is the same for both tests mode A and B, the order $Q = 2$ yields the best results, this value gives a good balance between noise robustness and information preservation.

Table 2: Accuracy percentages on LASA with different orders of the ARMA filter

Filter order	Test mode A			Test mode B		
	clean	20-0dB	-5dB	clean	20-0dB	-5dB
1	99.22	95.77	67.51	97.47	91.70	60.41
2	**99.36**	**96.02**	**70.02**	**97.07**	**91.78**	**60.84**
3	99.19	95.49	68.83	95.57	91.60	60.72
4	98.78	94.81	65.98	96.83	91.28	62.88
5	98.75	94.54	66.15	96.47	90.27	61.10
6	98.64	94.24	64.36	96.07	89.60	60.28
7	98.64	93.76	62.09	96.23	89.59	59.72
8	98.25	93.44	63.12	95.83	89.17	57.65
9	98.00	92.80	61.85	95.03	88.19	56.93

In order to show the performance of PNRF algorithm on the LASA database, comparison is made against the extraction methods frequently used in isolated word recognition systems. Tables 3 and table 4 show the detailed experimental results obtained in test mode A and test mode B respectively. The table column denoted by 'Avg 0–20dB' expresses the results recognition rate averaged from 0 dB to 20 dB over a four noise types. The fourth column of the two tables below contains values of a Relative Improvement (Rel_Imp) of the proposed algorithm extraction methods frequently used in ASR systems. Relative improvement is defined as

$$Rel\_Imp = \frac{r_p - r_t}{r_t} \times 100\% \qquad (15)$$

r_p is the recognition rate of our proposed algorithm and r_t is the recognition rate of comparison feature extraction methods.

In test mode A, at the SNR 0-20dB, PNRF with ($Q = 2$) achieve an error rate reduction of 21.54% as compared to MFCC baseline, 22.67% as compared to RASTA-PLP, 10.28% as compared to MFCC+MVA and 3.22% as compared to PNCC feature. In test mode B, at the SNR 0-20dB, PNRF with ($Q = 2$) gives 21.22% error rate reduction as compared to MFCC baseline, 21.21% as compared to RASTA-PLP, 9.41% as compared to MFCC+MVA and 1.79% compared to PNCC feature. PNRF gives a relative improvement over the baseline MFCC of 28.92% in the test mode A, and 30.07% in the test mode B. In the clean test, PNRF feature does not affect the performance of the recognition like the other features (MVA, RASTA-PLP and PNCC) on the contrary, it yields more performance as the standard MFCC features for both test modes. At very low SNR -5db, the PNRF feature gives a substantial improvement accuracy recognition compared to the standard MFCC, more than 42.49% in test mode B and more than 48.76% in test mode A.

Table 3: Test mode A word accuracy (%) using different feature sets

Features	Clean	AVG 0-20dB	Rel_Imp	-5 dB
MFCC	99.25	74.48	28.92	21.26
RASTA_PLP	98.11	73.35	30.90	19.43
MFCC + CMN	99.00	82.03	17.05	22.12
RASTA_PLP + CMN	98.42	75.73	26.79	19.06
MFCC + MVN	98.89	85.48	12.33	40.17
MFCC + MVA	99.00	85.74	11.99	42.79
PNCC	98.34	92.80	3.46	58.47
PNRF (ord2)	**99.36**	**96.02**	–	**70.02**

Table 4: Test mode B word accuracy (%) using different feature sets

Features	Clean	AVG 0-20dB	Rel_Imp	-5 dB
MFCC	96.63	70.56	30.07	18.35
RASTA_PLP	97.23	79.57	15.34	21.65
MFCC + CMN	97.67	82.00	11.92	36.27
RASTA_PLP + CMN	95.43	76.57	19.86	23.50
MFCC + MVN	95.40	80.07	14.62	25.26
MVA	97.43	82.36	11.43	38.99
PNCC	95.29	89.99	1.99	55.73
PNRF (ord2)	**97.07**	**91.78**	–	**60.84**

4.2 Evaluation on AURORA 2 Database

Aurora 2 database contains a set of continuous digit strings spoken in English; it contains two training modes: clean condition training and multi-condition training. The first mode uses clean speech data only for acoustic model training, while the second mode uses both clean and noisy speech data for acoustic model training. The training data contain 8440 utterances spoken by 110 adult speakers. Test data contain 4,004 utterances spoken by 104 author speakers who have not contributed in the training phase. Database contains 3 test sets: test set A, test set B and test set C. The first two test sets are corrupted by 4 different types of noise, where different noises have been artificially added. The last test set is corrupted by 2 types of noise. The speech and the noise signals are filtered with a G.712 frequency characteristic except the signals of test set C filtered with the MIRS (modified intermediate reference system) characteristic before adding the two types of noise (see table 5).

In our experiments, we use HMM model with the same configuration as [24]. Each digit HMM has 16 emitting states with three Gaussian mixtures per state. Silence and short pause models have three and one stat respectively, with six Gaussian mixtures per state.

Table 5: A summary description of Aurora 2 database

Training data	Clean condition training	Training utterances: 8,440
	multi-condition training	8,440 training utterances are split into 20 subsets with 422 utterances in each subset. Additive noises: suburban train, babble, car, and exhibition hall. SNR levels: clean, 20 dB, 15 dB, 10 dB and 5 dB.
Test set	Set A	Test utterances: 28,028 Additive noises: babble, car, suburban train, and exhibition hall SNR levels: clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB.
	Set B	Test utterances: 28,028 Additive noises: street, restaurant, airport and train station SNR levels: clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB.
	Set C	Test utterances: 14,014 Additive noises: suburban train and street SNR levels: clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB.

Table 6 shows the effect of ARMA filter order on the recognition accuracy for Aurora 2 database obtained with clean training condition. We changed the order Q of the non causal ARMA filter until 9.
We note that, small Q gives better recognition performance in clean test subset, because it preserves the short-term cepstral information. However, it gives bad performance in noisiest test subset (-5dB), because the short-term cepstral information is more sensitive to noise. The reverse is true. The best performance of the system has been obtained with order $Q = 6$, this value establish a good balance between information preservation and noise robustness.

For clean set, accuracy rate is 99.11%. At Avg 0-20, the average accuracy rate is 85.33%, which is very encouraging. At the SNR -5 dB accuracy recognition rate is 27.42%.

Table 6: Accuracy percentages on aurora 2 with different orders of the ARMA filter

Filter order	**Clean**	**Avg 0-20db**	**-5 db**
2	99.29	81.20	16.88
3	99.24	81.52	17.86
4	99.21	81.81	17.72
5	99.15	84.37	21.80
6	**99.11**	**85.33**	**27.42**
7	98.91	85.44	29.27
8	**98.77**	**85.72**	**32.41**
9	98.49	85.02	32.59

Table 7 shows that PNRF performs very well in the case of speech corrupted by car noise, which present some stationarity. The average recognition rate for the three test sets is given by the last column.

Table 7: Detailed recognition rate (%) for the PNRF feature ($Q = 6$) for aurora 2 database

	Set A				Set B				Set C		
	Subway	**Babble**	**Car**	**Exhibition**	**Restaurant**	**Street**	**Airport**	**Station**	**Subway**	**Street**	**Avg**
Clean	99.23	98.97	99.02	99.35	99.23	98.97	99.02	99.35	99.14	98.79	**99.11**
20dB	97.36	97.73	97.64	97.19	98.07	97.40	98.18	98.18	97.21	97.10	**97.61**
15dB	95.67	96.13	96.48	95.46	96.04	95.95	96.54	96.70	95.36	95.59	**95.99**
10dB	91.25	93.02	93.86	91.24	91.71	92.44	93.77	93.18	90.08	91.41	**92.20**
5dB	80.26	80.89	88.34	78.65	77.89	82.01	84.25	83.77	79.24	81.53	**81.68**
0dB	55.05	52.72	75.48	58.35	48.94	61.12	62.21	63.10	54.16	60.43	**59.16**
-5dB	23.83	18.35	42.38	28.88	17.47	30.32	26.78	33.23	23.30	29.66	**27.42**
Avg 0–20dB	**83.92**	**84.10**	**90.36**	**84.18**	**82.53**	**85.78**	**87.00**	**86.99**	**83.21**	**85.21**	**85.33**

4.3 Comparison With Other Methods

In order to evaluate the effectiveness of the proposed algorithm, several comparisons are made against the MFCC baseline [1], RASATA-PLP [4], MFCC+MVA [10] and minimum mean-square error (MMSE) spectral amplitude estimator [15]. We added to our comparison two newly effective techniques [27, 28], the first technique denoted (TECH1) proposes the implementation of the 2D psychoacoustic models to MFCC features and the second technique denoted (TECH2) investigates the distribution of Mel-filtered log-spectrum of speech signals in noisy environments.

Table 8: Recognition results for state-of-the-art ASR algorithms (%)

SNR(dB)	**Clean**	**Avg 0-20dB**	**Rel-Imp**	**-5dB**
MFCC + CMN	98.89	59.08	44.43	8.74
RASTA-PLP + CMN	98.93	68.27	24.99	10.24
MFCC + CMVN	99.29	79.12	7.85	14.93
MMSE	99.26	80.12	6.50	20.31
PNCC	99.32	82.90	2.93	18.13
MFCC + MVA	99.20	83.60	2.06	24.50
NMCC	99.09	83.83	1.79	*
TECH1	99.26	84.38	1.12	24.73
TECH2	99.20	84.89	0.52	24.70
PNRF (proposed)	**99.11**	**85.33**	–	**27.42**

*: the value was not mentioned by the author in [17].

Table 8 presents the advantages of the proposed algorithm compared to other comparison targets, it shows statistical results in terms of recognition accuracy rates (%) with the different methods. The proposed algorithm PNRF gives better recognition rate for SNR for Avg 0-20 dB and -5dB. Even for clean set, accuracy rate is higher than 99%. At Avg 0-20 dB, the relative improvements are 44.43% over MFCC+CMN, 24.99% over RASTA-PLP+CMN, 7.85% over MFCC+CMVN, 6.5% over MMSE, 2.93% over PNCC, 2.06% over MFCC+MVA, 1.12% over TECH1 and 0.52% over TECH2.

For the SNR -5 dB, our proposed algorithm performs better than all algorithms, the improvements are much more significant, 18.68% over MFCC-CMN, 17.18% over RASTA-PLP+CMN, 9.29% over PNCC, 7.11% over MMSE and 2.92% over MFCC+MVA.

We see well that, the proposed algorithm has a major contribution compared to the target algorithms, more especially, in the noise interest interval which varies between 0 and 20dB.

5.0 CONCLUSION

In this paper, a new cepstral features have been proposed by introducing a differential power spectrum and power-law nonlinearity that replaces the existing technique of log nonlinearity used in MFCC processing. The proposed feature used the MVA technique as an integral part of the algorithm and not as post-processing phase. It was verified that the effectiveness of the proposed feature depends on the choice of ARMA filter order, the optimal order of the latter is equal to 2 for the recognition of the isolated digits and 6 for the recognition of the connected digits. According to the experimental results conducted on the Aurora 2 and LASA database, the proposed feature provides a significant improvement in robustness, it outperforms state-of-the-art algorithms in noisy environments without losing performance in clean environment as well. The proposed feature is effective and slightly greater computational cost than baseline MFCC processing. It can be used easily in real-time embedded systems. The future work seeks to experiment our algorithm on large vocabulary ASR databases such as DARPA Wall Street Journal.

REFERENCES

[1] Davis S., & Mermelstein P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no 4, pp. 357-366, 1980.

[2] Hossan M. A., Memon S., & Gregory M. A., "A novel approach for MFCC feature extraction," In Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on IEEE, pp. 1-5. 2010.

[3] Shao Y., Jin Z., Wang D., & Srinivasan S., "An auditory-based feature for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 4625-4628, 2009.

[4] Hermansky H., & Morgan N., "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578-589, 1994.

[5] Furui S., "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254-272, 1981.

[6] Huang X., Acero A., Hon H. W., & Foreword By-Reddy R., *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice hall PTR, 2001.

[7] Pujol P., Macho D., & Nadeu C., "On real-time mean-and-variance normalization of speech recognition features," In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings,* vol. 1, pp. 773-776, 2006.

[8] Techini, Elhem, Sakka, Zied, & Bouhlel MedSalim. "Robust front-end based on MVA processing for Arabic speech recognition, " Engineering & MIS (ICEMIS), pp. 1-5. 2017.

[9] Chen C. P., Bilmes J. A., & Kirchhoff, K., "Low-resource noise-robust feature post-processing on Aurora 2.0," In *Seventh International Conference on Spoken Language Processing*, 2002.

[10] Chen C. P., & Bilmes, J. A., "MVA processing of speech features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257-270, 2009.

[11] Miao, Y., Gowayyed, M., & Metze, F. "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," In Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on, IEEE, pp. 167-174, 2015.

[12] Qian, Y., Bi, M., Tan, T., & Yu, K.. "Very deep convolutional neural networks for noise robust speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol 24, no 12, pp. 2263-2276, 2016.

[13] Ovtcharov, K., Ruwase, O., Kim, J. Y., Fowers, J., Strauss, K., & Chung, E. S. "Accelerating deep convolutional neural networks using specialized hardware. Microsoft Research Whitepaper, Vol 2, no 11, 2015.

[14] Korba M. C. A., Messadeg D., Bourouba H., & Djemili R., "Noise Robust Features Based on MVA Post-processing," In IFIP International Conference on Computer Science and its Applications, vol. 456, pp. 155-166, 2015.

[15] Ephraim Y., & Malah D., "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[16] Kim C., & Stern R. M., "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no 7, pp. 1315-1329, 2016.

[17] Mitra V., Franco H., Graciarena M., & Mandal A., "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 4117-4120, 2012.

[18] Chen J., Paliwal K. K., & Nakamura S. "Cepstrum derived from differentiated power spectrum for robust speech recognition," *Speech Communication*, vol. 41, no. 2, pp. 469-484, 2003.

[19] Glasberg B. R., & Moore B. C., "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103-138, 1990.

[20] M Slaney, Auditory toolbox version 2. [online] https://engineering.purdue.edu/~malcolm/interval/1998-010/

[21] Ellis, D. PLP and RASTA in MATLAB. [online], 2006. http://labrosa.ee.columbia.edu/matlab/rastamat/

[22] PNCC code. Online http://www.cs.cmu.edu/~robust/archive/algorithms/PNCC_IEEETran/PNCC_IEEETran.tar.gz, 2012.

[23] Korba, Mohamed Cherif Amara, et al. "Robust speech recognition using perceptual wavelet denoising and mel-frequency product spectrum cepstral coefficient features," Informatica journal, vol. 32, no. 3, pp. 283-288, 2008.

[24] Hirsch H. G., & Pearce D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[25] NOISEX-92 Database. [online] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html , 1992.

[26] Yung S., Evermann G., Gales M., Hain, T., Kershaw D., Moore G., & Woodland P., The HTK book. *Cambridge University Engineering Department, Cambridge*, 2005.

[27] Dai P., Rudzicz F., Soon Y., Mihailidis A., & Ding H., "2D Psychoacoustic modeling of equivalent masking for automatic speech recognition," Signal Processing, vol. 115, pp. 9-19, 2015.

[28] Seyedin S., Gazor S., & Ahadi S. M., "On the distribution of Mel-filtered log-spectrum of speech in additive noise," Speech Communication, vol. 67, pp. 8-25, 2015.