Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 1–8
DOI: 10.15439/2025F2032
ISSN 2300-5963 ACSIS, Vol. 44

Enhancing Arabic ASR in Noisy and Transcoding EVS Conditions: A Multimodal Deep Learning Study

Lallouani Bouchakour
0000-0003-2070-5115
Scientific and Technical Research Center for the Development of the Ar- abic Language (CRSTDLA) Algiers, Algeria.
l.bouchakour@crstdla.dz
lbouchakour@usthb.dz

Khaled Lounnas
0000-0003-2649-4419
University of Sciences and Technol- ogy Houari Boumediene (USTHB) Speech Communication and Signal Processing Laboratory (LCPTS), P.O. Box 32, Bab Ezzouar, 16111 Algiers, Algeria.
k.lounnas@crstdla.dz

Ahmed Krobba
0000-0002-7197-1870
University of Sciences and Technol- ogy Houari Boumediene (USTHB), Speech Communication and Signal Processing Laboratory (LCPTS), P.O. Box 32, Bab Ezzouar, 16111 Algiers, Algeria.
akrobba@usthb.dz

***Abstract*—In this paper, we investigate the impact of speech transcoding and noise on the performance of Arabic automatic speech recognition (ASR) systems based on deep learning. We apply Non-negative Matrix Factorization (NMF) as a denoising preprocessing step to enhance robustness to noise. Three deep architectures—CNN-LSTM, LSTM, and DNN—are evaluated using fused acoustic features including MFCCs, Mel- spectrograms, and Gabor filter representations. Experiments are conducted under four signal-to-noise ratio (SNR) conditions (−5 dB, 0 dB, 5 dB, and 10 dB) on both transcoded and non-transcoded speech. Results show that the CNN-LSTM model achieves the highest accuracy of 87% at 10 dB SNR on clean (non-transcoded) speech using multimodal features. However, speech recognition performance degrades by 2–4% when using the Enhanced Voice Services (EVS) codec, especially in high-noise environments. Specifically, accuracy drops from 65.00% to 61.43% at −5 dB SNR, and from 87.00% to 84.00% at 10 dB SNR due to transcoding. These findings highlight the negative impact of mobile codec compression on ASR systems, particularly under low-SNR conditions. Our study confirms the effectiveness and stability of NMF-based feature fusion and denoising in improving recognition, offering insights into deploying Arabic ASR in real-world scenarios such as mobile and VoIP communications.**

***Index Terms*—Audio transcoding, Noise, Arabic speech, NMF; CNN-LSTM, LSTM, DNN, SNR.**

I. Introduction

AUTOMATIC Speech Recognition (ASR) technologies have achieved remarkable performance in clean, controlled environments with the advancement of deep learning and sophisticated feature extraction techniques. Their prowess in real-world environments under hostile conditions such as mobile communication, Voice over IP (VoIP) services, and low-bandwidth channels remains an enduring challenge. In these situations, speech signals are usually distorted by not only background noise but also compression distortions due to speech codecs, such as those employed in Enhanced Voice Services (EVS). The dual distortions greatly impair speech intelligibility and acoustic coherence, leading to drastic degradation of ASR performance. Traditional automatic speech recognition (ASR) systems, being predominantly Hidden Markov Model (HMM)- and Gaussian Mixture Model (GMM)- based [1][2], are plagued with limited robustness in mildly noisy environments. Their accuracy significantly with nonlinear distortions via lossy speech compression. Deep learning models—Deep Neural Networks (DNNs), Long Short-Term Memory (LSTM) networks, and hybrid Convolutional Neural Network–LSTM (CNN-LSTM) architecture—have overwhelmed such traditional practices in recent years due to their strong ability to learn complicated speech patterns [16]. Due to all these developments, current state-of-the-art ASR engines are still very susceptible to non-stationary noise and encoding artifacts, particularly in the absence of any special preprocessing. A hard problem arises in mobile and internet communication systems, where speech signals are typically compressed by low-bitrate codecs (EVS), perceptually optimized rather than acoustically faithful [18,19]. The compression causes time-frequency distortions that mask important phonetic information, significantly degrading ASR performance. Moreover, these distortions become exacerbated under low signal-to- noise ratio (SNR) conditions, such as –5 dB, significantly making it difficult to obtain correct speech recognitionIn order to address these challenges, a speech enhancement method based on Non-negative Matrix Factorization (NMF) is put forward in this research. As an unsupervised learning algorithm, NMF decomposes the magnitude spectrogram of noisy speech into low-rank, non-negative bases and temporal activations [8,9,10,11]. With separate modeling of speech and noise components, efficient noise reduction can be achieved without prior noise training. This feature makes the approach extremely adaptive to dynamic and changing acoustic environments. Moreover, we investigate the impact of Enhanced Voice Services (EVS) transcoding on the performance of Arabic automatic speech recognition (ASR), which is an under investigated area considering the widespread use of EVS deployment in mobile wireless networks. In this regard, we compare the performances of three deep learning- based architectures: deep neural networks (DNNs), long short-term memory (LSTM) networks, and a hybrid convolutional-LSTM (CNN-LSTM) network [16]. These models are acquired on the basis of a multimodal fusion of acoustic features like Mel-frequency Cepstral Coefficients (MFCCs), Mel-spectrograms, and Gabor filter-based descriptors.

Topical area: Network Systems and Applications

By fusing complementary spectral and temporal representations of speech, our work achieves increased robustness in adverse acoustic conditions.

This paper advances the understanding of Automatic Speech Recognition (ASR) robustness in challenging acoustic environments through the following key contributions:

- A thorough analysis of ASR performance under simulated combined noise and Enhanced Voice Services (EVS)-induced distortions, considering serious degradation trends under actual-like unfavorable conditions.
- A novel preprocessing system with NMF as the underlying framework to enhance the quality of noisy and transcoded speech, significantly enhancing downstream ASR accuracy.
- A comparative study of deep learning-based ASR models through integrated acoustic representations, demonstrating their ability to successfully counter compounded speech distortions.

The remainder of this paper is organized as follows: Section II presents speech enhancement techniques based on Non-Negative Matrix Factorization (NMF) in order to establish the theoretical framework for our preprocessing approach. Section III presents the feature extraction methods investigated in this work, noting their suitability for noise and transcoded speech. Section IV describes the deep learning-based Automatic Speech Recognition (ASR) models used in this study. Section V describes the speech corpus, experiment setup, and discusses the results in various degradation conditions. Section VI summarizes the paper with the most significant results and future directions of work.

II. Automatic Speech Recognition Over Mobile Network And Speech Enhancement

Today, with rapid expansion of *cellular networks* for voice services, system design for making speech recognition systems reliable and solid in the environment is a paramount issue of research.

Noise is introduced by cellular network transmission, bandwidth constraint, signal degradation, all of which are certain to impact recognition. In an effort to combat these factors, strategies from effective robust automatic speech recognition techniques to advanced speech enhancement approaches have been developed. This section explains these strategies in depth, beginning with the exploration of how the performance of speech recognition systems under mobile network conditions, followed by implementing techniques such as Non-negative Matrix Factorization for enhancing the intelligibility and quality of speech signals

A. Speech recognition over mobile Network

The incredible developments in computing and networking have spurred a huge interest in deploying Automatic Speech Recognition on Mobile Devices and Over Communication Networks, and this trend is growing.

B. Client-server architectures for Speech Recognition

Many studies focus on architectures where speech recognition is performed on a remote server, while the mobile device acts as a lightweight client. For instance, Aggarwal et al. [1] proposed optimized protocols for real-time transmission of compressed audio streams, reducing latency and bandwidth consumption. These architectures leverage the computational power of cloud data centers to run complex recognition models

C. Audio Compression and Transmission

Recognition accuracy heavily depends on the quality of the transmitted audio signal. Research has explored compression methods tailored for speech recognition, such as specialized codecs (AMR-WB, Opus) that preserve essential speech features while minimizing bitrate. Lukas et al. [2] studied the impact of different codecs on recognition performance over mobile networks.

D. Robustness to Variable Network Conditions

Mobile networks (3G, 4G, 5G) experience fluctuating bandwidth, latency, and packet loss. Kumar et al. (2020) proposed adaptive mechanisms that dynamically adjust audio quality and recognition model complexity based on network conditions to ensure smooth user experience.

Cloud-Based Speech Recognition Models

With cloud computing advances, platforms like Google Speech-to-Text, Microsoft Azure Speech Services, and IBM Watson provide APIs accessible via mobile networks. These services utilize deep learning models trained on large multilingual datasets, offering high accuracy even in noisy environments.

E. On-device vs Network-Based Recognition

Research comparing on-device and network-based speech recognition highlights trade-offs. Chen et al. (2021) showed that on-device recognition reduces latency and enhances privacy but is limited by mobile hardware constraints, justifying cloud usage for more demanding applications.

F. Speech enhancement

Speech signals under real acoustic conditions are mostly corrupted by forms of acoustic interference. Speech enhancement techniques, particularly those using spectral subtraction, have proved to significantly improve the performance of Automatic Speech Recognition (ASR) systems under noisy conditions. The observed noisy speech signal can be modeled in the time domain as:

$$y(t) = x(t) + n(t) \tag{1}$$

where $x(t)$ denotes the clean speech signal, $y(t)$ represents the observed noisy speech, and $n(t)$ is the additive noise component. By applying the Short-Time Fourier Transform (STFT).

The signals are represented in the time-frequency domain as y(f, m), $\hat{x}$(f,m), and $n(f, m)$, corresponding to the noisy speech, estimated clean speech, and noise spectrum, respectively. The basic spectral subtraction method estimates the clean speech spectrum as follows:

$$x(f,m) = y(f,m) - n(f,m) \quad (2)$$

G. Non-negative matrix factorization

Non-negative Matrix Factorization (NMF) is a widely used technique for speech enhancement that decomposes the training data of noisy speech— typically represented as a magnitude or power spectrogram—into the product of two non-negative matrices: a basis matrix and an activation (or weight) matrix. This decomposition enables the independent reconstruction of the magnitude spectrograms of both speech and noise components [8]-[9]-[10]-[11]. Formally, given a non-negative matrix $V \in R \geq 0\ n \times m$ NMF seeks to find two non-negative matric$W \in R \geq 0\ n \times r$ and $H \in R \geq 0\ r \times m$ such that:

$$V = W * H \quad (3)$$

Here, W contains the basis vectors (e.g., spectral patterns), and H contains their corresponding activations over time. The rank r is typically chosen such that $r < min(n, m)$, resulting in a low-rank approximation of the original matrix $\boldsymbol{V}$. This decomposition allows for the modeling and separation of speech and noise components in the spectrogram domain using NMF-based reconstruction techniques [9]. After segmenting the time-domain signal, each segment is transformed into the frequency domain using the Fast Fourier Transform (FF).

III. HYBRID DEEP LEARNING ARCHITECTURES FOR ASR

In this study, the Deep Neural Network (DNN) architecture comprises three hidden layers, following the design proposed in [10]. The network is trained to perform speech enhancement by mapping noisy speech inputs to their clean counterparts. Each input sample consists of a log-magnitude spectrogram computed over a window of consecutive frames, providing temporal context. The dimensionality of the input layer corresponds directly to the size of the feature vector. The output layer generates an estimated log-magnitude spectrogram of clean speech, aiming to suppress noise components effectively. Each hidden layer activation hi is calculated through a linear transformation of the input, using a weight matrix , followed by a nonlinear activation function. This layer- wise transformation allows the DNN to learn complex mappings between noisy and clean speech spectra. The network is trained using a mean squared error loss between the predicted and target clean spectrograms.

where. $Z(v) = (w)^{\mathrm{T}} v + a$, and W and a represent the weight matrix.
respectively.

$$h_i^l = \sigma\left(\left(w_i^l\right)^T v^l + a_i^l\right) \quad (5)$$

where w^l and a^l are the weight matrix and bias, respectively, at the hidden layer l, h_i^l is the output of the neuron.

A. LSTM (Long Short-Term Memory) Model for Speech Recognition

The Long Short-Term Memory (LSTM) network is a highly evolved version of the recurrent neural network (RNN) that was originally created to mitigate the short comings of standard RNNs—most notably the vanishing and exploding gradient issues hindering learning over long sequences. LSTM architecture consists of a memory cell and three gate mechanisms—input, forget, and output gates—which manage the flow of information into, through, and out of the cell. This architecture enables the network to retain meaningful information on large time steps and thus is most appropriate for sequence data modeling of long-term dependencies such as speech. This gating architecture allows the model to effectively extract long-term temporal relationships by discarding or main-training them suitably. Because of this capability, LSTM networks have proven to be particularly beneficial in sequential data modeling applications such as voice processing, where retaining context over time is critical. In speech recognition, it is essential to preserve the temporal context of phonemes and words to correctly interpret them. The Long Short-Term Memory (LSTM) model meets this need by processing input sequences of acoustic feature vectors, for instance, Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram slices, or Gabor-based features, that represent the speech signal as a function of time. Using its internal memory characteristics, the LSTM effectively captures dynamic temporal patterns and transitions of spoken language without any need for spatial structure analysis.

Major advantages of the LSTM architecture are:

Ability to manage long-term temporal dependencies, which play a significant role in the context of continuous speech understanding.

Noise and variability insensitivity in the speech sequence length, enhancing performance under real-world conditions.

A reasonably simple and computationally efficient architecture, and thus suitable for real-time and embedded speech processing tasks.

Overall, LSTM networks continue to offer a robust and interpretable approach to sequence modeling in speech recognition tasks.

B. CNN-LSTM Model for Speech Recognition

Convolutional Long Short-Term Memory (CNN- LSTM) is a deep learning hybrid architecture which integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to efficiently learn and represent speech signals. In this case, CNNs operate on time-frequency representations such as spectrograms to learn spatial features—identifying significant acoustic patterns such as formants, harmonics, and local frequency changes. These high-level feature maps are then fed into the LSTM, where it extracts their temporal dynamics and sequential dependencies inherent in natural speech. This combined architecture has strong points, particularly in noisy acoustic scenarios. CNNs are insensitive to noise and local deformations, whereas LSTMs preserve long temporal dependencies well. In contrast to traditional models relying on hand-designed features, CNN-LSTM models learn discriminative feature representations automatically from raw inputs, reducing the demands of manual feature engineering [17]. Generally, the CNN-LSTM architecture demonstrates superior performance in speech recognition tasks by leveraging spatial and temporal modeling capabilities in combination. It is well suited for application in visual time-series input tasks and has potential in real-world and multilingual speech processing.

IV. FEATURES EXTRACTION (FRONT-END)

The front-end analysis is the preliminary step of Automatic Speech Recognition (ASR), wherein the acoustic in- put signal is mapped into a series of acoustic feature vectors. This typically involves inspection of the short-term signal spectrum, which effectively characterizes the acoustic realizations of phonetic events. The optimal front- end analysis method must be able to retain all perceptually pertinent information needed for phonetic discrimination while remaining tolerant of variations that are linguistically or phonetically insignificant. We utilize two techniques for feature extraction in this paper. One technique is perceptually motivated representations of speech that we use to align the extracted features with human perception. The second is the utilization of Gabor filter-based representations because such representations suit extracting localized spectro-temporal patterns from the speech signal [7].

A. Perceptual Speech Approach

This approach is perceptually centered on speech modeling, with the focus laid on how humans interpret and process auditory signals. Methods such as: Fourier Analysis: Used to decompose the speech signal into its frequency constituents, providing a spectral description over time. Mel-Frequency Cepstral Coefficients (MFCCs): A widely employed feature extraction algorithm that maps frequencies to the Mel scale—a more perceptually human auditory scale. MFCCs capture perceptually relevant spectral information and perform best at phoneme-level discrimination. In parallel, Gabor filter banks are used as a second alter- native, particularly for extracting spectro-temporal features from time-frequency representations. Originally designed for image analysis, Gabor filters mimic the response characteristics of visual cortex neurons by extracting local frequency, orientation, and texture details. In speech processing, they are employed to promote feature representation by identifying fine-grained spectrogram patterns for better classification performance in both clean and noisy conditions [7]. The Gabor features are employed here to retrieve robust spectro-temporal information from the speech signal. Two-dimensional (2-D) Gabor modulation filters are employed to manipulate the input spectro- gram. These filters operate in frequency and time domains and produce 2-D feature vectors that capture the patterns of localized modulation. Gabor representation describes the envelope width as a function of modulation frequency in order to possess the same number of periods at every frequency. It possesses this property so that Gabor features can be used as a wavelet-like representation in frequency and time domains too [13]-[14]. The convolution of the Gabor functions $g_{u,v}(t, f)$ with the power spectrum $X(t, f)$ is given by:

$$G_{u,v}(t, f) = |X(t, f) * g_{u,v}(t, f)| \tag{6}$$

where $*$ represents the 2-D convolution operation. These resulting feature maps constitute a collection of image-like representations, each for different time-frequency modulations and filter parameters. The underlying spectro-temporal representation utilized for Gabor filtering is often obtained from the Short-Time Fourier Transform (STFT) or Mel spectrogram. The STFT is widely used for speech analysis, where the signal is segmented into overlapping frames and transformed via the Discrete Fourier Trans- form (DFT). This complex-valued STFT obtained has both magnitude and phase. The magnitude spectrogram is created by computing the absolute value of each STFT coefficient. In situations where the amplitude spectrum is modified—e.g., by masking methods—reconstruction of the time-domain signal will typically involve retaining the original phase and applying the inverse DFT. Alternatively, Non-negative Matrix Factorization (NMF) is more likely to be applied in the Mel-frequency spectral domain, which offers a frequency resolution inspired by perception aligned with human hearing.

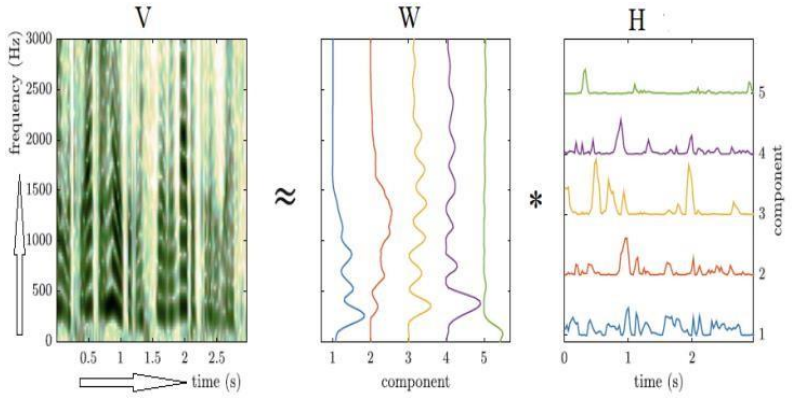

Fig. 1. The NMF model, represents the magnitude spectrum matrix V as the product of basis matrix W and H

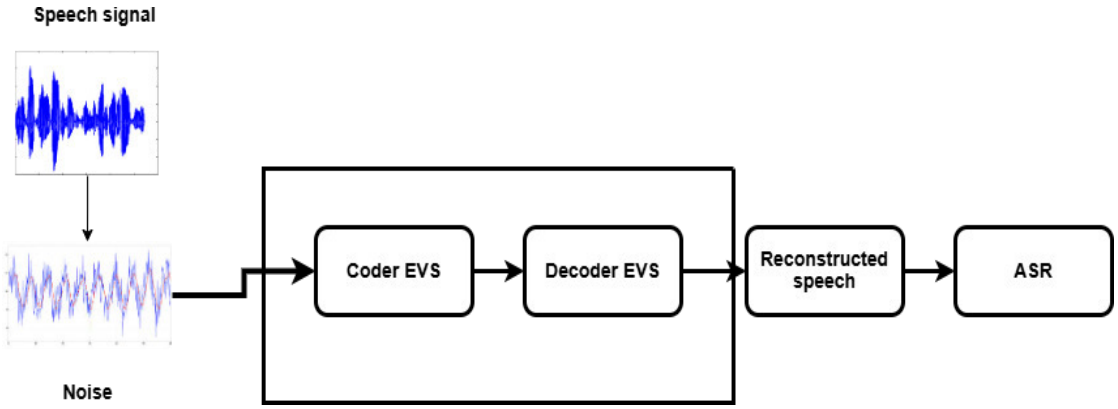

Fig.2. Speech recognition in mobile communication

TABLE. 1. SUMMARY OF ARADIGITS-BASED SPEECH DATABASES

Attribute	ARADIGIT_NOISE_NMF	ARADIGIT_EVS_NOISE_NMF
Content	Arabic digits (0 to 9)	Arabic digits (0 to 9)
Speakers	110 Algerian speakers (both genders)	Same as ARADIGIT_NOISE_NMF
Repetitions	3 repetitions per digit	Same as ARADIGIT_NOISE_NMF
Speaker Age Range	18 to 50 years	Same as ARADIGIT_NOISE_NMF
Recording Environment	Quiet room, ambient noise < 35 dB	Same as ARADIGIT_NOISE_NMF
File Format	WAV, sampled at 16 kHz, downsampled to 8 kHz	Same as ARADIGIT_NOISE_NMF
Developed By	LCPTS Laboratory	LCPTS Laboratory
Noise Type	Babble noise	Babble noise
Processing Steps	Noise added + NMF-based noise removal	Noise added + EVS transcoding + NMF-based noise removal

V. EXPERIMENTAL SETUP

In this section, we introduce the datasets, evaluation metrics.

A. Datasets

This section describes the database used to train the speech recognition models. The speech database used in this paper is the ARADIGITS database [4]. It consists of a set of 10 digits of the Arabic language (zero to nine) spoken by 110 speakers of both genders with three repetitions for each digit. This database was recorded by Algerian speakers from different regions aged between 18 and 50 years in a quiet environment with an ambient noise level below 35 dB, in .wav format, with a sampling frequency equal to 16 kHz and converting to 8kHz. We used two datasets:

1. ARADIGIT_NOISE_NMF

- ***Content***: Arabic digits from 0 to 9.
- ***Creation***: Developed at the LCPTS laboratory.
- ***Processing***:
  - This database is contaminated with various levels of babble noise.
  - The noise is then estimated and removed using the Non-negative Matrix Factorization (NMF) technique.

2. ARADIGIT_EVS_NOISE_NMF

- ***Content***: Arabic digits from 0 to 9.
- ***Creation***: Developed at the LCPTS laboratory.
- ***Processing***:
  - This database is also contaminated with various levels of babble noise.
  - It is then transcoded using an EVS.
  - Finally, the noise is estimated and removed using the NMF technique.

These databases (as illustrated in table 1) are used to evaluate the performance of our feature extraction approache under various noisy conditions by implementing advanced noise reduction techniques. We used EVS (Enhanced Voice Services) as the speech codec standardized by 3GPP for voice communication over LTE networks (VoLTE) [18]-[19]. It was developed to significantly improve audio quality compared to earlier codecs like AMR- NB and AMR-WB, while offering greater robustness to packet loss and more efficient compression.

B. Recognition Accuracy (RA)

A set of experiments was conducted to test the Recognition Accuracy (RA) by measuring the ASR performance. The Recognition Accuracy is calculated by the following equation

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

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

IV. RESULTS OF SPEECH RECOGNITION USING HYBRID DEEP LEARNING ARCHITECTURES

The following table presents the speech recognition results for speech corrupted by different levels of SNR with babble noise and estimated using the NMF technique. Two parameterization approaches are used: MFCC representing the perceptual approach and Gabor filter representing the approach. The recognition system used is based on DNNs (Deep Neural Networks).

TABLE. 2. DNN RECOGNITION ACCURACY

Model and features	**Signal**	**SNR (-5dB)**	**SNR (0dB)**	**SNR (5Db)**	**SNR (10dB)**
MFCC	Non-trans-coded	62%	69.21%	75.65%	83.05%
MFCC	Transcoded	**54.39%**	**62.15%**	**69.82%**	**77.08%**
GFMFCC	Transcoded	**55.34%**	**64.28%**	**73.17%**	**78.77%**

This table provides an overview of the speech recognition system's performance under various noise conditions, highlighting a comparison between MFCC and Gabor filter-based feature extraction methods. The use of Non-negative Matrix Factorization (NMF) for noise reduction is essential for enhancing recognition accuracy in noisy environments. SNR Level (dB): This column denotes the Signal-to-Noise Ratio levels at which babble noise was introduced. MFCC (Perceptual Approach): This column shows the recognition accuracy achieved using Mel-Frequency Cepstral Coefficients, which capture the perceptual features of speech. Gabor Filter: This column presents the recognition accuracy achieved with Gabor features, which are designed to improve the representation of speech signal parameters by analyzing time-frequency resolution.

A. Description and Analysis of Results

The Table.3 presents a comparative analysis of the performance of three speech recognition models— CNN-LSTM, LSTM, and DNN—using various feature sets (MFCC, Mel spectrogram, and Gabor filter) under different noise conditions. The models are evaluated on both non-transcoded and transcoded speech signals, with the Signal-to-Noise Ratio (SNR) and Noise-to-Speech Ratio (NSR) values reported at -5 dB, 0 dB, 5 dB, and 10 dB for each condition. Non-Transcoded Speech:

The CNN-LSTM model, using the combination of MFCC, Mel spectrogram, and Gabor filter, shows the best performance across all noise levels, achieving a significant improvement in recognition accuracy, particularly under higher noise conditions (NSR -5 dB to 5 dB), with the highest recognition accuracy of 87.00 at SNR 10 dB.

TABLE. 3. PERFORMANCE COMPARISON OF SPEECH RECOGNITION MODELS WITH DIFFERENT FEATURE S AND RECOGNITION MODEL UNDER VARYING NOISE CONDITIONS

Model and features	Signal	SNR (-5dB)	SNR (0 dB)	SNR (5 dB)	SNR (10dB)
CNN-LSTM (MFCC+mel_db+Gabor)	Non-transcoded	**65.00**	**71.00**	**77.00**	**87.00**
LSTM (MFCC+Gabor)	Non-transcoded	63.78	70.12	75.55	86.74
DNN (MFCC+Gabor)	Non-transcoded	63.50	70.00	74.33	85.00
CNN-LSTM (MFCC+mel_db+Gabor)	Transcoded	61.43	68.00	75.00	84.00
LSTM (MFCC+Gabor)	Transcoded	60.00	67.32	73.95	82.74
DNN (MFCC+Gabor)	Transcoded	59.50	65.00	70.33	82.00

This suggests that the inclusion of Mel spectrogram and Gabor filter features provides enhanced robustness against noise. The LSTM model with MFCC and Gabor filter features also demonstrates good performance, but it falls behind the CNN-LSTM model in terms of recognition accuracy, particularly as the noise level increases. Its best performance is 86.74 at SNR 10 dB. The DNN model, while still effective, shows the lowest performance compared to the CNN-LSTM and LSTM models across all noise conditions. This model achieves its best result (85.00) at SNR 10 dB. Transcoded Speech: when the speech signal undergoes transcoding, performance degrades across all models. The CNN-LSTM model still outperforms the other two models but with a notable drop in accuracy, especially under lower noise conditions (NSR -5 dB to 5 dB). It achieves a maxi- mum recognition accuracy of 84.00 at SNR 10 dB. In the similarly, the LSTM and DNN models exhibit reduced accuracy in the transcoded speech condition, with the LSTM reaching a maximum of 82.74 at SNR 10 dB, and the DNN reaching 82.00.

Overall, the CNN-LSTM model with the combination of MFCC, Mel spectrogram, and Gabor features offers the best performance across both non-transcoded and trans- coded speech, showing strong resilience against noise. However, the performance degradation with transcoding highlights the impact of signal distortion on model effectiveness, and future work could explore improving robust- ness under transcoding scenarios.

V. Conclusion

In this study, we evaluated the robustness of Arabic automatic speech recognition (ASR) systems under challenging conditions, focusing on the combined effects of noise and speech transcoding using the Enhanced Voice Services (EVS) codec. The proposed approach incorporated Non- negative Matrix Factorization (NMF)-based denoising and multi-acoustic feature fusion as a preprocessing strategy. Experimental results demonstrated that the hybrid CNN- LSTM model, combined with the proposed preprocessing pipeline, achieved the highest recognition accuracy of 87% at 10 dB SNR on clean speech. However, EVS transcoding led to a performance drop of 2–4%, particularly in low- SNR scenarios. These findings underscore the effective- ness of NMF-based denoising and the benefit of combining multiple spectral representations to enhance ASR robust- ness in real-world environments. Future work will explore advanced speech enhancement techniques and more sophisticated architectures, including self-supervised learning models, to further improve robust-ness especially in mobile telephony and multilingual con- texts.

REFERENCES

[1] C. Aggarwal, D. Olshefski, D. Saha, Zon-Yin Shae and P. Yu, "CSR. (2005): Speaker Recognition from Compressed VoIP Packet Stream,". IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands,, pp. 970-973.

[2] Drude, L., Heymann, J., Schwarz, A., & Valin, J. M. (2021). Multi- channel Opus compression for far-field automatic speech recognition with a fixed bitrate budget. arXiv preprint arXiv:2106.07994.

[3] Dong, P., Wang, S., Niu, W., Zhang, C., Lin, S., Li, Z., ... & Tao, D. (2020). Rtmobile: Beyond real-time mobile acceleration of rnns for speech recognition. In 2020 57th ACM/IEEE Design Automation Con- ference (DAC) (pp. 1-6). IEEE

[4] Amrouche, A., Debyeche, M., Taleb Ahmed, A., Rouvaen, J. M., & Ya- goub, M. C. E. (2010). Efficient system for speech recognition in ad- verse conditions using nonparametric regression. Engineering Applica- tions of Artificial Intelligence, 23(1), 85–94.

[5] Ryumin, D., Ivanko, D., & Ryumina, E. (2023). Audio-visual speech and gesture recognition by sensors of mobile devices. Sensors, 23(4), 2284.,

[6] Bouchakour, L., & Debyeche, M. (2022). Noise-robust speech recogni- tion in mobile network based on convolution neural networks. Interna- tional Journal of Speech Technology, 25(1), 269-277.

[7] Bouchakour, L., Debyeche, M., & Krobba, A. (2024). Robust Features in Deep Neural Networks for Transcoded Speech Recognition DSR and AMR-NB. In 8th International Conference on Image and Signal Pro- cessing and their Applications (ISPA) (pp. 1-5). IEEE.

[8] M. Schmidt and R. Olsson, (2006).“Single-channel speech separation using sparse non-negative matrix factorization,” in Proc. Interspeech, pp. 3111–3119.

[9] R. J. Weiss and D. P. Ellis, (2006). “Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking,” in Proc. SAPA,, pp. 31–36.

[10] Rohlfing, C., Becker, J. M., & Wien, M. (2016,). NMF-based informed source separation. In IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 474-478). IEEE.

[11] Tuomas Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 3, pp. 1066–1074, 2007.

[12] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recogni- tion. IEEE/ACM Transactions on audio, speech, and language pro- cessing, 22(10), 1533-1545.

[13] Schädler, M. R., & Kollmeier, B. (2012) Normalization of Spectro- Temporal Gabor Filter Bank Features for Improved Robust Automatic Speech Recognition Systems. In : In Thirteenth Annual Conference of the International Speech Communication Association.

[14] Schädler, Marc René; Meyer, Bernd T.; Kollmeier, Birger (2012) Spec- tro-temporal modulation subspace-spanning filter bank features for ro- bust automatic speech recognition. In : The Journal of the Acoustical Society of America, vol. 131, n° 5, p. 4134–4151. DOI: 10.1121/1.3699200.

[15] Zhao, J., Li, R., Tian, M., & An, W. (2024). Multi-view self-supervised learning and multi-scale feature fusion for automatic speech recogni- tion. Neural Processing Letters, 56(3), 168.

[16] A. Mahmoudi and M. Deriche, (2004). "CNN-BiLSTM Architectures for Arabic Speech Recognition under Noise and Compression," Neural Computing and Applications, 2024.

[17] Djeffal, N., Addou, D., Kheddar, H., & Selouani, S. A. (2023). Noise- robust speech recognition: A comparative analysis of

LSTM and CNN approaches. In 2023 2nd International Conference on Electronics, En- ergy and Measurement (IC2EM) (Vol. 1, pp. 1-6). IEEE.

[18] Dietz, M., Multrus, M., Eksler, V., Malenovsky, V., Norvell, E., Pobloth, H., ... & Zhu, C. (2015). Overview of the EVS codec architecture. In 2015 IEEE International Conference on Acoustics, Speech and Sig- nal Processing (ICASSP) (pp. 5698-5702). IEEE.

[19] Wankhede, N., & Wagh, S. (2023). Enhancing biometric speaker recog- nition through MFCC feature extraction and polar codes for remote ap- plication. IEEE Access, 11, 133921-133930.