International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:9, 2014

Robust Features for Impulsive Noisy Speech Recognition Using Relative Spectral Analysis

Hajer Rahali, Zied Hajaiej, Noureddine Ellouze

***Abstract*—**The goal of speech parameterization is to extract the relevant information about what is being spoken from the audio signal. In speech recognition systems Mel-Frequency Cepstral Coefficients (MFCC) and Relative Spectral Mel-Frequency Cepstral Coefficients (RASTA-MFCC) are the two main techniques used. It will be shown in this paper that it presents some modifications to the original MFCC method. In our work the effectiveness of proposed changes to MFCC called Modified Function Cepstral Coefficients (MODFCC) were tested and compared against the original MFCC and RASTA-MFCC features. The prosodic features such as jitter and shimmer are added to baseline spectral features. The above-mentioned techniques were tested with impulsive signals under various noisy conditions within AURORA databases.

***Keywords*—**Auditory filter, impulsive noise, MFCC, prosodic features, RASTA filter.

I. Introduction

SPEECH parameterization is an important step in modern automatic speech recognition systems (ASR). The speech parameterization block is used to extract from the speech waveform the relevant information for discriminating between different speech sounds. The information is presented as a sequence of parameter vectors. In this paper, two acoustic features are found: MFCC and RASTA-MFCC. Generally, both methods are based on three similar processing blocks: firstly, basic short-time Fourier analysis which is the same for both methods, secondly, auditory based filterbank, and, thirdly, cepstral coefficients computation. The RASTA method belonging to the second category was proposed to extract robust speech features for recognition by processing temporal trajectories of frequency band spectrum using a band-pass filter [9]. The principle of RASTA method comes from the human auditory perception which indicates the relative insensitivity of human hearing to slowly and quickly varying auditory stimuli. Thus, the RASTA band-pass filter is designed with an IIR filter with a sharp spectral zero at the zero frequency in the modulation frequency domain. The most interesting point of RASTA method is to emphasize the important part of speech signal by human hearing perception which is definitely more immune to noise.

MFCC are used extensively in ASR. MFCC features are derived from the FFT magnitude spectrum by applying a filterbank which has filters evenly spaced on a warped frequency scale.

Hajer Rahali, Zied Hajaiej, and Noureddine Ellouze are with the National Engineering School of Tunis (ENIT), Laboratory of Systems and Signal Processing (LSTS), BP 37, Le Belvédère, 1002 Tunis, Tunisie (e-mail: hajer.rahali@enit.rnu.tn, zied.hajaiej@enit.rnu.tn, N.ellouze@enit.rnu.tn).

The logarithm of the energy in each filter is calculated and accumulated before a Discrete Cosine Transform (DCT) is applied to produce the MFCC feature vector. There are many similarities between the two methods. The difference however lies in the shape of the filterbank. In this paper we present the proposed modifications of MFCC method, and it will be shown that the performance of MFCC and RASTA-MFCC, are also compared to MODFCC which integrate a new model. In the current paper, prosodic information is first added to a spectral system in order to improve their performance. Such prosodic characteristics include parameters related to the fundamental frequency such as the jitter and shimmer. The MODFCC shows consistent and significant performance gains in various noise types and levels. For this we will develop a system for automatic recognition of isolated words with impulsive noise based on HMM\GMM. We propose a study of the performance of parameterization techniques including the prosodic features proposed in the presence of different impulsive noises. Then, a comparison of the performance of different used features was performed in order to show that it is the most robust in noisy environment. The sounds are added to the word with different signal-to-noise SNRs (20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB). Note that the robustness is shown in terms of correct recognition rate (CRR) accuracy. The evaluation is done on the AURORA database.

This paper is organized as follow; in the next section we describe the proposed modifications of MFCC. An experimental study performed to compare the performance of the different parameterization methods in various acoustic environments is described in Section III. Finally, the major conclusions are summarized in Section IV.

II. New Proposed Technique

In this work, we present some modifications of the original MFCC in its recognition accuracy under noisy environments. This method is based on a study of differences between MFCC and RASTA-MFCC parameterizations.

A. MFCC and Relative Spectral RASTA

The relative spectral analysis technique (RASTA) is based on the idea that the rate of changing of the short-term spectrum for linguistic and non-linguistic components in speech is different [9]. This means that the spectral components of the communication channel vary more quickly or more slowly than the spectral components of the speech and they could be separated (filtered). The core part of RASTA processing is a band-pass filtering of the spectral parameters trajectories by an IIR filter. The convolutive (in the time domain) distortions in the communication channel can be

reduced by using the RASTA filtering in the logarithmic domain (spectral or cepstral). The RASTA approach can be combined with the mel cepstral coefficient method (so called RASTA-MFCC approach) or can directly be applied to the cepstral trajectories [9]. Fig. 1 shows a block diagram for extracting MFCC and RASTA-MFCC. The steps of RASTA-MFCC are as follows. First, we pre-emphasis the input speech signal using a pre-emphasis filter. The Hamming window is applied to the pre-emphasized signal and then, it is processed by short-time Fourier transform (STFT). In the next step, we have used logarithmic amplitude transformation as a compressing static non linearity in step one of RASTA. The RASTA-MFCC is derived using a band-pass filter where more slowly and quickly changing parts for each spectral component are suppressed. The expanding static non linearity in step 3 of RASTA was an antilogarithmic transformation. Finally, the logarithm of the energy in each filter is calculated and accumulated before DCT is applied to achieve the cepstral coefficient.

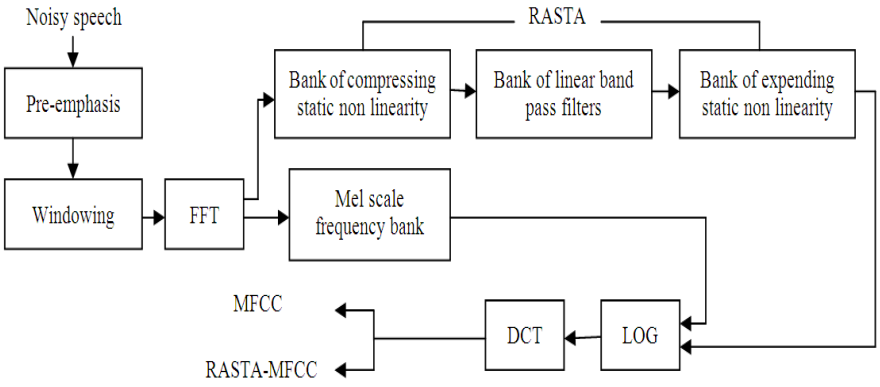

Fig. 1 Block diagram for extracting MFCC and RASTA-MFCC

B. Proposed Modifications of MFCC

In this paper, we present some modifications of the standard MFCC feature extraction method. The proposed modifications are presented in the following section. A schematic diagram of the proposed technique is shown in Fig. 2. In this proposed algorithm MODFCC an application of pre-emphasis is applied to the speech signal before the short term spectral analysis. In the second step, the digitized noisy speech is segmented into overlapping frames, each of length 20 ms with 10 ms overlap [7], in speech processing a Hamming window are mostly used. Next, the FFT is taken of these segments. Afterwards, the segmented signal is filtered using the non linear model of the external and middle ear which is given by the following analytical expression [5]:

$$H(f) = -2.184 * f^{-0.8} + 6.5 * e^{-0.6(f-3.3)^2} - 10^{-3} * f^{-3.6}. \quad (1)$$

The transfer function of the external and middle ear model is shown in Fig. 3. The next processing step applies a filterbanks. Many different types of filterbanks exist but for MODFCC features the gammachirp filterbank is used. In this study, our objective is to introduce new speech features that are more robust in noisy environments. We propose a robust speech feature which is based on the method with gammachirp filterbank. The output signal of the outer and middle ear model filter is applied to a gammatone filterbank. On each sub-band we calculate the sound pressure level Ps (dB) in order to have the corresponding sub-band chirp term C. Those values of chirp term C corresponding to each sub-bands of the gammatone filterbank lead to the corresponding gammachirp filterbank. The proposed auditory use filters that are smoother and broader than the Mel filterbank (the bandwidth of the filter is controlled by the ERB curve and the bandwidth multiplication factor F). The main differences between the proposed filterbank and the typical one used for MFCC estimation are the type of filters used and their corresponding bandwidth. In this paper, we experiment with one parameter to create a family of gammachirp filterbanks: the number of filters. An example of the gammachirp filterbank employing 32 filters is shown in Fig. 4. The free parameter in gammachirp filterbank as noted above is the number of filters. By increasing the number of filters they become narrow but with a small number of filters the loss of information is introduced. A new filterbank is presented where the width of the filters is fixed to 226 filter and the number of them is equal to the number of spectral coefficients (in our case we used 265 filters). The RASTA filter removes variations in the signal that are outside the rate of change of speech by filtering the log-spectrum at each frequency band. Both very slow and very fast changes in sound are ignored by the human ear, so RASTA processing attempts to filter these components out. The filter also helps to eliminate noise due to channel variation in the data. That is why we use the RASTA filtering technique to process the cepstral coefficients, and then we get the features coefficients which we need. In the next stage, we calculate tonal and non tonal components. This step begins with the determination of the local maxima, followed by the extraction of the tonal components (speech) and non tonal components (impulsive noise), in a bandwidth of a critical band. If frequency exceeds neighboring

components within a bark distance by at least 6 dB then it will treated as "speech" otherwise it will be considered as "noise". The selective suppression of tonal and non tonal components of masking is a procedure used to reduce the number of maskers taken into account for the calculation of the global masking threshold. The tonal and non tonal components remaining are those which are above the hearing absolute threshold. Individual masking threshold takes into account the masking threshold for each remaining component [6]. Speech signals contain two types of information, time and frequency. In time space, sharp variations in signal amplitude are generally the most meaningful features. In the frequency domain, although the dominant frequency channels of speech signals are located in the middle frequency region, different speakers may have different responses in all frequency regions [3].

Thus, the traditional methods which just consider fixed frequency channels may lose some useful information in the feature extraction process. The characteristic of multiple frequency channels and any change in the smoothness of the signal can then be detected to perfectly represent the signals. Then, the MFCC are applied to these channels to extract features characteristics. MFCC as previously stated has the advantage that they can represent sound signals in an efficient way because of the frequency warping property. In this way, the advantages of this technique are combined in the proposed method. For the final acoustic modeling we extended the modified MFCC-cepstral representation with derived delta and delta-delta features. The following features were extracted: 12 MFCC, 12 RASTA-MFCC, and 39 MODFCC. The energy of the frame, first (Δ) and second temporal derivatives ($\Delta\Delta$) extracted from each enumerated parameter. At the end, we have 9 distinct feature vectors that can be categorized into three categories according to its length. Firstly, feature vector has length 13: (12 MODFCC and Energy). Secondly, feature vector has length 26: (12 MODFCC, Energy, and 13 Δ). Thirdly, feature vector has length 39: (12 MODFCC, Energy, 13 Δ, and 13 $\Delta\Delta$). We note that the addition of delta-cepstral features to the static 13 dimensional MODFCC features strongly improves speech recognition accuracy, and a further (smaller) improvement is provided by the addition of double delta-cepstral features. The feature vector constructed on the basis of MODFCC is applied in the statistical classifier. This classifier is based on Gaussian mixture model (GMM) and Hidden Markov Model (HMM). Following the above modifications a new acoustic features (named MODFCC are derived. In the next section, we investigate the robustness and compare the performance of the proposed features to that of RASTA-MFCC with the different prosodic parameters by artificially introducing different levels of impulsive noise to the speech signal and then computing their correct recognition rate.

III. Experimental Framework and Results

In this section, we investigate the robustness of MODFCC in noise by artificially injecting various types of impulsive noise to the speech signal. We then present speech recognition experiments in noisy recording conditions. The results are obtained using the AURORA databases.

A. AURORA Task

AURORA is a noisy speech database, designed to evaluate the performance of speech recognition systems in noisy conditions. The AURORA task has been defined by the European Telecommunications Standards Institute (ETSI) as a cellular industry initiative to standardize a robust feature extraction technique for a distributed speech recognition framework. The initial ETSI task uses the TI-DIGITS database down sampled from the original sampling rate of 20 kHz to 8 kHz and normalized to the same amplitude level [4]. Two different noises (Explosion and door slams) have been artificially added to different portions of the database at signal-to-noise (SNR) ratios ranging from clean, 20 dB to -5 dB in decreasing steps of 5dB. The training set consists of 8440 different utterances split equally into 20 subsets of 422 utterances each. Each split has one of the three noises added at one of the seven SNRs (Clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB). The test set consists of 4000 test files divided into four sets of 1000 files each. Each set is corrupted with one of the three noises resulting in a total of (2 x 1000 x 7) 14,000 test utterances. In spite of some drawbacks of the current AURORA task such as the matched test and training conditions [1], or the absence of natural level variations and variable linear distortions, the AURORA task is of interest since it can demonstrate the potential benefits of using noise robust feature extraction techniques towards improving the recognition performance on a task which (though with matched training and test conditions) has substantial variability due to different types of additive noise at several SNRs.

C. Experimental Setup

To evaluate the suggested techniques, we carried out a comparative study with different baseline parameterization techniques of MFCC and RASTA-MFCC implemented in HTK. For the performance evaluation of our feature extractors, we have used the two noise of the AURORA corpus at different SNRs. The features extracted from clean and noisy database have been converted to HTK format using "VoiceBox" toolbox [2] for Matlab. In our experiment, there were 21 HMM models trained using the selected feature MODFCC, MFCC and RASTA-MFCC. Each model had 5 by 5 states left to right. The features corresponding to each state occupation in an HMM are modeled by a mixture of 12 Gaussians [8]. In all the experiments, 39 vectors are used as the baseline feature vector. Jitter and shimmer are added to the baseline feature set both individually and in combination.

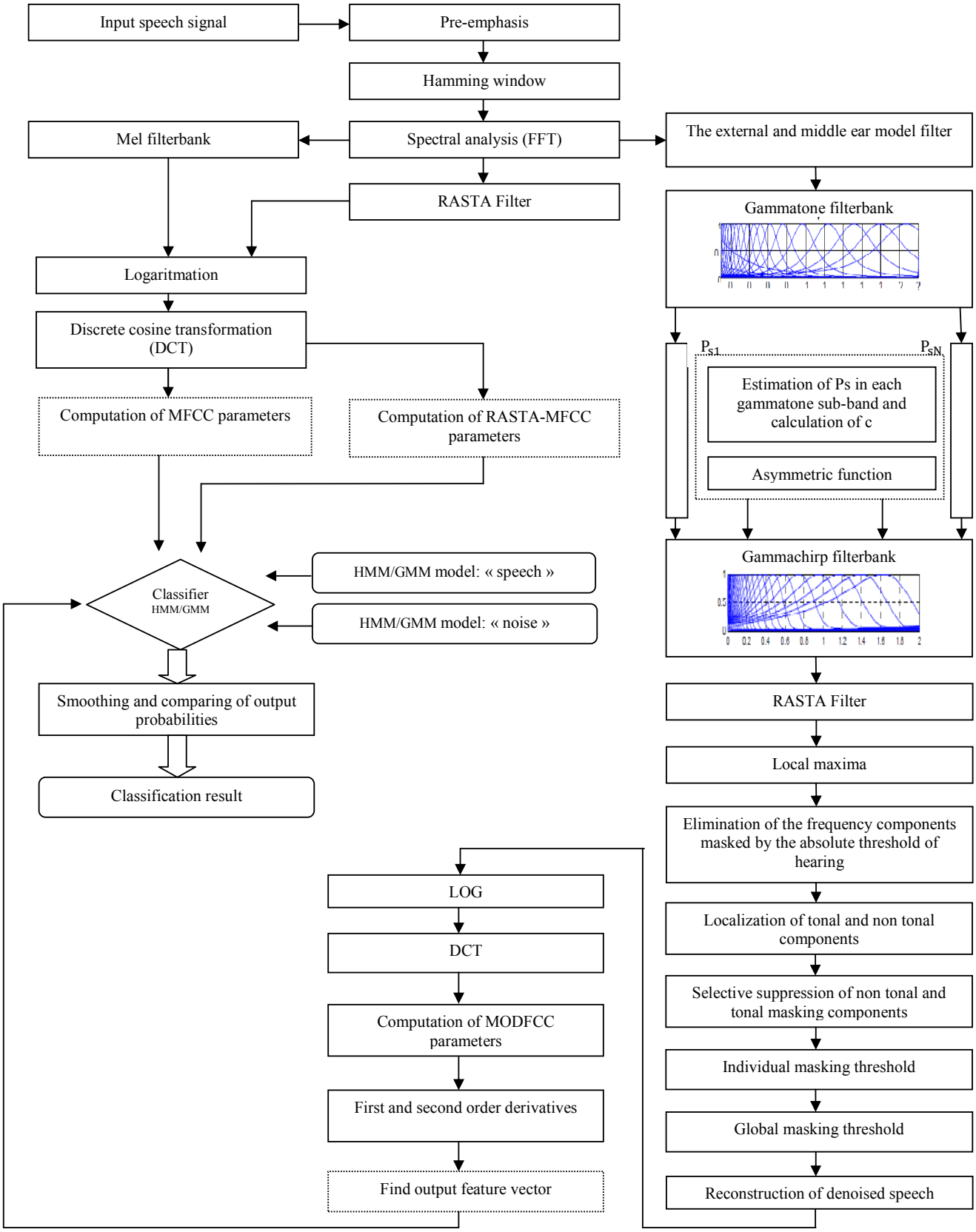

Fig. 2 The structure of MODFCC features extraction

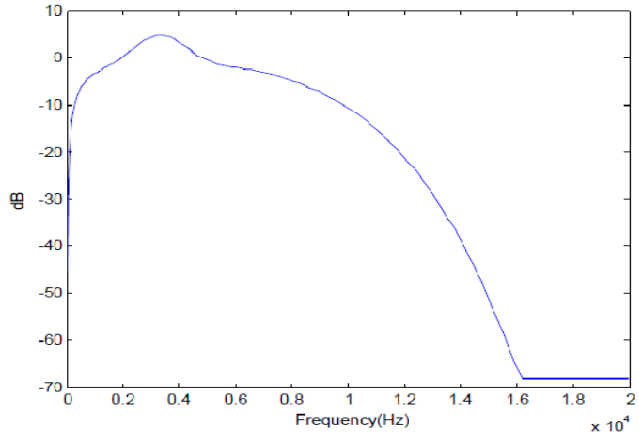

Fig. 3 The transfer functions of the external and middle ear model

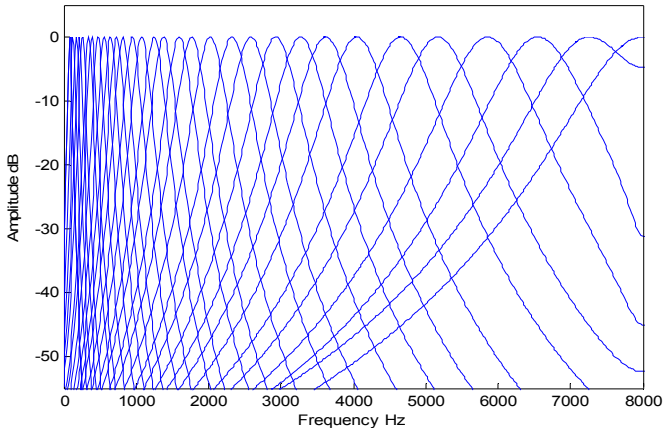

Fig. 4 Characteristics of gammachirp filterbank

D. Results and Discussion

The performance of the suggested parameterization methods is tested on the AURORA databases using HTK. We use the percentage of word accuracy as a performance evaluation measure for comparing the recognition performances of the feature extractors considered in this paper. %: The percentage rate obtained. One Performance measures, the correct recognition rate (CORR) is adopted for comparison. They are defined as:

% CRR = no. of correct labels/no.of total labels * 100%. (2)

Comparison of phoneme recognition rates is shown in the Tables I-IV.

TABLE I
WORD ACCURACY (%) USING DIFFERENT PARAMETERIZATION TECHNIQUES

Features	Explosions							Door slams						
SNR(dB)	∞	20	15	10	5	0	-5	∞	20	15	10	5	0	-5
MFCC (24 filters)	72.46	62.03	60.30	58.43	46.33	33.61	33.32	74.20	60.52	55.98	44.06	40.19	35.32	32.76
RASTA-MFCC (24 filters)	73.04	63.85	61.90	59.07	47.99	34.25	34.04	75.94	61.16	56.56	45.33	41.77	36.94	33.34
MODFCC (24 filters)	74.45	64.25	62.28	60.44	48.38	35.67	35.33	76.35	62.59	57.97	46.71	42.12	37.30	34.71
MODFCC (265 filters)	75.51	65.34	63.32	61.51	49.48	36.75	36.44	77.44	63.66	58.76	47.86	43.22	38.42	35.86

TABLE II
WORD ACCURACY (%) OF RASTA-MFCC USING THE PROSODIC FEATURES

Features	Explosions							Door slams						
SNR(dB)	∞	20	15	10	5	0	-5	∞	20	15	10	5	0	-5
RASTA-MFCC (Baseline)	85.40	82.25	78.29	78.44	66.38	50.65	33.35	84.32	80.56	78.98	77.76	70.12	55.32	42.76
RASTA-MFCC+Jitter	88.03	84.84	80.80	80.03	68.99	53.24	35.93	86.93	83.17	81.59	80.36	72.72	57.90	45.31
RASTA-MFCC+Shimmer	88.45	85.25	81.21	81.46	69.38	53.65	36.34	87.34	83.56	81.91	80.74	73.14	58.32	45.76
RASTA-MFCC+Jitter+Shimmer	89.55	86.35	82.39	82.54	70.45	54.72	37.44	88.44	84.66	82.76	81.86	74.22	59.42	46.86

TABLE III
WORD ACCURACY (%) OF MODFCC USING PROSODIC FEATURES

Features	Explosions							Door slams						
SNR(dB)	∞	20	15	10	5	0	-5	∞	20	15	10	5	0	-5
MODFCC (Baseline)	92.43	90.17	88.20	85.74	77.21	71.54	50.07	90.35	89.96	88.98	87.70	78.99	70.23	68.45
MODFCC+Jitter	95.05	92.78	90.83	88.33	79.84	74.74	52.68	92.94	92.51	91.50	90.31	81.59	72.82	71.03
MODFCC+Shimmer	95.43	93.17	91.25	88.71	80.28	74.55	53.06	93.35	92.94	91.95	91.70	84.59	75.82	74.03
MODFCC+Jitter+Shimmer	96.53	94.27	92.30	89.84	81.31	75.64	54.17	94.45	94.06	93.08	92.83	85.67	76.92	75.13

TABLE IV
RECOGNITION RATE (%) OF MODFCC USING DYNAMIC PROPERTIES

Features	Explosions							Door slams						
SNR(dB)	∞	20	15	10	5	0	-5	∞	20	15	10	5	0	-5
MODFCC (13)	84.77	82.55	80.18	75.74	71.65	65.33	43.87	82.54	80.64	80.58	79.71	65.87	54.02	34.74
MODFCC+Δ (26)	84.53	82	82.65	80.71	79.68	70.45	59.50	90.91	89.53	88.98	85.33	80.45	72.49	66.87
MODFCC+Δ+ ΔΔ (39)	93.92	91.11	91.20	90.09	89.56	85.67	77.87	94.21	92.87	91.98	90.10	81.90	78.10	72.76

The new filterbank with fixed filter width and a large number of filters has also been applied and tested with the standard MFCC method. It can be seen (Table I) that the recognition accuracy improves slightly for a relative value of 1%. From the Table I it can be observed that improvement (3% relative increase of recognition accuracy) is achieved with the new MODFCC method of parameterization over the baseline MFCC method. Tables II and III presents the performance of two voice features in presence of various levels of additive noise. We note that the MODFCC features that are extracted using the gammachirp containing frequency-domain noise and speech detection exhibit the best CRR. Also, it is observable that the performance of the two features decreases when the SNR decreases too, that is, when the speech signal becoming more noisy. Similarly, the performance of RASTA-MFCC shows a decrease, but it is a relatively small decrease, whereas the MODFCC features have the overall highest recognition rate throughout all SNR levels. In additive noise conditions the proposed method provides consistently better word accuracy than all other methods. Jitter and shimmer are added to the baseline feature set both individually and in combination. The absolute accuracy increase is 2.6% and 3.0% after appending jitter and shimmer individually, while there is 4.1% increase when used together. As we can see in the tables, the identification rate increases with speech quality, for higher SNR we have higher identification rate, the MODFCC based parameters are slightly more efficiencies than standard RASTA-MFCC for noisy speech (94.27% vs 86.35% for 20 dB of SNR with jitter and shimmer) but the results change the noise of another. From the above Table IV, it can be seen that the recognition rates are above 90%, this is recognition rates are due to the consideration of using 39 MODFCC features.

IV. CONCLUSION

In this paper we proposed a novel robust method-based feature extraction algorithm for speech recognition. The proposed features called MODFCC have been shown to be more robust than MFCC and RASTA-MFCC in noise environments for different SNRs values. Jitter and shimmer features have been evaluated as important features for analysis for speech recognition. Adding jitter and shimmer to baseline spectral and energy features in an HMM\GMM based classification model resulted in increased word accuracy across all experimental conditions. The results gotten after application of this features show that this method give acceptable and better results by comparison at those gotten by other methods of parameterization.

REFERENCES

[1] H. G. Hirsch, D. Pearce, "The AURORA Experiment Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition", *ISCA ITRW ASR2000 Automatic Speech Recognition*: Challenges for the Next Millennium, France, 2000.

[2] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB", Software, available (Mar, 2011).

[3] C. Hsieh, E. Lai, and Y. Wang, "Robust speaker identification system based on wavelet transform and Gaussian mixture model", *Journal of Information Science and Engineering*, 19, pp. 267-282, 2003.

[4] Schlüter, R., Bezrukov, I., Wagner, H., Ney, H, "Gamma tone features and feature combination for large vocabulary speech recognition", In *ICASSP 2007. Honolulu* (HI, USA), April 2007, p. 649-652.

[5] Irino. T, E. Okamoto, R. Nisimura, Hideki Kawahara and Roy D. Patterson, "A Gammachirp Auditory Filterbank for Reliable Estimation of Vocal Tract Length from both Voiced and Whispered Speech", The *4th Annual Conference of the British Society of Audiology*, Keele, UK, 4-6, Sept, 2013.

[6] T. Irino and M. Unoki, "An Analysis Auditory Filterbank Based on an IIR Implementation of the Gammachirp", *J. Acoust. Soc Japan*. 20(6): 397-406, November, 1999.

[7] Daniel PW Ellis and Byunk Suk Lee, "Noise robust pitch tracking by subband autocorrelation classification", in *13th Annual Conference of the International Speech Communication Association*, 2012.

[8] D. Povey, L. Burget, et al., "The Subspace Gaussian Mixture Model–A Structured Model for Speech Recognition", *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, April 2011.

[9] H. Hermansky and N. Morgan, "RASTA Process ing of Speech", *IEEE. Trans. on Speech and Audio Processing*, Vol.2, No.4, Oct. 1994.