

Improving Continuous Arabic Speech Recognition over Mobile Networks DSR and NSR Using MFCCs Features Transformed

L. Bouchakour , M. Debyeche

Abstract—In this paper, we argue that the improved the performance of speech recognition in mobiles communication system, in order to improve the performance of Automatic Speech Recognition (ASR) systems, we have achieved by two modules Front-End or feature extractor used and a Back-End or recognizer. The Front-End we have used MFCC-MT (Multitaper Frequency Cepstral Coefficients features) and Gabor features GF-MFCC, are the result of their ability to extract discriminative internal representations that are robust to the many sources of variability in speech signals so to reduce spectral variations and correlations. In the back-end we have investigated different systems of classification in the field of speech using three systems: CHMM (Continues Hidden Markov Models), DNN (Deep Neural Network) and HMM-DNN hybrid. We have examined the DNN, which is usually used to reduce spectral variations and the spectral correlations that exist in model signals. Furthermore, we focused particularly on HMM-DNN in continuous speech recognition tasks of the large Arabic vocabulary, and we gave more emphasis to the optimal number of the hidden units and the best feature of input for DNN as well. Our findings show that HMM-DNN can achieve consistently almost 8% of clean speech, 13% of AMR-NB coder and 8.5% of DSR coders. The system was trained using the 3hour training set 440 sentences with 20 speakers with labels generated by Viterbi alignment from a maximum likelihood ML trained CHMM system using the HTK toolkits.

Keywords— ASR, MFCC-Multitaper , Gabor features, DNN.

I. INTRODUCTION

IN these last decades, the research in the field of automatic speech recognition over mobile communication networks has led to the development of client-server recognition systems, also known as Network Speech Recognition (NSR) (Kim et al. 2001; Peláez-Moreno et al. 2001) and Distributed Speech Recognition (DSR) (Pearce 2000). This research

brought numerous methods to improve the recognition performance by increasing the robustness against variability of speech signals (speech coding, DSR, AMR-NB Adaptive Multi-Rate Narrow-Band). On the performance viewpoint, the mobile technologies provide access to communication networks anytime, anywhere, and from any device. However, there are several sources degradation regarding the performance of speech recognition. Usually be caused by the resulting distortions of low-bit-rate in speech coders of networks and the arising distortions from the transmission errors that occur through the associated communication channels. The ASR system may be divided into two modules: a front-end or feature extractor and a back-end or recognizer. The objective of ASR is to recognize human speech such as sentences or words and phonemes which transforms a speech signal into a compact representation. A feature extractor (front-end) methods have been developed for robust ASR. The use of feature extraction techniques inspired by the auditory system has previously demonstrated a boost in speech recognition performance. The Mel Frequency Cepstral Coefficients features (MFCC) are the most popular and they demonstrate good performance in ASR. MFCC computation begins by multiplying a short-term frame of speech by a tapered window function. A major source of problems are the spectral front-ends based on either discrete Fourier transform (DFT). Despite having low bias, a consequence of the windowing is increased estimator variance. In this study, the focus lies on the improvement of feature extraction that employ a better MFCC estimator based in technique for reducing the spectral variance we proposed two steps. These two steps allow the reduction of spectral amplitude variation of the obtained spectrum that are, the multitaper, that to replace a windowed periodogram estimate with a multiple windowed spectrum estimate and second step is the Gabor features. In the multitaper spectral estimation method, a set of orthogonal tapers is applied to the short-time speech signal and the resulting spectral estimates are averaged, which reduces the spectral variance. The Gabor features were first used by Kleinschmidt et al , in 2002 for ASR improved feature extraction by using a set of physiologically inspired filters (Gabor filters), which is applied to a spectro-temporal representation of the speech signal. With considerable ASR improvements. Second module that, back-end or recognizer,

L. Bouchakour University of Sciences and Technology Houari Boumediene (USTHB) Faculty of Electronics and Computer Sciences (FEI), Department of Telecommunication, Speech Communication and Signal Processing Laboratory (LCPTS), P.O. Box 32, Bab Ezzouar, Algiers, Algeria lbouchakour@usthb.dz

M. Debyeche National Polytechnic School (NPS) and University of Sciences and Technology Houari Boumediene (USTHB) Faculty of Electronics and Computer Sciences (FEI), Department of Telecommunication, Speech Communication and Signal Processing Laboratory (LCPTS), P.O. Box 32, Bab Ezzouar, Algiers, Algeria mdebyeche@gmail.com

the Hidden Markov models (HMMs) are widely used in many systems, a statistical framework that supports both acoustic and temporal modeling. However, the HMMs make a number of suboptimal modeling approximations that limit their potential effectiveness. The Neural networks avoid many of these approximations, can learn complex functions, generalize effectively, tolerate noise, and support parallelism. Recent studies have shown that deep neural networks (DNNs) perform significantly better than shallow networks and HMMs on continuous speech recognition. In this paper, we are conducting research on the speech recognition over mobile network DSR and NSR based on two modules a front-end and a back-end. The front-end based on MFCC-Multitaper and features Gabor and the back-end based on Hidden Markov Model (HMM), deep neural networks (DNN) and HMM-DNN. The remainder of the paper is organized as follows. In Section 2, a detailed description of the Speech codec DSR and AMR-NB. Front-End (features extraction) in Section 3 we then propose two techniques for reducing the spectral variance. back-end (recognition system) in section4, we present three techniques HMM, DNN and HMM/DNN. The experiments are presented in section 5. Finally, we summarize the conclusion of the presented work in section 6.

II. SPEECH CODEC

A. AMR-NB Codec

Adaptive Multi-Rate Narrow-Band (AMR-NB) speech coding research has progressed substantially in recent years and several algorithms are rapidly finding their way into consumer products. Research and development in Algebraic Code-Excited Linear Prediction (ACELP) yielded algorithms that have been adopted for several standards and employed in many network and cellular telephone applications [4]-[13]. The AMR speech coder consists of the multi-rate speech coder, a source controlled rate scheme including a voice activity detector and a comfort noise generation system, and an error concealment mechanism to combat the effects of transmission errors and lost packets. The coder is capable of operating at 8 different bit-rates denoted coder modes. The multi-rate speech coder is a single integrated speech codec with eight source rates from 4.75 kbit/s to 12.2 kbit/s [13]-[14], and a low rate background noise encoding mode. The 12.2 kbit/s mode is equivalent to the GSM EFR (Global System for Mobile Enhanced Full Rate) coder while the 7.4kbit/s mode is equivalent to the EFR coder for the IS-136 system. The frame size is 20 ms with 4subframes of 5 ms. A look ahead of 5 ms is used. In the encoder, the speech signal is analyzed and the parameters of the ACELP speech synthesis model are extracted. The set of linear prediction filter coefficients are calculated for each frame. The indices for the adaptive and Fixed CodeBooks as well as their gains are extracted for each subframe. The speech synthesis is computed by filtering the excitation signal through synthesis filter. The function of the decoder consists of decoding the transmitted parameters (LSF parameters, adaptive codebook vector,

adaptive codebook gain, fixed codebook vector, fixed code book gain) and performing synthesis to obtain the reconstructed speech. The Figure 1 shows a series of processing blocks applicable to ASR over mobile networks.

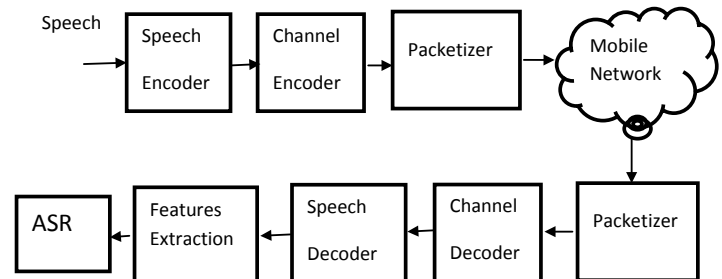


Fig. 1 Block diagram of the ASR over mobile networks

B. Distribution Speech Recognition DSR

The performance of speech recognition systems receiving speech that has been transmitted over mobile channels can be significantly degraded when compared to using a speech clean. The degradations are as a result of both the low bit rate speech coding and channel transmission errors. A Distributed Speech Recognition (DSR) [14]-[15] system overcomes these problems by eliminating the speech channel and instead using an error protected data channel to send a parameterized representation of the speech, which is suitable for recognition. The processing is distributed between the terminal and the network. The terminal performs the feature parameter extraction, or the front-end of the speech recognition system. The pre-emphasis and windowing the short term spectrum is obtained by an Fast Fourier Transform (FFT). This linear spectrum is then warped into a non-linear spectral distribution of 24 bins using triangular weighting filters on a Mel-scale. The 12 cepstral coefficients are obtained by retaining the 12 lowest frequency coefficients after taking the cosine transform of the logarithm of the 24 Mel-spectrum bins. The chosen frame rate is 10 ms. The total energy of each frame is also computed before the preemphasis is filter. The final output feature vector consists of 12 cepstral coefficients (C1-C12), log Energy and C0. The final feature vector consists of 14 coefficients: the log-energy coefficient and the 13 cepstral coefficients. The C0 coefficient is often redundant when the log-energy coefficient is used. These features (14 coefficients) are transmitted over a data channel to a remote "back-end" recognizer [14,15]. The end result is that the degradation in performance due to transcoding on the voice channel is removed and channel invariability is achieved. The feature compression method selected uses split vector quantisation (SVQ). The 14 coefficients are split into 7subvectors each consisting of a pair of cepstral coefficients.

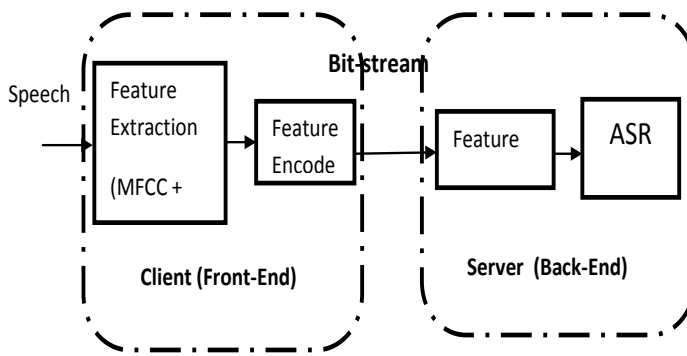


Fig. 2 Block diagram of the front-end algorithm DSR

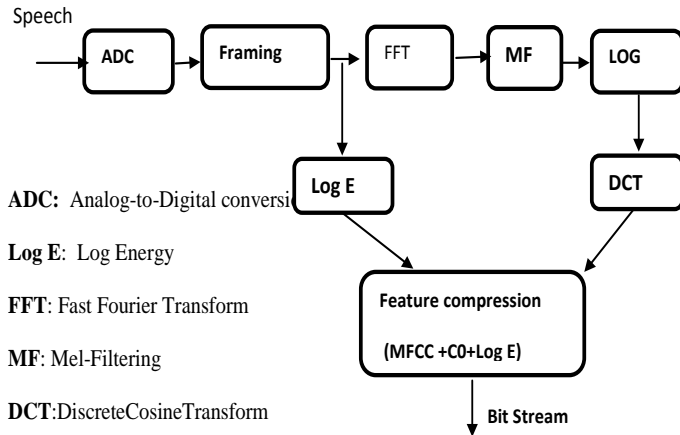


Fig. 3 Block diagram of the front-end algorithm DSR

III. FEATURES EXTRACTION (FRONT-END)

The term “front-end analysis” refers to the first stage of ASR [15], whereby the input acoustic signal is converted to a sequence of acoustic feature vectors. The short-term spectrum provides a convenient way of capturing the acoustic consequences of phonetic events. Ideally the method of front-end analysis should preserve all the perceptually important information for making phonetic distinctions, while not being sensitive to acoustic variations that are irrelevant phonetically. As a general policy for ASR, it seems desirable not to use features of the acoustic signal that are not used by human listeners, even if they are reliably present in human productions, because they may be distorted by the acoustic environment or electrical transmission path without causing the perceived speech quality to be impaired.

A. MFCC standard

The MFCC coefficient is a representation of the short term power spectrum of a sound (Davis and Mermelstein 1980). The frequency bands in MFCC are equally spaced on the mel scale which closely approximates the human auditory system response [15]-[3]. The Mel scale can be calculated by Eq.(1).

$$Mel(f) = 2595 \times \log_{10} \left(1 + \frac{f}{700} \right). \tag{1}$$

The MFCCs are commonly computed from FFT power coefficients filtered by a triangular band pass filter bank, where A_j the output of the j -th filter bank and N is the number of samples in a basic unit.

$$C_n = \sqrt{\frac{2}{N}} \sum_{j=1}^N A_j \cos\left(\frac{n\pi j}{2N}\right), \quad n = 1, \dots, P. \tag{2}$$

MFCC-Multitaper

Hamming windowed DFT spectrum is the used for power spectrum estimation. For m -th frame and l -th frequency an MFCC of the windowed periodogram can be expressed as:

$$S(m, l) = \left| \sum_{j=0}^{N-l} w(j) s(m, j) e^{\frac{2\pi i j l}{N}} \right|^2 \tag{3}$$

Where $l \{0, 1, \dots, l-1\}$ denotes the frequency index, N is the frame length, $s(m, j)$ is the time domain speech signal and $w(j)$ denotes the time domain window function called Taper. Windowing reduces bias difference between estimated spectrum and actual spectrum but it does not reduce variance of the estimated spectrum therefore variance of MFCC. To reduce variance of estimated, replace the windowed DFT spectrum estimation by Multitaper spectrum estimate [16]-[25]. The Multi-taper spectrum estimator is given by:

$$w_p(j) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi p(j+1)}{N+1}\right), \quad j = 0, 1, 2, \dots, N-1 \tag{4}$$

$$\sum_j w_p(j) w_q(j) = \delta_{pq} = \begin{cases} 1, & \text{for } p = q \\ 0, & \text{for } p \neq q \end{cases} \tag{5}$$

$$S(f) = \sum_{j=1}^K \lambda(j) \left| \sum_{l=1}^{N-l} w_j(t) x(t) e^{\frac{-i 2 \pi f t}{N}} \right|^2 \tag{6}$$

That, k multitapers $w_j = [w_j(0) \dots w_j(N-1)]$, $j = 1, \dots, k$, are used with corresponding weights $\lambda(j)$. The multitaper estimate is therefore obtained as a weighted average of k individual sub-spectra. A number of different tapers have been proposed in literature for spectrum estimation.

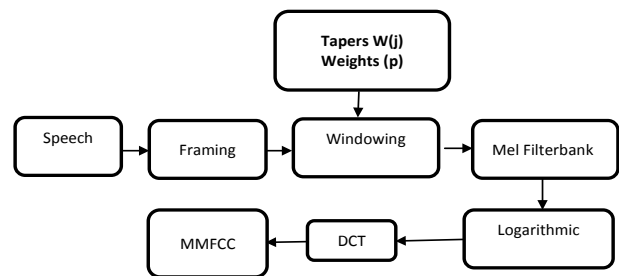


Fig. 4 block diagram for the multi-taper spectrum estimation-based MFCC feature extraction

B. Gabor features

In this section, we describe the characteristics of the Gabor features. Feature extraction was proved the successful application of Gabor features to noise-robust ASR [26]-[29]. The features were first calculated by convolving the log mel spectrogram of speech with a set of 2D Gabor filters. Each Gabor filter $g(n, k)$ is a product of a complex sinusoid $s(n, k)$ with a Hann envelope function $h(n, k)$.

$$s(n, k) = e^{[iw_n(n-n_0) + iw_k(k-k_0)]} \quad (7)$$

$$h(n, k) = 0.5 - 0.5 \cos\left(\frac{2\pi(n-n_0)}{w_n+1}\right) \cos\left(\frac{2\pi(k-k_0)}{w_k+1}\right) \quad (8)$$

The ωn and ωk terms represent the time and frequency modulation frequencies of the complex sinusoid, while W_n and W_k represent time and frequency window lengths of the Hann window.

C. Dimensionality Reduction Techniques LDA

The feature vectors collected from the speech signal for continuous speech classification. If the technique such as DNN is used, the computational complexity and the memory requirement have been increased. Hence, the vectors are mapped from the feature dimensional space to the lower-dimensional space. This is known as dimensionality reduction technique. The most commonly used dimensionality reduction technique is the Linear Discriminant Analysis (LDA) [35]. The LDA, the variances of the individual elements of the projected centroid vectors are maximized. Simultaneously, the variances of the individual elements of the projected vectors corresponding to the identical clusters are minimized. Hence, LDA helps in bringing down the vectors closer to each other and simultaneously separating the vectors farther from each other in the projected lower-dimensional space. The LDA consists of two positive definite scatter matrices, namely between class scatter matrix (SB) and within-class scatter matrix (SW) defined as follows[34]. An LDA is applied in order to obtain a better discrimination of clusters in feature space, e.g. phonemes. It is a linear transform realized by a matrix multiplication. The result is a compact representation of each cluster with an improved spatial discrimination with respect to other clusters. Furthermore, a dimension reduction can be achieved.

IV. SPEECH RECOGNITION SYSTEM

The speech recognition is basically a pattern recognition problem, although most state-of-the-art approaches to speech recognition are based in the use of HMMs and GMMs, also called Continuous Density HMMs (CD-HMMs)[30], these models are all based on probability estimates and maximization of the sequence likelihood. While the neural network based in Maximum A Posteriori criterion,

A. Hidden Markov Model (HMM)

The HMM has been the dominant technique for ASR for at least two decades. One of the critical parameters of HMM is the state observation probability distribution. The Gaussian mixture HMMs are typically trained based on maximum likelihood criterion [30]-[33].

The input speech from a microphone is converted into a sequence of fixed size acoustic vectors Y in a process called feature extraction (MFCC). The decoder then attempts to find the sequence of words W , which is most likely to have generated Y , the decoder, tries to find

$$\hat{W} = \underset{W}{\text{argmax}} \{P(W/Y)\} \quad (9)$$

The Y model can be characterized by the transitions A_{ij} and emitting matrix probabilities $B_j(X_i)$.

B. Multi-Variate Continuous Distributions

We have just defined a Continuous Density HMM (CDHMM), many natural processes involve variable quantities which approximate reasonably well to the normal (or Gaussian) distribution. The normal distribution has only two independently specifiable parameters, the mean, μ_k , and Σ_k the covariance matrix [30]. The definition of the multi-variate normal distribution gives the output probability compactly in matrix notation:

$$b_j(o_t) = p\left(\frac{x}{s_j}\right) = \sum_{k=1}^K g_k N(o, \mu_k, \Sigma_k) \quad (10)$$

$$N(o, \mu_k, \Sigma_k) = \frac{1}{2\pi^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(o-\mu_k)^t |\Sigma_k|^{-1} (o-\mu_k)\right) \quad (11)$$

where :

μ_k and Σ_k are means vector and the covariance matrix respectively:

$|\Sigma_k|$ is the determinant of Σ_k .

$(o-\mu_k)^t$ is the transpose of $(o-\mu_k)$.

C. Deep Neural Networks

The Deep Neural Networks (DNNs)[11], [38]-[42] is simply a multi-layer perceptron (MLP) with many hidden layers between its inputs and outputs. In this section, we review fundamental ideas of the DNN, that can be used as an acoustic model for speech recognition. DNNs have achieved tremendous success continuous speech recognition. The Pre-training DNNs performing back propagation training from a randomly initialized network can result in a poor local optimum, especially as the number of layers increases. To remedy this, pre-training methods have been proposed to better initialize the parameters prior to back propagation. The most well-known method of pre-training grows the network layer by layer in an unsupervised manner. This is done by treating each pair of layers in the network as a restricted

Boltzmann machine (RBM) that can be trained using an objective criterion called contrastive divergence.

The DNN are typically trained based on posterior probability criterion Eq (6) of a class S given an observation vector X , as a stack of $(L + 1)$ layers of log-linear models. Each hidden activation h_i is computed by multiplying the entire input V by weights W in that layer.

$$\begin{aligned} p_{y/x}(y=S/x) &= P^L(y=S/x) \\ &= \frac{e^{z^L(v^L)}}{\sum_j e^{z_j^L(v^L)}} \\ &= \text{softmax}_S(v^L) \end{aligned} \quad (12)$$

Where $z(v) = (w)^T v + a$, and W and a represent the weight matrix and bias vector in the l -th layer, respectively.

D. Hybrid DNN/HMM

The DNN-HMM [43] is a special case of the artificial neural network (ANN) HMM hybrid system developed in 1990s, for which several adaptation techniques have been developed. The DNN accepts an input observation, which typically consists of 6 frames of acoustic features, and process it through many layers of nonlinear transformation [43].

$$h_i^l = \sigma\left(z_i^l(v^l)\right) = \sigma\left(\left(w_i^l\right)^T v^l + a_i^l\right) \quad (13)$$

Where w^l and a^l are the weight matrix and bias, respectively, at hidden layer l , h_i^l is the output of the i -th neuron

$$z^l(v^l) = (w^l)^T v^l + a^l \quad (14)$$

is the excitation vector given input v^l , and $v^0 = x$ and $\sigma(x) = \frac{1}{1 + \exp(-x)}$ is the sigmoid function

applied element-wise. At the top layer, the softmax function is used to estimate the state posterior probability $p(y=s/x)$ which is converted to the HMM state emission probability as:

$$p(x/y=s) = \frac{p(y=s/x)}{p(y=s)} p(x) \quad (15)$$

V. EXPERIMENTS

A. Corpus analyses

Our work consists of a sequence of two phases shown in Fig. They consist of: Corpus Acquisition and Phoneme Segmentation.

1) 1-Corpus acquisition a 3-hour recording of standard Arabic speech [44]. It contains a list of Arabic sentences, an Arabic dictionary and script files used for manipulating

corpus information.

- 2) Phoneme Segmentation, the corpus is processed by the HTK (Hidden Markov Model) speech recognition engine to produce the phoneme and word segmentation output files for all utterances.
- 3) Arabic phoneme, the Arabic phoneme set used in the corpus is shown in Table I. Every phoneme is corresponding to an English symbol. The regular Arabic short vowels /AE/, /IH/, and /UH/ correspond to the Arabic pronunciation Fatha, Damma, and Kasra respectively.

Table. I The Arabic phoneme list

Phoneme	Arabic Letter	Phoneme	Arabic Letter
/AE/	◌َ FATHA	/S/	س
/UH/	◌ُ DAMMA	/SH/	ش
/IH/	◌ِ KASRA	/SS/	ص
/Sh/	◌ْ SKON	/DD/	ض
/E/	ء	/AI/	ع
/B/	ب	/GH/	غ
/T/	ت	/F/	ف
/TH/	ث	/Q/	ق
/G/	ج	/K/	ك
/HH/	ح	/L/	ل
/KH/	خ	/M/	م
/D/	د	/N/	ن
/DH/	ذ	/H/	ه
/R/	ر	/W/	و
/Z/	ز	/Y/	ي

B. Analysis of results

To test the speech recognition performance of the DNN and HMM-DNN, we conducted series of experiments on Arabic database. These experiments are performed with the 8 kHz multi-condition, speech clean, and speech transcoded AMR and DSR. The training set consists of 440 utterances (about three hours for continuous speech and one hour for testing).

The input layer was formed by a context window of 6 frames which create it for the MFCC features and 1000*2000*1500*2500 hidden units for each layer and the final soft-max output layer had 40 units (phonemes). The networks have been initialized with layer-by-layer generative pre-training, and then they are discriminatively trained using

600 iterations of back propagation. A learning rate of 0.15 used for the first 5000 epochs and 0.004 for the remaining 4000 epochs.

These experiments were conducted using CHMM of $N=3$ states and 6 numbers of Gaussian components. The frame length is 240 samples. They aim to study the influence of the AMR-NB and DSR speech coder on the performance of the speech recognition system.

C. Recognition Accuracy (RA)

The Recognition Accuracy (RA) gives the recognition results.

$$RA\% = \frac{N - D - S}{N} \times 100 \quad (16)$$

The Error Rate is given by the following equation:

$$ER = 100\% - RA\% \quad (17)$$

N: is the total number of units (words).

D: is the number of deleted errors.

S: is the number of substituted errors.

Table. II Recognition Accuracy (RA) of Arabic Continuous Speech with Clean and Transcoded AMR/DSR trained DNN

codec	Speech clean	Speech transcoded AMR-NB 5.9 Kbits/s	Speech transcoded DSR 4.8 Kbits/s
Parametre			
MFCC	90	82	85
gabor	91	84	88
Gabor + MFCC	94	89	90
Multin taper	96	91	

Table. III Recognition Accuracy (RA) of Arabic Continuous Speech with Clean and Transcoded AMR/DSR trained DNN/HMM

Speech condition	Speech clean	Speech transcoded AMR-NB 5.9 Kbits/s	Speech transcoded DSR 4.8 Kbits/s
Parametre			
MFCC	94	88	89
gabor	93	90	91
Gabor + MFCC	95	92	93.5
Multin taper	98	95	

Several experiments have been conducted to test the performance of the proposed work. As shown in Table II, the overall classification accuracy using the transformed MFCCs that are GF-MFCC, MFCC-MT and MFCC+GF-MFCC is

91% and 96% and 94 for speech clean, 84%,91%,89 % for speech transcoded AMR-NB, respectively and for speech transcoded DSR 88% GF-MFCC and 90% MFCC+GF-MFCC by DNN. On the other hand, the speech accuracy classification accuracies using the traditional MFCC are calculated as 90% speech clean, 82% speech transcoded AMR-NB and 85% speech transcoded DSR by DNN classifier. The statistical analysis of the MFCCs features is studied [44]-[47] in age and gender problems. These results proved that the speech transcoded AMR/DSR decreases the ASR performance compared to the speech clean. They also showed that the DSR database achieved the batter rate % compared to AMR database rate %. The result of speech transcoded AMR can be explained by the degradation of signal quality which is caused by the effect of excitation codebooks quantification fixed and adaptive, and the quantized spectral parameters LSF quantized. It is clear that the DSR transcoded when we used the 14 coefficients, 12 MFCC (C1-C12), log Energy and C0, and the rate % is increased compared to the AMR transcoded. The transformed MFCCs GF-MFCC, MFCC-MT and GF-MFCC+MFCC that are generated in this work increased the overall classification accuracy about 6% for speech clean, 11% for speech transcoded AMR-NB and 5% for speech transcoded DSR.

Table III shows the results of DNN-HMM classifier with different speech conditions (clean and transcoded DSR/AMR) using MFCCs and its transformation GF-MFCC and MFCC-MT.

These results proved that the speech transcoded AMR/DSR decreases the ASR performance compared to the speech clean. Based on the obtained results which are displayed in the above tables, we can conclude that the DNN-HMM systems with MG-MFCC and MFCC-MT gives the better RA % for communication client-server in mobile networks.

VI. CONCLUSION

The present paper aims at improving the communication client-server on mobile networks NSR and DSR and minimizing the impact of degraded performance ASR which is introduced by speech coder AMR-NB/DSR. For this purpose, the major contributions are made in the area of Front-End and Back-End. Starting with the Front-end which is a new approach introduced to generate transformed MFCCs feature set. Second, the classifier based on DNN and DNN/HMM. As one of the most popular feature sets in the speech signal processing, MFCCs are proved ineffective in speech transcoded in literature. The DNN that performs speech recognition in a manner that is more robust to mismatches in the speech recognition. The obtained results suggest that the MFCCs transformed GF-MFCC and MFCC-MT parameters with DNN-HMM classifiers could improve the ASR performance in mobile communication.

ACKNOWLEDGMENT

The authors would like to thank the LCPTS laboratory team

for providing contribution and making many suggestions, which have been exceptionally helpful in carrying out this research work.

REFERENCES

- [1] Deller, J., Hansen, J., and Proakis, J. *Discrete-Time Processing of Speech Signals*, 2nd Edition. Wiley-IEEE Press (1999).
- [2] Davis, S.B. and Mermelstein, P. Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences. *IEEE Transactions on ASSP*, vol. ASSP-28, no. 4, pp. 357–366 (1980).
- [3] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*. ETSI ES 201 108.
- [4] Zheng-Hua Tan and Børge Lindberg “Automatic speech recognition on mobile devices and over communication networks,” Springer.vol. pp 41-117.2008.
- [5] Kim, H. K. and Cox, R. V. A bitstream-based front-end for wireless speech recognition on IS-136 communications system. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 558–568. (2001).
- [6] Euler, S., and Zinke, J. The influence of speech coding algorithms on automatic speech recognition. In *Proceedings of ICASSP, Adelaide, Australia*. (1994).
- [7] Lilly, B.T., and Paliwal, K.K. Effect of speech coders on speech recognition performance. In *Proceedings of ICSLP*, pp. 2344–2347, Philadelphia, PA, USA. (1996).
- [8] Peinado, A., and Segura, J.C. *Speech Recognition Over Digital Channels*. Wiley. (2006).
- [9] Pearce, D. Enabling Speech and Multimodal Services on Mobile Devices: The ETSI Aurora DSR standards and 3GPP Speech Enabled Services. *VoiceXML Review*, Nov/Dec 2004.
- [10] Pearce, D. Robustness to Transmission Channel—the DSR Approach. *COST278 & ICSA Research Workshop on Robustness Issues in Conversational Interaction*, Aug 2004.
- [11] Tan, Z.-H., Dalsgaard, P., and Lindberg, B. Automatic speech recognition over errorprone wireless networks. *Speech Communication*, vol. 47, no. 1–2, pp. 220–242. (2005).
- [12] Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio, Speech Lang. Process.* 20(1), 30–42 (2012)
- [13] *Universal Mobile Telecommunications System (UMTS); AMR speech Codec*. ETSI TS 126 090.
- [14] Vladimir Fabregas Surigué de Alencar and Abraham Alcaim “On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese,” *IEEE*. pp 693-697. 2009.
- [15] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms* ETSI ES 201 108.
- [16] Jose Trangol and Abel Herrera “Traditional Method and Multi-Taper to Feature Extraction Using Mel Frequency Cepstral Coefficients”*International Journal of Information and Electronics Engineering*, Vol. 5, No. 1, January 2015.
- [17] Md Jahangir Alam, Tomi Kinnunen “Multitaper MFCC and PLP features for speaker verification using i-vectors » *Speech Communication* vol 55. Pp 237–251. 2013.
- [18] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, P. Borgnat: “Multitaper estimation of frequency-warped cepstra with application to speaker verification”. *IEEE Signal Processing Letters*, vol. 17(4), pp. 343–346 (2010).
- [19] K. S. Riedel and A. Sidorenko: Minimum bias multiple taper spectral estimation. *IEEE Trans. on Signal Proc.*, vol. 43(1), pp. 188–195 (1995).
- [20] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P.Flandrin, and P. Borgnat, Multitaper estimation of frequency warped cepstra with application to speaker verification, *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 343346, 2010.
- [21] T. Kinnunen, R. Saeidi, F. Sedlk, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and Haizhou Li, Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification, *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, VOL. 20, NO. 7, SEPTEMBER 2012.
- [22] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten, “What else is new than the Hamming window? Robust MFCC for Speaker Recognition via Multitapering,” 2010.
- [23] Md. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet “Shaughnessy, “Multitaper MFCC feature for speaker verification using I-vectors,” *Automatic Speech Recognition and Understanding (ASRU) IEEE Workshop*, pp. 11-15, 2011.
- [24] T. Kinnunen et al., “Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990, 2001, Sept. 2012
- [25] M. J. Alam, P. Kenny, and D. O’Shaughnessy, “A study of low-variance multi-taper features for Distributed speech recognition, springer,” *Lecture Notes in Computer Science*, vol. 7015, pp. 239-245, 2011.
- [26] Bernd T. Meyer *, Birger Kollmeier “Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition” *Speech Communication* vol 53. Elsevier. Pp 753–767. 2011
- [27] A.M. Castro Martinez et al “ On the relevance of auditory-based Gabor features for deep learning in robust speech recognition” *Computer Speech & Language* vol 45. Pp 21-38. 2017.
- [28] Kleinschmidt, M., Gelbart, D. “Improving word accuracy with Gabor feature extraction”. In: *Proc. ICSLP*. 2002.
- [29] Meyer, B., Kollmeier, B.,” Optimization and evaluation of Gabor feature sets for ASR” . *Proc. Interspeech*.. 2008.
- [30] Mark Gales and Steve Young ‘the application of hidden markov models in speech recognition’ foundation and trends in signal processing, vol 1 No 3, 2008.
- [31] Gernot A. Fink “Markov models for pattern recognition,” *Springer*.vol. pp. 61-92. 2008.
- [32] Sadaoki Furui “Digital speech processing, synthesis and recognition,” *Second Edition*. Vol. pp 243-328.2001.
- [33] Antonio M. Peinado. Jose C.Segura “Speech recognition over digital channels”*JohnWiley & Sons Ltd*. Vol. pp 7-29, 2006.
- [34] John Holmes. Wendy Holmes “Speech synthesis and recognition” *Taylor & Francis e-Library*, Second Edition. Vol. pp 161-164.2003.
- [35] Matthias Wölfel. John McDonoug “Distant Speech Recognition” *Wiley*.2009. pp 169-179.
- [36] L. Bouchakour, M. Debyeche: (Prosodic Features and Formant Contribution for Speech Recognition System over Mobile Network). *International Joint Conference SOCO’13-CISIS’13-ICEUTE’13 - Salamanca, Spain, September 11th-13th, 131-140* 2013 Proceedings.
- [37] L Bouchakour, M Debyeche (Using the Compensation of LSP Coded and Auxiliary Parameters for Improving the Speech Recognition System in Mobile Environment) *Mediterranean conference on information and communication technologies MedICT 2015,Morocco may 7-9 2015*
- [38] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [39] B. Kingsbury, T. N. Sainath, and H. Soltau, “Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization,” in *Proc.Interspeech*, 2012.
- [40] F. Seide, G. Li, and D. Yu, “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks,” in *Proc. Interspeech*, 2011.
- [41] N. Jaitly, P. Nguyen, A. W. Senior, and V. Vanhoucke, “Application Of Pretrained Deep Neural Networks To Large Vocabulary Speech Recognition,” in *Proc. Interspeech*, 2012.
- [42] Karel Vesel’y, Arnab Ghoshal, Luk’a’s Burget, Daniel Povey “ Sequence-discriminative training of deep neural networks” *Interspeech* 2013
- [43] Frank Seide, Gang Li and Dong Yu ‘Conversational Speech Transcription Using Context-Dependent Deep Neural Networks’ *Interspeech* 2011. pp 337-340.2011
- [44] Mohammad Abushariah, Raja Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa “Arabic Speaker-Independent Continuous

Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus” International Arab Journal of Information Technology, Vol. 9, No. 1, January 2012.

- [45] Eiman Alsharhan , Allan Ramsay “ Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions” Information Processing and Management. Elsevier. pp 1–11. 2017
- [46] Mansour Alsulaiman, Awais Mahmood , Ghulam Muhammad “ Speaker recognition based on Arabic phonemes” Speech Communication vol 86. Elsevier. Pp 42–51.2017.
- [47] Mohamed Amine Menacer et al “Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect “ 3rd International Conference on Arabic Computational Linguistics, , 5–6 November 2017, Dubai, United Arab Emirates . Procedia Computer Science . pp 81–88. 2017.

L. BOUCHAKOUR received his Engineer degree in electronics from University of Bordj Bou Arreridj, and the M.Sc. degree in Speech Communication from USTHB, Algeria, in 2008 and 2012 respectively. He is now working as a researcher at the Technical Research Center for the Development of Arabic Language (CRTDLA), Algeria. He is also a member of Speech Communication and Signal Processing Laboratory (LCPTS) at USTHB. His main research area includes signal and speech processing, human machine interaction, deep learning, communication mobile network, speech enhancement, speech recognition. speaker recognition and Optical character recognition.

M. Debyeche received the engineering degree in electronics from ENP, Algeria, the magister degree in signal processing, and the PhD degree in speech recognition from USTHB University, in 1982, 1991 and 2007, respectively. He is currently a professor of electronics at the Faculty of Electronics and computer science at USTHB. His research interests include speech and speaker recognition, and multi-modal pattern recognition applied to Arabic language.