EasyChair Preprint

№ 13167

Robust Features in Deep Neural Networks for Transcoded Speech Recognition DSR and AMR NB

Lallouani Bouchakour, Mohamed Debyeche and Ahmed Krobba

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

May 2, 2024

Robust Features in Deep Neural Networks for Transcoded Speech Recognition DSR and AMR-NB

Lallouani Bouchakour
Speech Communication and Signal Processing Laboratory, Université des Sciences et de la Technologie Houari Boumediene (USTHB), and Centre for Scientific and Technical Research on Arabic Language Development (CRSTDLA) Algiers, Algeria
lbouchakour@usthb.dz,
l.bouchakour@crstdla.dz

Mohamed Debyeche
Speech Communication and Signal Processing Laboratory, Université des Sciences et de la Technologie Houari Boumediene (USTHB), Algiers, Algeria
mdebyeche@usthb.dz

Ahmed Krobba
Speech Communication and Signal Processing Laboratory, Université des Sciences et de la Technologie Houari Boumediene (USTHB), Algiers, Algeria
akrobba@usthb.dz

***Abstract*— Automatic speech recognition (ASR) performance in mobile communications degrades significantly when the environment contains many sources of variability. For example, when the test environment differs from the training environment, and when the acoustic environment contains disturbances such as noise, channel distortion, speaker differences, and mobile codecs. In this work, we have used two mobile network speech recognition architectures. The first one is Distributed Speech Recognition based on the DSR codec, and the second architecture is based on the Adaptive Multi-Rate Narrow-Band (AMR-NB) codec. We propose a novel robust feature extraction (Front-End) technique to improve speech recognition performance in noisy mobile communications. This technique utilizes special parameters such as Gabor features, Power Normalized Spectrum Gabor filter (PNS-Gabor), and Power Standardized Cepstral Coefficients (PNCC). These features consider psychoacoustic effects like the temporal masking effect and have different distributions of filter banks and filter forms to better model human perception. In the back end, we investigated speech classification systems using Continuous Hidden Markov Models (CHMM) and Deep Neural Networks (DNN). Based on the results obtained in noisy mobile communications, the proposed features PNS-Gabor and PNCC show significant improvements over conventional acoustic features such as Mel frequency cepstral coefficients (MFCC)**

Keywords—ASR, DSR, AMR-NB, PN-Gabor, PNCC, MFCC, HMM, DNN

I. Introduction

The goal of automatic speech recognition (ASR) is to accurately convert human speech, including sentences, words, or phonemes, into written text. However, speech recognition on mobile devices can be influenced by various factors that can affect its accuracy, such as codecs, channel transmission, and background noise. These factors can create challenges and hinder the performance of the speech recognition engine.

To overcome these challenges, mobile devices often employ noise-reduction algorithms and specialized microphones. These technologies work together to eliminate unwanted noise and enhance the quality of the captured speech, thereby improving the accuracy of the speech recognition system. Additionally, users can play a role in improving speech recognition accuracy on their mobile devices by speaking clearly and utilizing robust front-end features for speech recognition. Furthermore, minimizing background noise as much as possible can also contribute to achieving better results in speech recognition on mobile devices. [1][2].
The European Telecommunications Standards Institute (ETSI) [3] has developed a specific standard for automatic speech recognition in mobile communications, leading to client-server communication. Two systems are proposed: Network Speech Recognition (NSR) and Distributed Speech Recognition (DSR) [3][4]. The ASR system consists of two modules: a Front-End and a back-end. Feature extraction in the Front-End is inspired by the human auditory system's ability to analyze speech under challenging acoustic conditions. Principles of auditory signal processing have been integrated to improve ASR system performance.

While Mel frequency cepstral coefficients (MFCC) are the most popular features for ASR and perform well in ideal operating conditions (clean speech), they are not suitable for challenging conditions. In this study, our goal is to improve the Front-End by extracting spectro-temporal features with independent spectral and temporal processing. Physiological and psychoacoustic studies have shown that primary neurons in the auditory cortex are sensitive to spectro-temporal modulations, leading us to use Gabor features and power standardized cepstral coefficients (PNCC) for improved feature extraction in ASR.

Gabor features are applied to a spectro-temporal representation of the speech signal, using physiologically inspired filters (Gabor filters) [23]. PNCC coefficients aim to obtain more robust speech recognition characteristics in the presence of acoustic variability. This study aims to evaluate and compare the robustness of these proposed features (Front-End) for speech recognition.

We conducted experiments by transcoding speech using two codecs: Adaptive Multi-Rate Narrow-Band (AMR-NB) and Advanced Front-End for Distributed Speech Recognition (ETSI-AFE). We also introduced noise to the speech at different signal-to-noise ratios (SNRs) ranging from 10 dB to 20 dB to simulate noisy speech. The transcoded speech was then processed using the AMR-NB and DSR codecs [22]. For the back-end, we employed a continuous hidden Markov model (CHMM) [12] and Deep Neural Networks (DNN) [13]. Our results indicate that the DSR codec is the most suitable for speech recognition over a mobile network in a noisy environment. Additionally, the DNN classification outperforms HMM. The remaining sections of this paper are organized as follows: Section 2 provides a detailed description of the speech codecs DSR and AMR-NB. Section 3 focuses on the Front-End and proposes two techniques for reducing spectral variance. Section 4 presents the back-end recognition system utilizing both HMM and DNN technologies. Experimental results are presented in Section 5, and finally, the paper concludes with a summary in Section 6.

II. SPEECH CODEC

A. *Adaptive Multi-Rate Narrow-Band AMR-NB Codec*

Adaptive Multi-Rate Narrow-Band (AMR-NB) speech coding research has progressed substantially in recent years and several algorithms are rapidly entering consumer products. Several cellular telephone standards adopted the Algebraic Code-Excited Linear Prediction (ACELP) algorithms. The AMR speech coder consists of a 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 transmission errors and lost packets. The coder operates at eight different bit rates that are referred to as coder modes from 4.75 Kbit/s to 12.2 Kbit/s. For each frame of 20 ms with 4 sub-frames the speech signal is analyzed and the ACELP parameters are extracted, as linear prediction coefficients (LPC) and indices of the adaptive and Fixed Codebook. The decoder consists of decoding the transmitted ACELP parameters and performing synthesis to obtain the reconstructed speech.

B. *Distribution Speech Recognition DSR*

Mobile networks can degrade speech recognition systems due to low bit rate speech coding and channel transmission errors. Distributed Speech Recognition eliminates these problems by removing the speech channel. Instead of using an error-protected data channel, sending a parameterized representation of the speech, which is suitable for recognition. Speech recognition processing is distributed between the terminal and the network. The DSR is based on the Mel-Cepstrum representation used extensively in speech recognition systems. The feature vector consists of 14 coefficients: the log-energy coefficient and the 13 cepstral coefficients (C0, .. C12), and compressed by split vector quantification (SVQ) and then transmitted over a data channel to a remote "back-end" recognizer.

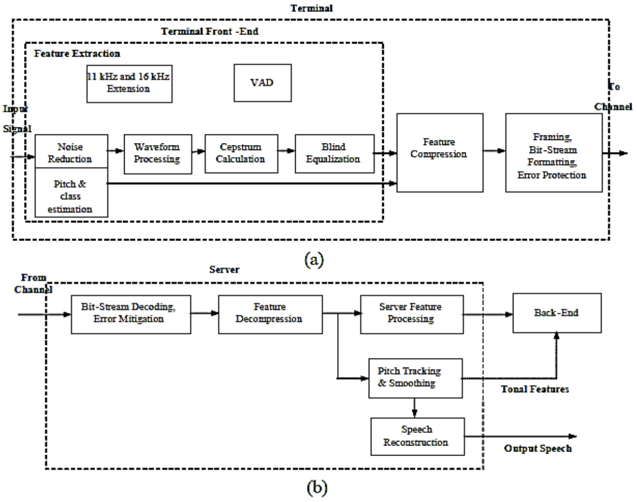

Figure 1 Architecture DSR extended front-end (a) blocks implemented at the terminal side (b) blocks implemented at the server side

III. FEATURES EXTRACTION (FRONT-END)

The input speech from a microphone is converted into a sequence of fixed-size acoustic vectors (Front-End). The term "front-end analysis" refers to the first stage of ASR.

A. *Mel Frequency Cepstral Coefficients Features (Mfcc)*

Mel Frequency Cepstral Coefficients Features (MFCC) can be used as a good feature vector to represent human speech. The frequency bands in MFCC are equally spaced on the mel scale which closely approximates the human auditory system response [15]-[3]. MFCC has two types of filters, which are spaced linearly at low frequency <1khz and logarithmic spacing > 1000Hz. The Mel scale can be calculated by Eq.(1) [12].

$$Mel(f) = 2595 \times \log_{10}(1+\frac{f}{700}) \qquad (1)$$

The MFCCs coefficients are given by the equation

$$C_n = \sqrt{\frac{2}{N}\sum_{j=1}^{N} A_j}\ \cos(\frac{n\pi j}{2N}) \qquad (2)$$

Where Aj is the output of the j-th filter bank and N is the number of samples in a basic unit [14].

B. *Power Normalized Cepstral Coefficient*

The structure of Power Normalized Cepstral Coefficient PNCC is similar to MFCC, but this feature extraction system is more adapted to representing physiological observations. The motive of extraction features PNCC is to obtain a set of practical features for speech recognition that is more robust concerning acoustical variability in their native form, without loss of performance when the speech signal is undistorted, and with a degree of computational complexity that is comparable to that of MFCC coefficients [9].

$$P[m,l] = \sum_{k=0}^{(\frac{K}{2})-1} \left|X[m,e^{jw_k}]H_l(e^{jw_k})\right|^2 \qquad (3)$$

Where K is the DFT size m, and l represents the frame and channel indices, respectively, and $\omega k = 2\pi k/Fs$, with Fs representing the sampling frequency. $X[m, e^{jw_k}]$ Is the short-time spectrum of the mth frame of the signal and $H_l(f)$ is the frequency response of the lth gammatone channel. Structure of the PNCC algorithm, the major innovations of PNCC processing include the redesigned nonlinear rate-intensity function, along with the series of processing elements to suppress the effects of background acoustical activity based on medium-time analysis. The processing we estimate a quantity we refer to as "medium-time power" $\tilde{Q}[m,l]$ by computing the running average of $P[m,l]$, the power observed in a single analysis frame, according to the equation:

$$\tilde{Q}[m,l] = \frac{1}{2M+1}\sum_{m'=m-M}^{m+M} P[m',l] \tag{4}$$

C. Mean Power Normalization

Mean power normalization is an important part of this method when power-law nonlinearity is used. Because additive shifts can't be removed by cepstral mean normalization. We normalize the power by a running mean power computed from the following equation:

$$\mu[m] = \lambda_\mu \mu[m-1] + \frac{(1-\lambda_\mu)}{L}\sum_{l=0}^{L-1} T[m,l]$$

Where m and l are the frame and channel index. In addition, L represents the number of frequency channe (5) set the forgetting factor λ_μ= 0:999 as in PNCC [15]. Then we obtain the normalized power by the following equation:

$$U[m,l] = k\frac{T[m,l]}{\mu[m]}$$

We use the value 1/15 for the pressure exponent as described in PNCC processing to get reasonable accuracy in both white noise and clear speech. (6)

$$V[m,l] = U[m,l]^{1/15} \tag{7}$$

D. Gabor Features

The usage of Gabor filters in speech recognition was driven by physiological technology and has been obtained from the measurements of so-called spectro-temporal response fields (STRF) of primary auditory cortex (AI) cells, which summarizes the way neuron cell responds to the stimulus. The PNS Gabor features are obtained by convolving two-dimensional modulation filters and the power-normalized cepstral. The PNS-Gabor uses gammatone filters followed by power bias subtraction and power nonlinearity.

The convolution of the Gabor functions $g_{u,v}(t,f)$ with the power spectrum $X(t,f)$ is given by:

$$G_{u,v}(t,f) = |X(t,f) * g_{u,v}(t,f)|, \tag{8}$$

The convolution results $G_{u,v}(t,f)$ are spectro-temporal features with different filter characteristics[11], which investigate the multilinear feature space [15][25]. Each Gabor filter $g(n,k)$ is a product of a complex sinusoid $s(n,k)$ with a Hann envelope function $h(n,k)$.

$$s(n,k) = e^{[i\omega_n(n-n_0)+i\omega_k(k-k_0)]} \tag{9}$$

$$h(n,k) = 0.5 - 0.5\times\cos(\frac{2\pi(n-n_0)}{\omega_n+1})\cos(\frac{2\pi(k-k_0)}{\omega_k+1}) \tag{10}$$

The ωn and ωk terms represent the complex sinusoid's time and frequency modulation frequencies, while ω_n and ω_k represent time and frequency window lengths of the Hann window.

IV. Speech Recognition System (Back-End)

A. Continues Hidden Markov Models

Speech recognition is a pattern recognition problem. Although the most state-of-the-art approach to speech recognition is based on HMMs and GMMs, also called Continuous Density HMMs (CD-HMMs)[16], these models are all dependent on probability estimates and maximization of sequence likelihood. While the neural network is based on the Maximum A Posteriori criterion[18].

HMM has been the dominant ASR technique for at least two decades. One of the critical parameters of HMM is the state observation probability distribution. Gaussian mixture HMMs are typically trained based on maximum likelihood criteria. The decoder then attempts to find the sequence of words W, which is most likely to have generated Y, and the decoder tries to find

$$W=arg_w max\{P(w/Y)\} \tag{11}$$

The Y model can be characterized by the transitions Aij and emitting matrix probabilities Bj(Xi). We have just defined a Continuous Density HMM (CDHMM). The normal distribution has only two independently specifiable parameters, the mean, μ_k, and Σ_k the covariance matrix.

$$b_j(o_t) = p(x/s_j) = \sum_{k=1}^{K} g_k N(o,\mu_k,\Sigma_k) \tag{12}$$

$$N(o,\mu_k,\Sigma_k) = \frac{1}{2\pi^{d/2}|\Sigma_k|^{1/2}}\exp(-\frac{1}{2}(o-\mu_k)^t|\Sigma_k|^{-1}(o-\mu_k)) \tag{13}$$

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 $(o-\mu_k)$.

B. Deep Neural Networks

Deep Neural Networks (DNNs) are multi-layer perceptron (MLP) with many hidden layers between their inputs and outputs[13]. In this section, we review fundamental ideas of DNN that can be used as an acoustic model for speech recognition. DNNs have achieved tremendous success in continuous speech recognition. The pre-training DNNs performing backpropagation training from a randomly initialized network can result in a poor local optimum, especially as the number of layers increases. To solve this problem, pre-training methods have been proposed to initialize the parameters better before backpropagation. The most wellknown method of pre-training is to grow the network layer by layer unsupervised. Specifically, each pair of layers in the network is considered a restricted Boltzmann machine (RBM) that can be trained using an objective criterion referred to as contrastive divergence.

The DNN is typically trained based on posterior probability criterion Eq (14) of a class S given an observation vector X, like a stack of (L +1) layers of log-linear models. Each hidden activation hi is computed by multiplying the entire input V by weights W in that layer [15].

$$P\left({}^{y=s}/_{x}\right)=P^{l}\left({}^{y=s}/_{v^{L}}\right) = \frac{e^{z^{L}(v^{L})}}{\sum e^{z_{j}^{l}(v^{L})}} = softmax(v^{L}) \tag{14}$$

$$z(v) = (w)^{T} v + a \tag{15}$$

Where ` and **a**` represent the weight matrix and bias vector in the *l*-th layer, respectively.

V. EXPERIMENTS

A. *ARABIC PHONEME*

The Arabic phoneme set used in the corpus corresponds to English symbols. The regular Arabic short vowels /AE/, /IH/, and /UH/ correspond to the Arabic pronunciations Fatha, Damma, and Kasra, respectively.

The purpose of these experiments is to investigate the performance of noise-robust speech recognition systems in mobile communications, specifically using AMR-NB/DSR.

To evaluate the speech recognition performance of HMM and DNN, we conducted a series of experiments on an Arabic database. The experiments were conducted with 8 kHz multi-condition data, including speech clean, speech transcoded with AMR/DSR, and babble noise, with signal-to-noise ratios (SNRs) ranging from 10 dB to 20 dB and denoising by the Non-Négative Matrix Factorisation (NMF) techniques separation [24]. The training set comprised 360 utterances, totaling approximately three hours of continuous speech, with an additional hour set aside for testing [24].

For the input features, 12-dimensional MFCC features were used for speech clean and AMR-transcoded speech experiences, while 14-dimensional MFCC features (12 MFCC + C0 + E) were used for DSR-transcoded speech.

The input layer was constructed using a context window of 6 frames, resulting in an input layer of either 812 or 814 visible units for the MFCC features, and hidden units of 2000*1300*1500*5000 for each layer, respectively. The final softmax output layer consisted of 40

TABLE I. SYSTEMS RECOGNITION ACCURACY (RA) HMM BASED

	SNR	MFCC	GF-MFCC	MFCC + GF-MFCC	PN-Gabor	PNCC
Speech Clean	**clean**	**95.9**	**96.21**	**97**	**95**	**95.5**
	20	**80.8**	**82**	**82.7**	**85**	**86**
	15	**72.3**	**74**	**75.1**	**80.4**	**81**
	10	**66.6**	**70**	**70.9**	**72**	**73.2**
Speech transcoded DSR	**clean**	**81.42**	**86.11**	**88.52**	**86.35**	**83.53**
	20	**78**	**69**	**70.3**	**71**	**74**
	15	**71.6**	**62**	**63.05**	**63**	**72**
	10	**62.1**	**58.8**	**59.4**	**60.5**	**65**
Speech transcoded AMR	**clean**	**80.68**	**85**	**89.78**	**88.5**	**90.1**
	20	**75**	**75.9**	**78.7**	**80.8**	**82**
	15	**70**	**72**	**72**	**74.6**	**77**
	10	**62**	**66**	**67**	**67.8**	**70.6**

TABLE II. SYSTEMS RECOGNITION ACCURACY (RA) DNN BASED

	SNR	MFCC	GF-MFCC	MFCC + GF-MFCC	PN-Gabor	PNCC
Speech Clean	**clean**	**96.5**	**96.8**	**98**	**95.8**	**95.2**
	20	**82**	**82.7**	**84.1**	**86.3**	**86**
	15	**73.8**	**74.6**	**78**	**80**	**81**
	10	**68**	**67.54**	**70**	**74.9**	**78**
Speech transcoded DSR	**clean**	**85.11**	**86.88**	**90.55**	**85.4**	**84.4**
	20	**81**	**72.2**	**78.85**	**80.4**	**80**
	15	**70.62**	**68**	**67.6**	**69.4**	**70**
	10	**61.5**	**61**	**63.1**	**65.8**	**66**
Speech transcoded AMR	**clean**	**85**	**88**	**90**	**92.2**	**92.5**
	20	**79**	**80.5**	**82.7**	**83.8**	**83**
	15	**70**	**72**	**73**	**74.6**	**78.5**
	10	**62**	**66**	**67**	**67.8**	**74.5**

B. *Analysis of results obtained*

This work presents efficient feature extraction techniques for robust continuous speech recognition performance in noisy mobile communications. Multiple databases were trained and decoded using the DNN and HTK toolkit [20]. The databases included clean, transcoded, and noisy estimated by NMF. Two classification systems, HMM and DNN, were used for a continuous ARABIC database that was transcoded and corrupted by various noise ratios (SNRs) ranging from 10 dB to 20 dB. Table 1 shows the results for different feature sets, including MFCC, Mel-Gabor features, GF-MFCC+MFCC, PNS-Gabor features, and PNCC.

In the first system, HMM, the results for clean speech are presented in row 2. It was found that the ASR system achieved 97% accuracy (MFCC + GF-MFCC) with a clean database. However, the addition of noise between 10 dB and 20 dB resulted in a decrease in recognition accuracy to 66.6% (MFCC).

The results for transcoded DSR are presented in row 3, showing that the ASR system with the clean database transcoded achieved 88.52% accuracy (MFCC + GF-MFCC). However, the addition of noise between 10 dB and 20 dB led to a decrease in recognition accuracy to **58.8** % (GF-MFCC).

Row 4 presents the results for transcoded AMR-NB. The ASR system with the clean database transcoded using AMR-NB achieved 90.1% accuracy (PNCC). However, with the addition of noise between 10 dB and 20 dB, the recognition accuracy decreased to 62% (MFCC).

These results demonstrate that noisy speech, especially when transcoded using AMR/DSR, reduces the performance of ASR compared to clean speech. They also show that the DSR noisy database achieved a higher accuracy rate compared to the AMR database rate.

The degradation in signal quality can be explained by the effects of fixed and adaptive quantization on excitation codebooks, as well as quantized spectral parameters. It is evident that when DSR was transcoded using the 14 coefficients (12 MFCC, log Energy, and C0), the accuracy rate increased compared to transcoding with different noise ratios using AMR. The most effective results were obtained with the features MFCC + GF-MFCC for AMR and PNCC for transcoded DSR. In the second table, when using the DNN system, the best result was achieved by the MFCC + GF-MFCC features, with a classification accuracy of 98% for clean speech. For speech transcoded using DSR, the classification accuracies were 95.8% (PN-Gabor) and 90.55% (MFCC + GF-MFCC). For speech transcoded using AMR-NB, the classification accuracy was 92.5% (PNCC) and 92.2% (MFCC + GF-MFCC). Based on these results, we can conclude that the AMR codec is the most suitable for speech recognition over a mobile network in a noisy environment. Additionally, the DNN classification system outperforms the HMM system in terms of accuracy.

VI. Conclusion

The primary objective of this paper is to enhance client-server communication on noisy mobile networks, specifically focusing on NSR and DSR, while mitigating the negative impact of degraded ASR performance caused by noisy environments and the AMR-NB/DSR speech codec. To achieve this, we utilized more robust front-end spectro-temporal representations, namely MFCC + GF-MFCC features, and PNCC, and implemented two recognition systems (back-end): DNN and HMM. In previous studies, it has been observed that MFCCs, which are widely used in speech signal processing, are not effective in noisy environments or when speech is transcoded. Therefore, we sought to explore alternative approaches. Our findings indicate that employing the DNN classifier with the AMR coder significantly improved ASR performance in noisy environments. These results highlight the importance of robust feature sets and the utilization of advanced classification systems to overcome the challenges posed by noisy mobile networks. By adopting these approaches, we can enhance the accuracy and reliability of ASR systems in real-world scenarios.

References

[1] Euler, S., and Zinke, J. The Influence Of Speech Coding Algorithms On Automatic Speech Recognition. In Proceedings of ICASSP, Adelaide, Australia. (1994).

[2] 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).

[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] Yoma, N., McInnes, F., Jack, M.: Improving Performance Of Spectral Subtraction In Speech Recognition Using A Model For Additive Noise. IEEE Trans. Speech, Audio Processing 6 (6), 579–582 (1998)

[6] Stouten, V., Van hamme, H., Wambacq, P.: Application Of Minimum Statistics And Minima Controlled Recursive Averaging Methods To Estimate A Cepstral Noise Model For Robust ASR. In: Proc. IEEE Int. Conf. (ICASSP), vol. 1, pp. I–I (2006). DOI 10.1109/ICASSP.2006.1660133

[7] Stouten, V., Van hamme, H., Wambacq, W.: Model Based Feature Enhancement With Uncertainty Decoding For Noise Robust ASR. Speech Communication. 48(11), 1502–1514 (2006)

[8] Windmann, S., Haeb-Umbach, R.: Parameter Estimation Of A State-Space Model Of Noise For Robust Speech Recognition. Audio, Speech, and Language Processing, IEEE Transactions on 17(8), 1577 –1590 (2009)

[9] C. Kim and R. Stern 'Feature Extraction For Robust Speech Recognition Based On Maximizing The Sharpness Of The Power Distribution And On Power Flooring' ICASSP 2010

[10] Kleinschmidt, M., Gelbart, D. "Improving Word Accuracy With Gabor Feature Extraction". In: Proc. ICSLP. 2002.

[11] Shuo-Yiin Chang et al 'Spectro-Temporal Features for Robust Speech Recognition using Power-Law Nonlinearity and Power-Bias Subtraction' October 2013. ICASSP.2013. DOI: 10.1109/ICASSP.2013.6639032.

[12] 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.

[13] Frank Seide, Gang Li and Dong Yu 'Conversational Speech Transcription Using Context-Dependent Deep Neural Networks' INTERSPEECH 2011. pp 337-340. 2011

[14] Md Jahangir Alam, Tomi Kinnunen "Multitaper MFCC And PLP Features For Speaker Verification Using I-Vectors » Speech Communication vol 55. Pp 237–251. 2013.

[15] Chang, Shuo-Yiin, and Nelson Morgan. "Robust CNN-based speech recognition with Gabor filter kernels." Fifteenth annual conference of the international speech communication association. 2014.

[16] 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.

[17] Gernot A. Fink "Markov Models For Pattern Recognition," Springer.vol. pp. 61-92. 2008.

[18] Sadaoki Furui "Digital Speech Processing, Synthesis And Recognition," Second Edition. Vol. pp 243-328.2001.

[19] Antonio M. Peinado. Jose C.Segura "Speech Recognition Over Digital Channels"JohnWiley & Sons Ltd. Vol. pp 7-29, 2006.

[20] http://htk.eng.cam.ac.uk/

[21] 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.

[22] Bouchakour, Lallouani, and Mohamed Debyeche. "MFCCs and Gabor Features for Improving Continuous Arabic Speech Recognition in Mobile Communication Modified." ICAASE. 2018.

[23] Bouchakour, L.,M. Debyeche. "Improving Continuous Arabic Speech Recognition Over Mobile Networks DSR And NSR Using MFCCS

Features Transformed." International Journal of Circuits, Systems and Signal Processing 12 (2018): 1-8.

[24] Bouchakour, L., Debyeche, M. Noise-Robust Speech Recognition In Mobile Network Based On Convolution Neural Networks. Int J Speech Technol 25, 269–277 (2022). https://doi.org/10.1007/s10772-021-09950-9

[25] Bouchakour, L., Meziani, F., Latrache, H., Ghribi, K., Yahiaoui, M., 2021. Printed Arabic Characters Recognition Using Combined Features and CNN classifier. Proc. - 2021 IEEE Int. Conf. Recent Adv. Math. Informatics, ICRAMI 2021 1–5