Improving Continuous Arabic Speech Recognition in Mobile Environment Using MFCCs Features Modified

1st Lallouani Bouchakour
University of USTHB Faculty of Electronics and Computer Sciences (FEI), Department of Telecommunication, LCPTS
P.O. Box 32, Bab Ezzouar, Algiers, Algeria
lbouchakour@usthb.dz

2nd Mohamed Debyeche
University of USTHB Faculty of Electronics and Computer Sciences (FEI), Department of Telecommunication, LCPTS
P.O. Box 32, Bab Ezzouar, Algiers, Algeria
mdebyeche@gmail.com

Abstract – We argue that the improved the performance of automatic speech recognition (ASR) systems in mobiles communication system, we have achieved by two modules front-end or feature extractor used and a back-end or recognizer. the front-end we have used 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 systems of classification in the field of speech using: CHMM (continues hidden Markov models). Our findings show that HMM can achieve consistently almost 1.93% of clean speech, 5.23% of AMR-NB coder and 1.1% of DSR coders. The system was trained using the 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, Front-end, GF-MFCC, Back-end, CHMM, AMR-NB, DSR.

1. 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. 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. 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 the Gabor features for reduction of spectral amplitude variation in speech recognition on mobile communication. 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, the hidden Markov models (HMMs) are widely used in many systems, a statistical framework that supports both acoustic and temporal modeling.. 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 Mel features and Gabor features the second module is the back-end based on hidden Markov model (HMM). 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 the Front-End features extraction. The back-end (recognition system) in section4, we present the technique HMM . The experiments are presented in section 5. Finally, we summarize the conclusion of the presented work in section 6.

2. SPEECH CODEC

2.1. 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 [1]. The AMRspeech 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 [1,2], 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 synthesisis computed by filtering the excitation signalthrough 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.

2.2. 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 an 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) [2,3] 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 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 [2,3]. 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.

3. FEATURES EXTRACTION (FRONT-END)

The term “front-end analysis” refers to the first stage of ASR [3], 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.

3.1. MFCC standard

The MFCC coefficientis 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 [3]. The Mel scale can be calculated by Eq.(1).

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

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

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

3.2. 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 [6]. 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) .

$$s(n,k) = e^{\left[iw_n(n-n_0)+iw_k(k-k_0)\right]} \quad (3)$$

$$h(n,k) = 0.5 - 0.5\cos\left(\frac{2\pi(n-n_0)}{w_n+1}\right)\cos\left(\frac{2\pi(k-k_0)}{w_n+1}\right) \quad (4)$$

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

3.3. Dimensionality Reduction Techniques LDA

The feature vectors collected from the speech signal for continuous speech classification. If the technique such as HMM 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) [12]. 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[11]. 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.

4. SPEECH RECOGNITION SYSTEM

The speech recognition is basically a pattern recognition problem, although most state-of-the-art approache to speech recognition are based in the use of HMMs and GMMs, also called continuous density HMMs (CD-HMMs)[7],these models are all based on probability estimates and maximization of the sequence likelihood.While the neural network based in Maximum A Posteriori criterion,

4.1. Hidden Markov Model (HMM)

The HMM has been the dominant techniquefor ASR for at least two decades. One of the critical parametersof HMM is the state observation probability distribution. The gaussian mixture HMMs are typically trained based onmaximum likelihood criterion [7,10].
The input speech from a microphone is converted into a sequence of fixed size acoustic vectorsY 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

$$\hat{W} = arg_{w} \; max \{ P(w / Y) \} \quad (5)$$

The Y model can be characterized by the transitions Aij and emitting matrix probabilities Bj(Xi).

4.2. 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 [7]. The definition of the multi-variate normal distribution gives the output probability compactly in matrix notation:

$$b_j(o_t) = p\left(x / s_j \right) = \sum_{k=1}^{K} g_k N(o, \mu_k, \Sigma_k) \quad (6)$$

$$N(o, \mu_k, \Sigma_k) = \frac{1}{2\pi^{d/2} |\Sigma_k|^{1/2}} exp(-\frac{1}{2}(o-\mu_k)^t |\Sigma_k|^{-1} (o-\mu_k)) \quad (7)$$

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 $(o-\mu_k)$.

5. EXPERIMENTS

5.1. Corpus analyses

Our work consists of a sequence of tow phases shown in Fig. They consist of: Corpus Acquisition and Phoneme Segmentation.
1-Corpus acquisition a 440 sentences recording of standar Arabic speech [13]. It contains a list of Arabic sentence, an Arabic dictionary and script files used for manipulating corpus information.
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.
Arabic phoneme, the Arabic phoneme set used in the corpus is shown in Table 1. Every phoneme is corresponding English symbol. The regular Arabic short vowels /AE/,

/IH/, and /UH/ correspond to the Arabic pronunciation Fatha, Damma, and Kasra respectively.

Table 1. 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

5.2. Analysis of results

To test the speech recognition performance of the HMM, 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.
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.
5.3. Recognition Accuracy (RA)
The recognition accuracy (RA) gives the recognition results.

$$RA\% = \frac{N - D - S}{N} \times 100 \qquad (8)$$

The error rate ER is given by the following equation:

$$ER = 100\% - RA\% \qquad (9)$$

N: is the total number of units (words), D: is the number of deleted errors, S: is the number of substituted errors.

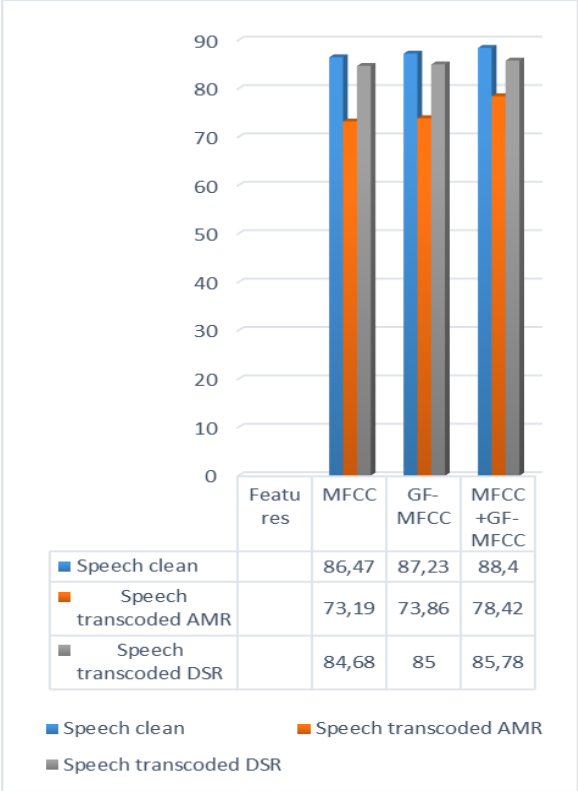

Figure 1: Recognition Accuracy of Arabic Continuous Speech with Clean and Transcoded AMR/DSR trained HMM

Table shows the results of HMM classifier with different speech conditions (clean and transcoded DSR/AMR) using MFCCs and its transformation GF-MFCC.
The overall classification accuracy using the MFCCs coefficients and transformed MFCCs that are GF-MFCC, and MFCC concatenated with MFCC transformed MFCC+GF-MFCC. (RA=86.47%) and (RA=87.23%) and (RA=88.4) for speech clean, (RA=73.19%), (RA=73.86%, (RA=84.42%) for speech transcoded AMR-NB, respectively and for speech transcoded DSR (RA=84.68% MFCC, (RA=85%) GF-MFCC and (RA=85.78%) MFCC+GF-MFCC. The statistical analysis of the MFCCs features is studied [4,5,13,14,15,16] 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 MFFCs concatenated GF-MFCC+MFCC that are generated in this work increased the overall classification accuracy about 1.93% for speech clean, 5.23% for speech transcoded AMR-NB and 1.1% for speech transcoded DSR.

6. 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 CHMM. As one of the most popular feature sets in the speech signal processing, MFCCs are proved ineffective in speech transcoded in literature. The obtained results suggest that the MFCCs transformed GF-MFCC parameters could improve the ASR performance in mobile communication.

REFERENCES

[1] Universal Mobile Telecommunications System (UMTS); AMR speech Codec. ETSI TS 126 090.

[2] 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.

[3] Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms ETSI ES 201 108.

[4] Addou Djamel, Selouani Sid-Ahmed “Optimisation of multiple feature stream weights for distributed speech processing in mobile environments ” IET Signal Process., 2015, Vol. 9, Iss. 4, pp. 387‐394

[5] 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.

[6] 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

[7] 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.

[8] Gernot A. Fink “Markov models for pattern recognition,” Springer.vol. pp. 61-92. 2008.

[9] Sadaoki Furui “Digital speech processing, synthesis and recognition,” Second Edition. Vol. pp 243-328.2001.

[10] Antonio M. Peinado. Jose C.Segura “Speech recognition over digital channels”JohnWiley & Sons Ltd. Vol. pp 7-29, 2006.

[11] John Holmes. Wendy Holmes “Speech synthesis and recognition” Taylor & Francis e-Library, Second Edition. Vol. pp 161-164.2003.

[12] Matthias W¨olfel. John McDonoug “Distant Speech Recognition” Wiley.2009. pp 169-179.

[13] 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.

[14] 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

[15] Mansour Alsulaiman, Awais Mahmood , Ghulam Muhammad " Speaker recognition based on Arabic phonemes" Speech Communication vol 86. Elsevier. Pp 42–51.2017.

[16] 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.