IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 2, April, 2021, pp. 349-358
Journal homepage: http://iieta.org/journals/ts

Self-Supervised Speech Enhancement for Arabic Speech Recognition in Real-World Environments

Bilal Dendani[1,2*], Halima Bahi[1], Toufik Sari[1,2]

[1] Computer Science Department, University Badji Mokhtar Annaba, Annaba 23000, Algeria
[2] Labged Laboratory, University Badji Mokhtar Annaba, Annaba 23000, Algeria

Corresponding Author Email: bilal.dendani@univ-annaba.org

https://doi.org/10.18280/ts.380212

Received: 30 December 2020
Accepted: 12 February 2021

Keywords:
Arabic language, deep autoencoder, deep learning, self-supervised speech enhancement, speech recognition, ubiquitous systems

ABSTRACT

Mobile speech recognition attracts much attention in the ubiquitous context, however, background noises, speech coding, and transmission errors are prone to corrupt the incoming speech. Therein, building a robust speech recognizer requires the availability of a large number of real-world speech samples. Arabic language, like many other languages, lacks such resources; to overcome this limitation, we propose a speech enhancement step, before the recognition begins. For the speech enhancement purpose, we suggest the use of a deep autoencoder (DAE) algorithm. A two-step procedure is suggested: in the first step, an overcomplete DAE is trained in an unsupervised way, and in the second one, a denoising DAE is trained in a supervised way leveraging the clean speech produced in the previous step. Experimental results performed on a real-life mobile database confirmed the potentials of the proposed approach and show a reduction of the WER (Word Error Rate) of a ubiquitous Arabic speech recognizer. Further experiments show an improvement of the perceptual evaluation of speech quality (PESQ), and the short-time objective intelligibility (STOI) as well.

1. INTRODUCTION

Extensive use of handheld and wearable devices exponentially increased the use of speech as communication means and favored the widespread of ubiquitous systems that aim to be accessible anywhere and at all times. Although these devices are widely used in Arabic countries, applications related to Arabic speech recognition are scarce, this is mainly due to the lack of resources such as mobile speech corpora; this contribution aims to overcome limitations related to the speech corpus scarcity and to develop an Arabic speech recognizer for real-life environments.

Commonly, automatic speech recognition (ASR) is implemented through two main stages: extraction of the acoustic features from the incoming signal (this stage is known as the front-end part) and the decoding/recognition (this part stands for the backend part) [1]. Meanwhile, mobile speech recognition can be deployed in three architectures according to the resources' availability, the components' complexity, and the application's location, namely network speech recognition (NSR), distributed speech recognition (DSR), and embedded speech recognition (ESR) [2]. The speech signal is always captured at the client-side while the application can reside either at the client or at the server-side. Server-based models (NSR and DSR) are concerned with speech coding which refers to the speech signal representation in a digital form with few bits, thereby enabling the signal transmission while preserving the quality required for further applications.

While the front-end processing is less resource demanding, the back-end requires more resources owing to the huge number of acoustic models' parameters [3] -usually, acoustic models are hidden Markov models: HMMs [4]. ESR systems are known as client-based systems where the two aforementioned parts are located at the client-side, this discarded them from ubiquitous use. In remote speech recognition (NSR and DSR), "the speech signal quality and … robustness are two important parameters for choosing DSR while the wide deployment of high-quality speech codecs makes NSR a favorite" [3]. Meanwhile, the use of NSR architectures is preferable as the decoding and recognition tasks are too complex for the client device. Moreover, NSR enables plug-and-play of the ASR system at the server-side without changes on the client devices, which is particularly suitable for ubiquitous systems.

Although NSR architecture is suitable for ubiquitous applications, low-bit-rate codecs, background noises, and transmission errors degrade the performances of the speech recognizer. As ubiquitous systems deal with users everywhere and anytime, the speech signal is prone to be corrupted by the user background and altered by the quantization noise as well as the transmission errors. To make the speech recognition outputs reliable, the speech signal needs to be enhanced before the recognition process begins, or the acoustic models should be robust enough to recognize the transcoded (coded and transmitted) noisy speech.

Noise reduction and speech enhancement (SE) are essential for languages where training corpus available for the construction of acoustic models are limited and/or do not allow the construction of robust ASR systems. Furthermore, important applications of remote ASR, such as pronunciation learning, remote control, or healthcare, need the input speech for the ASR system to be as close as possible to the uttered one. This paper tackles speech enhancement, the step before the recognition one, and deals with the Arabic language, in the

absence of clean data for the training of the SE model.

The recent emergence of deep learning algorithms (DL) has positively impacted research in speech enhancement. Research has dealt with speech enhancement and multiple deep architectures were applied [5]. A key component in these architectures is the speech corpus used in the training stage. Most of these researches have followed a supervised approach to train the deep architecture where the input speech is noisy, the output one is clean, and the SE model learns how to recover the clean speech given its corrupted version.

In this paper, the proposed approach leverages the few available resources for the Arabic language and does not dependent on background subtraction due to the nature of ubiquitous applications which are often subject to various challenging environments. Therefore, the speech recognizer is trained using a clean Arabic corpus, while the suggested SE deep neural network is trained in an unsupervised way, both input and output speech signals are noisy. Meanwhile, as the deep autoencoder (DAE) algorithm is well suited for unsupervised learning [6, 7], we suggest its use to train the SE model. Particularly, overcomplete autoencoders architectures are investigated for this purpose. In this paper, the research question is: "Can a deep autoencoder trained in an unsupervised way, and without access to any clean training data, improve the speech recognition performance in a ubiquitous context". To answer the question, a two-step procedure is suggested: in the first step, an overcomplete deep autoencoder (OAE) is trained in an unsupervised way using noisy/noisy pairs; in the second step, a denoising deep autoencoder (DDAE) is trained in a supervised way leveraging the previous step that produces a clean version of the speech. Experimental results performed on a mobile database confirmed the potentials of the proposed approach to improve the WER (Word Error Rate) of a ubiquitous Arabic speech recognizer in a real-life environment. To compare our results against state-of-the-art methods a traditional DDAE (an under complete AE) and a Fully Convolutional Networks (FCN) were built and their WER performances compared to the proposed approach. Further experiments are performed to explore the performances in terms of the perceptual evaluation of speech quality (PESQ), the short-time objective intelligibility (STOI) at different SNR (Signal-to-Noise Ratio) levels.

This paper is organized as follows. The next section outlines the various deep learning algorithms in speech enhancement. The theoretical description of the proposed approach for the noisy transcoded (coded and transmitted) speech enhancement is detailed in Section 3. Section 4 reports the experimental results performed on an Arabic mobile dataset. Conclusions and future works are drawn at the end.

2. RELATED WORKS

In the few past years, various deep learning algorithms known as data-driven models have been proposed for speech enhancement, such as deep neural networks [8, 9], deep autoencoders [10, 11], convolutional neural networks [12-14], or fully convolutional networks [15, 16]. Deep neural networks (DNNs) are artificial neural networks with many layers, and autoencoders (AEs) "are simple networks that are trained to reconstruct the input X on the output layer X' through one hidden layer H" [17]. Encoders reduce the dimensionality of the input data to represent them in a new space (encoding) while decoders reconstruct the data from the encoding [18]; DAEs consist of encoders followed by decoders. The denoising deep autoencoders (DDAE) are DAEs that given a corrupted version of a pattern, learn the submitted pattern's essence during the encoding, and thus eliminate the superfluous at the decoding. Traditionally, DDAEs are under complete AE, i.e., the coding layers have a lower dimension than the input layer.

Convolutional neural networks (CNNs) are DNNs that process data in local regions which drastically reduce their complexity compared to the fully connected models; CNNs are particularly suitable for image analysis and classification [19-21]. A typical CNN has convolutional layers interspersed with pooling layers, followed by fully connected layers as in a multilayer neural network [20, 22]. A convolution is a mathematical linear operation and convolutional layers are a set of filters that extract feature maps to describe the characteristics of input data, the pooling layers achieve translation and rotation invariance [23].

During the training stage, typical speech enhancement systems are fed with a noisy version of a signal as input and with its clean counterpart as output. To obtain the noisy version (input) of the clean signal (output), multiple noises (car, factory, gaussian), at different levels of SNR, are added to the clean signal. In the test stage, noisy signals are submitted to the SE system and it returns an enhanced version of the inputs, the obtained results outperform those obtained by conventional methods, such as Weiner filter [24] or the minimum-mean square error (MMSE) [25]. SE systems are mainly assessed in terms of measures related to speech quality and intelligibility, such as perceptual evaluation of speech quality [26] or short-time objective intelligibility [27]. Other speech evaluation criteria include noise reduction (NR), speech distortion (SD) [28], log spectral distortion (LSD), segmental SNR (SegSNR) [29], Mean Opinion Score (MOS) of the signal distortion (CSIG) and the MOS of background noise (CBAK) [30]. Herein, we mention that the widely used measures PESQ, as well as STOI, require the existence of a reference signal. Table 1 summarizes some key works for SE, notice that for almost all of them the supervised approach is followed during the training stage via noisy/clean speech pairs.

Lu et al. [11] have used a DAE algorithm for noise reduction and speech enhancement. They reported experiments where the input and the output of the DAE were clean speech. Thus, the DAE is expected to only encode statistical information of the clean speech. In their further experiments reported by Lu et al. [28], the DAE was trained using the noisy speech as input and clean speech as output. Thereby, DAE is expected to explicitly learn the difference between clean and noisy signals.

Fu et al. [12] investigated the CNN model to restore clean speech from a noisy version and improve denoising performance using SNR-aware algorithms. Later, they applied an FCN to model simultaneously the high and the low-frequency components of a raw waveform [16]. The results of the FCN outperformed those of the CNN and the DNN based on waveform inputs.

Zhao et al. [14] have implemented a CNN algorithm to enhance the coded speech. The proposed CNN architecture included three kinds of layers: convolutional layers, max-pooling layers, and up-sampling layers. The authors compared their solution to the G.711 post-processing as a baseline and found that their proposition improved the speech quality in terms of PESQ for G.711, G.726, and AMR-WB codecs.

Table 1. Summary of different deep learning algorithms for speech enhancement

Reference	Year	Language	Noise	Input / Output	DL algorithm	Measures
[11]	2012	Japanese	White, Car, Factory, Babble	Clean/ Clean	DAE	Phone recognition accuracy
[28]	2013	Japanese	Factory, Car	Noisy/ Clean	DDAE	NR, SD, PESQ
[8]	2014	English	AWGN, Babbles, CarRestaurant, Street	Noisy/ Clean	DNN	LSD, SegSNR, PESQ
[31]	2015	English	Car, Crowd, Traffic	Noisy/ Clean	DNN	PESQ
[32]	2015	English	Home (children, TV, Radio)	Noisy/ Clean	LSTM-RNN	WER, SDR
[10]	2016	English	Office environment (stationary and non-stationary)	Noisy/ Clean	DNN	PESQ, STOI, SD, NR
[12]	2016	Mandarin	Babble, Car, Jackhammer, Pink, Street, WGN, Engine	Noisy / Clean	CNN	MSE, SegSNR
[16]	2017	English	Bable, Car, Jackhammer, Pink,	Noisy/ Clean	FCN	PESQ, STOI
[33]	2017	English	Babble, Domestic, Office, Public, Transportation, Nature, Street	Noisy/ Clean	GANs	PESQ, SegSNR CSIG, CBACK
[34]	2018	English	More than 25 types of noises	Noisy/ Clean	RNN, CNN	PESQ, SNR, WER
[14]	2019	English, German	Cafeteria, Car, Traffic road, Coding	Noisy/ Clean	CNN	PESQ
[35]	2020	Arabic	G.711 Coding	Noisy/Clean	DDAE	Accuracy

Zhao et al. [34] used a convolutional-recurrent neural network to enhance the speech submitted to the speech recognizer. To overcome the "mismatch between clean data used to train the system and the noisy data encountered when deploying the system …[that] often degrade the recognition accuracy in practice" (p.1), the authors "created a synthetic dataset" (p.3) to obtain the clean/noisy pairs of speech. For the recognition purpose, an existing deep neural network was used. While the WER of the clean speech is 2.19%, that with speech corrupted with seen noise is 15.40%, and it reaches 14.64% after the enhancement pre-processing. For the speech with unseen noise, the WER is 18.4% and it grows to 16.71% after the proposed enhancement.

3. A UBIQUITOUS ARABIC SPEECH RECOGNITION SYSTEM

The implementation of the ubiquitous speech recognition system is a complex task that requires a holistic approach to apprehend it; Figure 1 depicts the system's modules and their location; the modules are detailed below as speech coding, speech enhancement, and speech recognition.

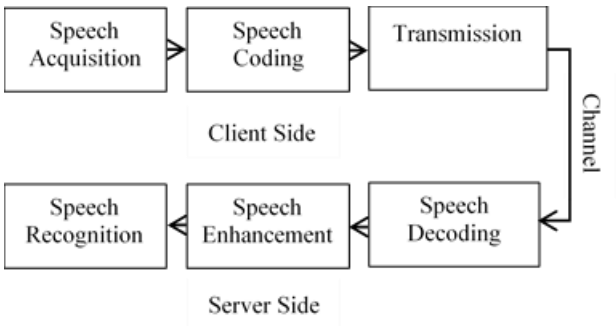


Figure 1. Block diagram for the proposed ubiquitous Arabic speech recognition system

3.1 Speech acquisition and coding

As depicted in Figure 1, the uttered speech is captured at the client-side. Afterward, the speech signal is coded before its transmission, using the software tool library G.191 standardized by ITU-T [36]. The used G.728 is an ITU-T standard for speech coding [37]; it is based on the Low-Delay Code Excited Linear Prediction (LD-CELP) compression principles, and it provides a bit rate at 16 kbps. G.728 was chosen for its low bit rate since a low bitrate allows, eventually, the use of the remaining bandwidth for video transmission (this is of amount interest for mobile applications involving other modalities than speech).

3.2 Speech enhancement

The transcoded speech signal is received at the server-side, then, the enhancement stage starts before the recognition one. The received signal was windowed into frames of 512 samples. To obtain the frequency representation, from each temporal frame were extracted coefficients corresponding to the log power spectrum [38]; as the SFFT (Short Fast Fourier Transform) produces a symmetric vector, only 257 values were kept for each frame.

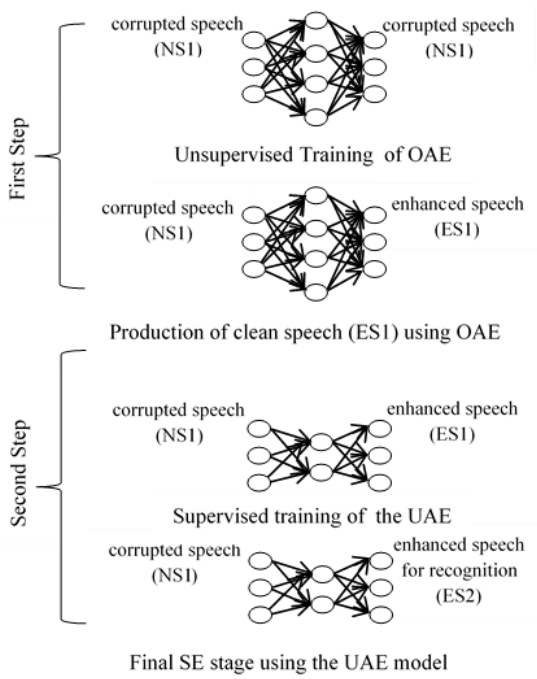


Figure 2. The two-step based DL proposed approach for speech enhancement

The obtained vector was submitted to a denoising deep autoencoder to produce the enhanced version; we called it UAE (for Under Complete AutoEncoder). The UAE has an input layer and an output layer of 257 neurons, each representing a log power spectrum coefficient (LPS). To build the UAE model, a two-step approach was followed. Figure 2 depicts the several phases of the proposed approach.

First, an overcomplete autoencoder (OAE) was trained in an unsupervised way using Adam optimizer [39] at a learning rate of 0.0001. An overcomplete AE is an AE "in which the hidden code has dimension greater than the input" [18]. During the training stage of the OAE, both the input and the output were transcoded noisy speech signals. As the OAE maps the data into a higher-dimensional space, it is intended to capture the stable structure from the inputs. Indeed, the speech signal is known as being redundant, thus the speech signal regularities should be "easy" to capture if compared to those relating to unexpected and complex noises. Once the OAE was trained, it served to produce clean data. These noisy/clean pairs of the speech signal stood for the training corpus for the denoising deep autoencoder (UAE). Finally, the speech enhancement stage is wherein the received signal is enhanced by the UAE model and sent to the ASR system.

The performance of the SE based on the two-step DAE proposed algorithm is indirectly assessed in terms of the WER score obtained after the recognition stage, given that the speech recognition task is the end-user application.

3.3 Speech recognition

Once the speech enhancement was performed, the resulted speech was fed to a speech recognizer. For the recognition, Hidden Markov Models (HMMs) stood for the acoustic models [4, 40]. Sphinxtrain was used to process and to create the acoustic models [41]; acoustic models were built from the clean modern standard Arabic corpus described by Almeman et al. [42]. During the test stage, the real-world noisy mobile utterances were decoded using PocketSphinx and the WER score computed using SphinxTrain tools. The HMM-based speech recognizer was used as a black box without tuning during our experiments related to speech enhancement.

4. EXPERIMENTS

The effectiveness of the proposed SE approach is validated by the improvement of the ubiquitous speech recognition WER score, and the evaluation of speech quality and objective intelligibility using PESQ and STOI metrics respectively.

The first experiments were conducted on a mobile dataset to show the validity of the proposition presented in Figure 2. The performance in terms of WER is reported in Table 2 through Table 4.

Once the effectiveness of the proposition was stated, other experiments were conducted in more challenging environments by considering complex noises (stationary and non-stationary) at different SNR levels applied to the recordings. These experiments focused on computing different metrics (PESQ, STOI, and WER) to compare the previous proposition against the state-of-the-art models. For that purpose, a denoising deep autoencoder is trained in a supervised way, and the UAE of the proposition (see Figure 2) is replaced by an FCN.

Additional experiments were carried out on a second corpus to confirm the results obtained with the first one.

4.1 Data sets

The Arabic mobile parallel multi-dialect speech corpus is a free Arabic corpus [43]. It contains four Arabic dialects: Modern Standard Arabic (MSA), Levantine, Gulf, and Egyptian. The corpus consists of 67132 wave files, sampled as 48 kHz with a precision of 16 bits, uttered by 52 speakers. For our experiments, the MSA subset is considered, it contains 15492 utterances from 12 speakers. The data were collected in four different environments, inside the home, in a moving car, in a public place, and in a quiet place. The chosen public areas and the streets used in the experiment varied between high noise and medium noise [43]. The noises that occur in the background can be divided into non-human (door closing, cutlery sounds, car horns, road traffic) noise, and human noise (crying, shouting, speaking). Besides, "Mobile call quality can be affected by many additional factors, such as network signal quality, recording quality, the distance between the mobile and the mouth, etc." [43]. The speech from the dataset [43] is called NS1.

Besides the inherent noises, the recordings from the study [43] were corrupted using some noises from 100 non-speech sounds [44]. Different noise types, ranging from stationary car noise to non-stationary noise (crowd and door moving) at different SNR levels, were considered. About 75% of the corrupted speech signals were used for the training set at SNR 0 dB, 15 dB, and -5 dB. For the testing set, the remaining 25% corrupted speech signals were considered at -5dB, 0 dB, and 15dB SNR levels. The NS1 speech corrupted with the noises from the study [44] is called NS2.

The second dataset is presented in the study [45]; it is an Arabic speech corpus for isolated words. The corpus has been developed by the Department of Management Information Systems, King Faisal University. It contains 9992 recorded utterances of 50 speakers pronouncing 20 words. Recordings from this corpus were corrupted using the real-world UrbanSound8K dataset [46] at different SNR levels (-15dB, -10dB, -5dB, 0dB, 5dB, and 10dB). The urban sound data set contains 8732 clip sounds. In our case, three environmental noises were considered the air conditioner, the engine idling, and the jackhammer noises.

4.2 Results and discussion

The first experiment deals with the real-world mobile corpus and aims to validate the proposition considering the final application, herein the ASR. First, we report the WER score on real-world mobile speech signals. The real-life transcoded speech signal was received at the server-side (NS1), and the HMM-based recognition took place without prior enhancement of the incoming signal. Figure 3 reports performances of the ASR system compared to those obtained after the recognition that took place at the client-side (before coding and transmission).

Figure 3 shows that the noisy background causes degradation of the WER. Indeed, the worst WER is seen in "public place" where the noise is unknown and from plural sources while the best WER is seen in the "inside home". Figure 3 also shows that the coding and transmission processes decrease yet the WER.

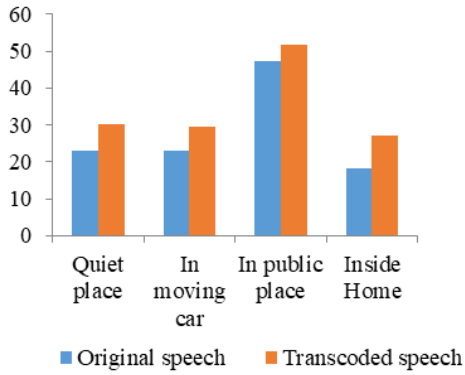


Figure 3. The ASR performance for original and transcoded speech

(1) The DAE-based SE effect on speech recognition performance

As already said, the DDAE used as the SE system has 257 neurons in the input layer as well as in the output one. The hidden layer has 200 neurons: it is UAE(200). The OAE, which served to produce clean speech for the UAE training, has the same input and output neurons' number, and it has 1024 hidden neurons: it is OAE(1024). Table 2 reports the WER values when UAE(200) is used as the SE system. Table 2 also reports the WER after the received speech was enhanced using the well-known MMSE method.

Table 2. WER (%) without and with SE

Environment	Without SE	DAE-based SE	MMSE-based SE
Quiet place	30.28	28.76	58.44
In moving car	29.59	25.19	57.02
In public place	51.76	47.56	84.46
Inside Home	26.97	20.82	57.50
Average WER	34.65	30.58	66.33

Table 2 shows the positive impact of the DAE-based SE on the WER score even in the absence of clean data to train the SE model. The use of the MMSE method makes the recognition worse than that performed without enhancement. It is known that the MMSE method does not perform well in presence of non-stationary or multiple sources' or noises.

(2) Investigating the DAEs architectures

When using autoencoders, the choice of the right degree of compression, i.e., dimensionality reduction is often a hyper-parameter that requires tuning for optimal results. Thus, once the positive effect of the proposed SE approach was proved, additional experiments were performed to fine-tune the DAEs models through the investigation of their depth and the number of the neurons in each layer as well.

Tables 3 and 4 report experiments where the OAE and UAE structures are investigated. In Table 3, the UAE hidden layer is set to 200 neurons, while multiple OAE configurations are tested. In Table 4, the UAE is set to two hidden layers each of 200 neurons.

In Table 3, all the configurations show an improvement of the WER score, stating the effectiveness of the suggested SE approach, this could be explained by the projection of the signal characteristics by the OAE in a higher dimensionality space which allows the isolation of noises' features.

Table 4 results confirm yet the potential of the SE proposed approach as all the tested configurations improved the WER of the ubiquitous speech recognition system.

Table 3. WER (%) of the ASR system for a DDAE of 200 neurons in the hidden layer

Models	Quiet place	Moving car	Public Place	Inside Home	Average
Without SE	30.28	29.59	51.76	26.97	**34.65**
OAE(1024) UAE(200)	28.76	25.19	47.56	20.82	30.58
OAE(1024,1024) UAE(200)	28.07	25.42	47.48	21.01	30.49
OAE(400) UAE(200)	27.91	24.96	48.09	20.51	30.36
OAE(400,400) UAE(200)	28.71	25.08	46.65	20.74	**30.29**

Table 4. WER (%) of the ASR system for a UAE of 200 neurons in each of the two hidden layers

Models	Quiet place	Moving car	Public Place	Inside Home	Average
Without SE	30.28	29.59	51.76	26.97	**34.65**
OAE(1024) UAE(200-200)	28.57	25.57	47.67	20.44	30.56
OAE(1024,1024) UAE(200-200)	28.18	24.96	47.60	21.01	30.44
OAE(400) UAE(200-200)	28.71	24.43	47.33	20.21	**30.17**
OAE(400,400) UAE(200-200)	28.97	24.85	47.67	20.89	30.59

The results reported in Tables 3 and 4 show an improvement of the WER score after the SE compared to that computed without applying the enhancement. This is due to the advantage of using fully connected-based models to model multiple complex real-world noisy environments. In particular, the use of an overcomplete autoencoder for unsupervised pretraining provides a solution to generate clean data and allowing the training of the classical DDAE.

To select the best deep learning model, another focus is the balance between the number of total parameters and the WER value. Indeed, one of the deep learning paradigm challenges is to optimize computational resources and the time needed for training. Table 5 reports the number of parameters for each configuration.

According to Table 5, model#2 is a good compromise between the number of parameters and the WER.

Table 5. Total number of parameters for each deep learning structure

#	Model	Number of parameters	WER %
1	OAE(400) UAE(200)	309514	30.37
2	OAE(400) UAE(200, 200)	349714	**30.17**
3	OAE(400,400) UAE(200)	469914	30.29
4	OAE(400,400) UAE(200, 200)	510114	30.59
5	OAE(1024) UAE(200)	630874	30.58
6	OAE(1024) UAE(200, 200)	671074	30.56
7	OAE(1024,1024) UAE(200)	1680474	30.49
8	OAE(1024,1024) UAE(200, 200)	2248291	30.44

(3) Comparative analysis

In addition to the WER used to assess the performance of the speech recognizer (the end-used application), the PESQ and the STOI are used to evaluate the quality and the intelligibility of the enhanced speech. For PESQ and STOI, a higher value is better, the PESQ values range in [-0.5, 4.5] while the STOI value range is [0,1].

For the following experiments, model#2 stands for the proposed model, and two additional DL models are considered. A denoising deep autoencoder trained in a supervised way. The denoising deep autoencoder (called UAE2) has the same architecture as the UAE of model#2 described in Table 5 and is trained with NS2 as input and with NS1 as output. The second model is based on a fully convolutional network that replaced the UAE in model #2, the OAE of the first step is called OAE1. However, the FCN is trained using NS2 as input and ES1 as output. The FCN architecture used for comparison is the one described in the study [47] with two and four convolutional layers instead of eight. The DDAE and the FCN models were trained with stationary and non-stationary noises at different SNR levels. The speech to recognize comprises both NS1 and NS2.

Table 6 reports the aforementioned measures on Mobile MSA corpus, it also reports the performances of the speech recognizer, in term of the WER, at different SNR levels. To compute the PESQ and the STOI, we considered the speech signal, in a quiet place as the reference one.

The two-step Model#2 outperformed the other competing models for all measures in different noisy types. For instance, the proposed approach achieved the best speech quality (PESQ) at different SNR levels, except at 15dB for car and crowd noises. Meanwhile, UAE2 provided better performances than the FCN-based models. The proposed approach achieved an average improvement of about 0.835 for the three SNR levels (0, 15, and -5)dB. Finally, Model#2 succeeded to decrease the WER by an average of 15.43% in complex noisy environments at different SNR levels.

4.3 Additional experiments

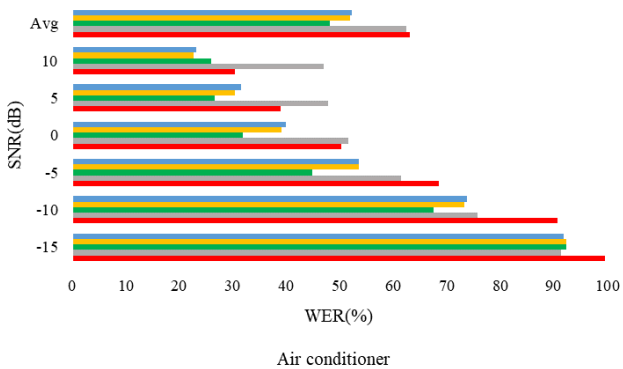
The following experiments aim to consolidate the proposition and to evaluate it on a second corpus. First, figures 4, 5, and 6 report the WER, the PESQ, and STOI respectively, for the considered SE methods, under different noise conditions (the air conditioner, engine idling, and jackhammer), and at different SNR levels.

Figure 4 shows that the WER obtained with the proposed model reached an overall value of 55.03%, and outperformed the other models. The lowest WER is seen with the engine idling noise at 33.05. In particular, often, model#2 outperformed the UAE2 model which was trained in a supervised way as the clean data is available for this corpus. Moreover, all unsupervised learning-based approaches outperformed the noisy version represented by the label (Without SE), except at engine idling under 10 dB SNR level.

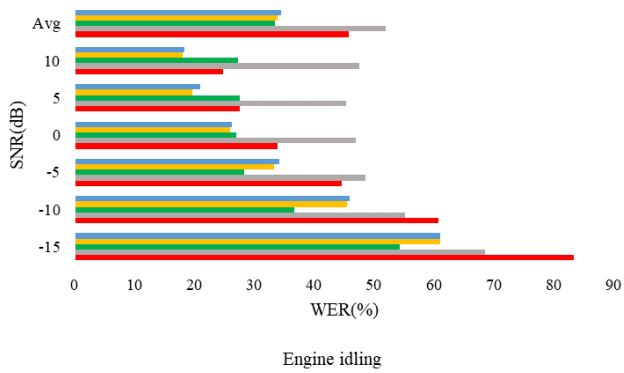
Figures 5, and 6 confirm the performances of the proposed model in terms of PESQ and STOI as well. In particular, for the low SNR values, the proposed model competes with the supervised UAE2.

Table 6. Performance measures for different models with multiple noises, and at different SNR levels

		Car Noise SNR (dB)			**Crowd Noise** SNR (dB)			**Moving Door** SNR (dB)			
Metrics	Model	-5	0	15	-5	0	15	-5	0	15	AVG
PESQ	Without SE	1.86	2.10	2.87	1.21	1.32	2.10	0.88	1.21	2.22	1.75
	UAE_2	2.33	2.57	3.06	2.52	2.61	2.85	2.15	2.26	2.72	2.56
	Model#2	2.35	2.59	3.08	2.53	2.63	2.81	2.19	2.31	2.80	**2.59**
	OAE_1-FCN_2	2.31	2.54	3.17	1.68	2.04	2.68	1.30	1.88	2.67	2.25
	OAE_1-FCN_4	2.25	2.46	3.14	1.74	1.92	2.78	1.41	1.62	2.63	2.22
STOI	Without SE	0.71	0.78	0.87	0.56	0.60	0.73	0.38	0.50	0.78	0.66
	UAE_2	0.70	0.75	0.80	0.66	0.67	0.73	0.63	0.67	0.76	0.71
	Model#2	0.70	0.75	0.81	0.67	0.69	0.75	0.64	0.68	0.77	**0.72**
	OAE_1-FCN_2	0.72	0.75	0.82	0.57	0.60	0.74	0.40	0.54	0.79	0.66
	OAE_1-FCN_4	0.68	0.75	0.81	0.61	0.67	0.76	0.45	0.58	0.79	0.68
WER	Without SE	78.50	51.89	18.25	100	100	84.94	99.62	96.26	38.83	74.25
	UAE_2	57.89	44.00	24.29	85.73	80.33	59.88	80.63	67.19	32.15	59.12
	Model#2	58.21	44.80	24.18	85.29	78.81	58.02	81.52	65.85	32.77	**58.82**
	OAE_1-FCN_2	80.24	66.97	28.32	100	100	85.87	99.69	95.59	39.85	77.39
	OAE_1-FCN_4	88.68	64.10	32.38	100	100	79.78	98.80	93.28	44.61	77.95



Air conditioner



Engine idling

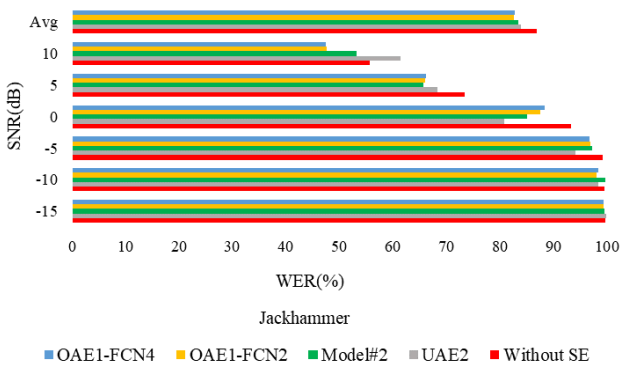


Figure 4. WER for different SE methods with multiple noises types and at different SNR levels

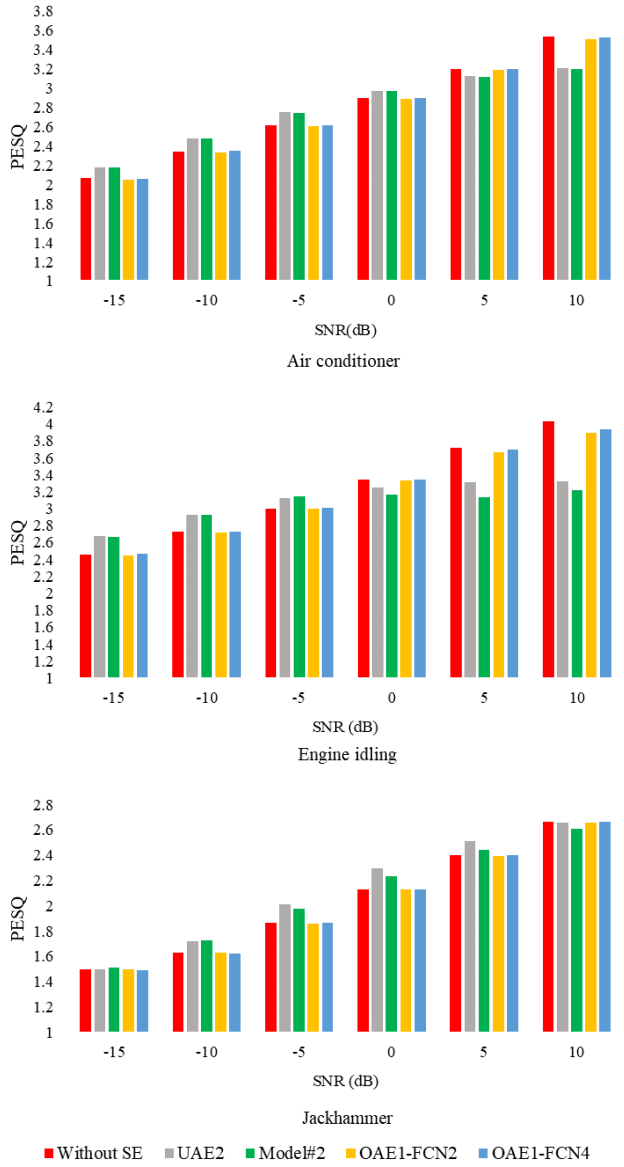


Figure 5. PESQ for different SE methods with multiple noises types and at different SNR levels

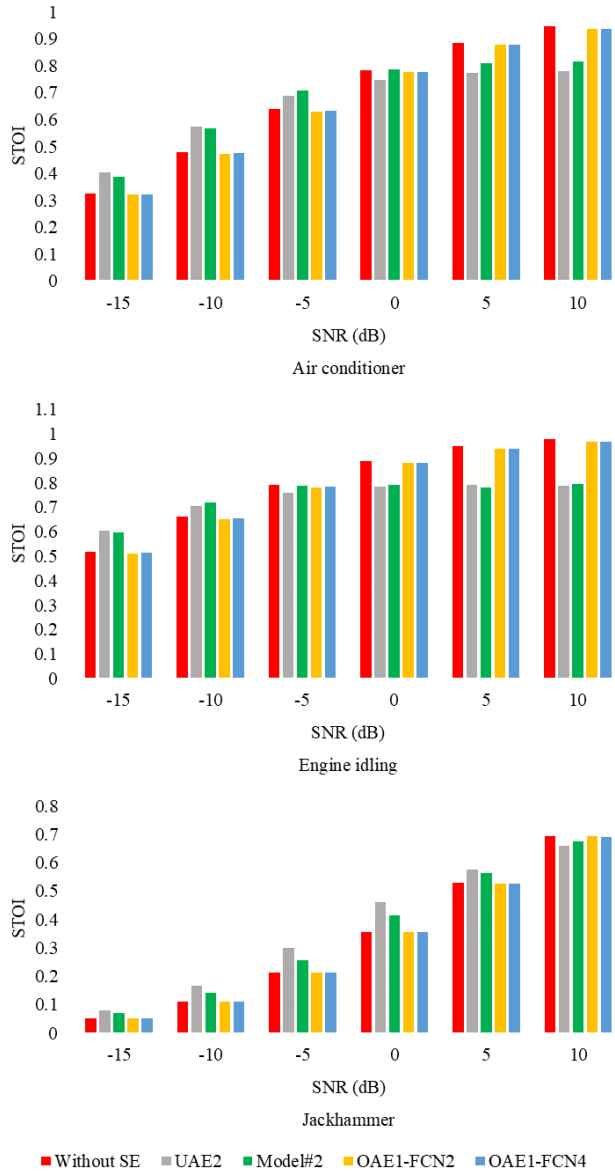
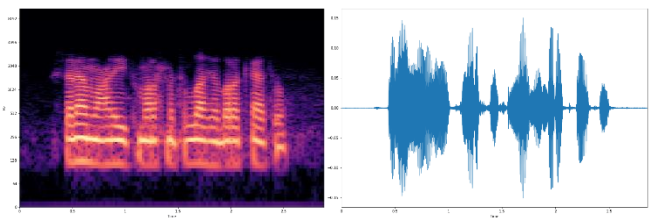


Figure 6. STOI for different SE methods with multiple noises types and at different SNR levels

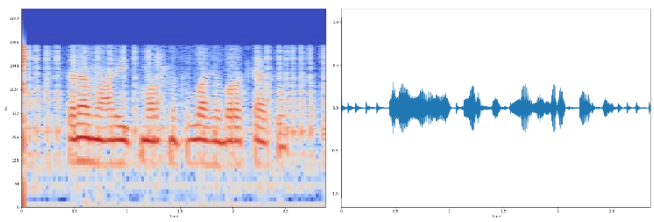
4.4 Spectrogram and waveform analysis

To visualize the effect of the different models, the reconstructed spectrograms and waveforms for a randomized selected sample from the test dataset of the mobile MSA are presented. Figure 7 shows the spectrograms and waveforms of the original speech, its noisy version, and the corresponding enhanced utterance.

Figure 7 shows that model#2 reduces the noise and conserves the speech components. UAE2 performed better than the OAE1-FCN2 and OAE1-FCN4 which are very close. The FCN-based models failed to effectively remove the noise as shown in oval red regions.



(a) Clean spectrogram-waveform



(b) Noisy spectrogram-waveform

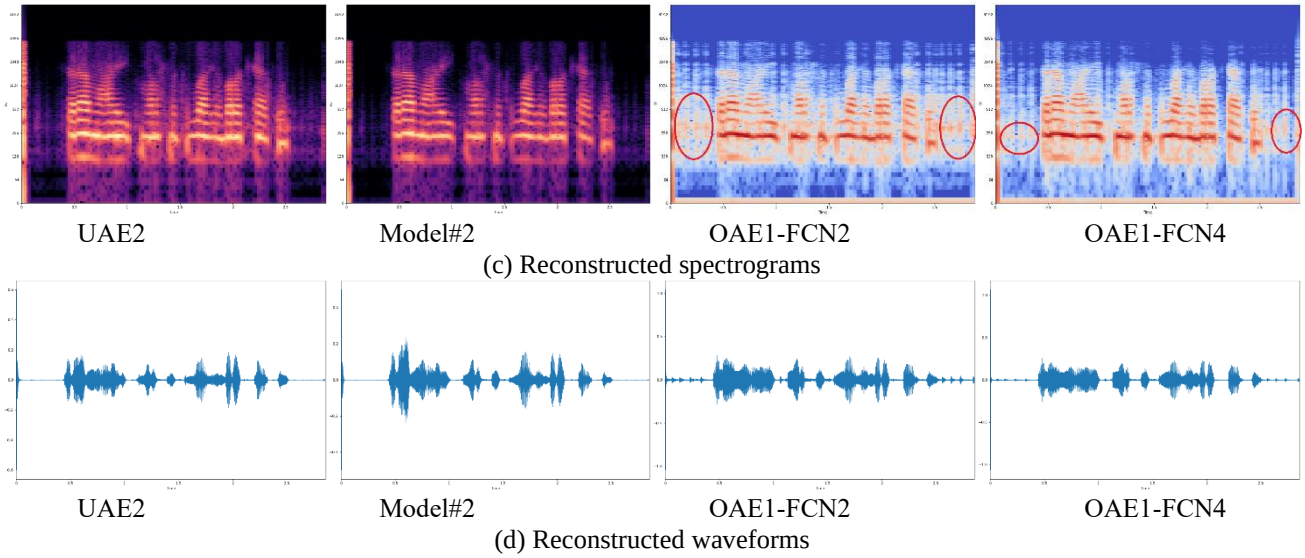


Figure 7. Examples of spectrograms and waveforms of a clean and noisy sample with their denoised versions using SE models

It is worth noting that the two steps approach (model#2) and the UAE2 alone, which are based on autoencoder models, outperformed the architectures including the FCN. In particular, the use of an overcomplete autoencoder for unsupervised pretraining provided a solution to generate clean data and allowed the supervised training of the classical DDAE.

5. CONCLUSIONS

This paper deals with Arabic speech recognition in a ubiquitous environment that aims to improve the WER score of the end-user ASR application. For that purpose, a speech enhancement approach is suggested. Speech enhancement is of paramount interest; however, it is not an easy task due to the lack of real-life labeled data (clean/noisy pairs). We proposed a two-step approach where an overcomplete deep autoencoder is trained in an unsupervised way to produce the enhanced speech, then a denoising deep autoencoder is used to produce the final enhanced speech signal to be recognized. The obtained results show an improvement of the WER of about 4.48% for the mobile MSA corpus, which makes the proposed approach an effective alternative to the implementation of robust ubiquitous speech recognition systems. Meanwhile, the two-steps unsupervised model achieves a significant improvement for speech quality (PESQ) and intelligibility (STOI) of about 0.835 and 0.06, respectively, on stationary and non-stationary noise, considering the real-world mobile dataset. For the Arabic isolated words speech corpus, an improvement of the WER of about 10% is seen.

On the other side, this work contributes to the practical speech enhancement problem by minimizing the requirements, i.e., without access to any clean training data. Indeed, the unsupervised and self-supervised SE approaches are considered as challenging topics that need more focus in future works. Meanwhile, as the indicator that measures the front-end algorithm and the accuracy of the back-end recognition are not positively correlated, the improvement of the front-end may not have a positive effect on the back-end recognition. Thus, it is expected that the back-end recognition results will feedback the front-end, which would make the system more robust.

Finally, the use of an overcomplete DAE model brings new perspectives in unsupervised learning, thus as future work, we plan to deeply explore their capabilities.

REFERENCES

[1] Bahi, H., Sellami, M. (2001). Combination of vector quantization and hidden Markov models for Arabic speech recognition. In Proceedings ACS/IEEE International Conference on Computer Systems and Applications, pp. 96-100. https://doi.org/10.1109/aiccsa.2001.933957

[2] Schmitt, A., Zaykovskiy, D., Minker, W. (2008). Speech recognition for mobile devices. International Journal of Speech Technology, Springer US, 11(2): 63-72. https://doi.org/10.1007/s10772-009-9036-6

[3] Tan, Z.H., Varga, I. (2008). Network, distributed and embedded speech recognition: An overview. In: Automatic Speech Recognition on Mobile Devices and over Communication Networks. Advances in Pattern Recognition Springer, London, 1-23. https://doi.org/10.1007/978-1-84800-143-5_1

[4] Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2): 257-286. https://doi.org/10.1109/5.18626

[5] Hepsiba, D., Justin, J. (2019). Role of deep neural network in speech enhancement: A review. Communications in Computer and Information Science, 890: 103-112. https://doi.org/10.1007/978-981-13-9129-3_8

[6] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A. (2008). Extracting and composing robust features with denoising autoencoders. Proceedings of ICML, Helsinki, Finland, pp. 1096-1103. https://doi.org/10.1145/1390156.1390294

[7] Wani, M.A., Bhat, F.A., Afzal, S., Khan, A.I. (2020). Unsupervised deep learning architectures. In: Advances in Deep Learning. Studies in Big Data, 57: 77-94. https://doi.org/10.1007/978-981-13-6794-6_5

[8] Xu, Y., Du, J., Dai, L.R., Lee, C.H. (2014). An experimental study on speech enhancement based on

deep neural networks. IEEE Signal Processing Letters, 21(1): 65-68. https://doi.org/10.1109/lsp.2013.2291240
[9] Saleem, N., Khattak. M.I. (2019). Deep neural networks for speech enhancement in complex-noisy environments. International Journal of Interactive Multimedia and Artificial Intelligence, 6(1): 84-90. https://doi.org/10.9781/ijimai.2019.06.001
[10] Kumar, A., Florencio, D. (2016). Speech enhancement in multiple-noise conditions using deep neural networks. INTERSPEECH, San Francisco, CA, USA. https://doi.org/10.21437/interspeech.2016-88
[11] Lu, X., Matsuda, S., Hori, C., Kashioka, H. (2012). Speech restoration based on deep learning autoencoder with layer-wised pretraining. In InterSpeech, Portland, OR, USA, 1504-1507
[12] Fu, S.W., Tsao, Y., Lu, X. (2016). SNR-aware convolutional neural network modeling for speech enhancement. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, CA, USA, pp. 3768-3772. https://doi.org/10.21437/interspeech.2016-211
[13] Park, S.R., Lee, J.W. (2017). A fully convolutional neural network for speech enhancement. Proceedings of INTERSPEECH, Stockholm, Sweden, pp. 1993-1997. https://doi.org/10.21437/interspeech.2017-1465
[14] Zhao, Z., Liu, H., Fingscheidt, T. (2019). Convolutional neural networks to enhance coded speech. IEEE/ACM Transactions on Audio Speech and Language Processing, 27(4): 663-678. https://doi.org/10.1109/taslp.2018.2887337
[15] Alamdari, N., Azarang, A., Kehtarnavaz, N. (2019). Self-supervised deep learning-based speech denoising. ArXiv, available at: http://arxiv.org/abs/1904.12069
[16] Fu, S.W., Tsao, Y., Lu, X., Kawai, H. (2017). Raw waveform-based speech enhancement by fully convolutional networks. Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, 2018(12): 6-12. https://doi.org/10.1109/apsipa.2017.8281993
[17] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, Elsevier B.V. https://doi.org/10.1016/j.media.2017.07.005
[18] Goodfellow, I., Bengio, Y., Courville, A.C. (2016). Deep Learning. MIT Press.
[19] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2323. https://doi.org/10.1109/5.726791
[20] Shen, D., Wu, G., Suk, H.I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, Annual Reviews, 19(1): 221-248.
[21] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386
[22] Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D. (1997). Face recognition: A convolutional neural-network approach. IEEE Transactions on Neural Networks, Institute of Electrical and Electronics Engineers Inc., 8(1): 98-113. https://doi.org/10.1109/72.554195
[23] Chiang, H.T., Hsieh, Y.Y., Fu, S.W., Hung, K.H., Tsao, Y., Chien, S.Y. (2019). Noise reduction in ECG signals using fully convolutional denoising autoencoders. IEEE Access, 7: 60806-60813. https://doi.org/10.1109/access.2019.2912036
[24] Sreenivas, T.V., Kirnapure, P. (1996). Codebook constrained wiener filtering for speech enhancement. IEEE Transactions on Speech and Audio Processing, 4(5): 383-389. https://doi.org/10.1109/89.536932
[25] Ephraim, Y., Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6): 1109-1121. https://doi.org/10.1109/tassp.1984.1164453
[26] Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Proceedigns of ICASSP, 2: 749-752. https://doi.org/10.1109/icassp.2001.941023
[27] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech and Language Processing, 19(7): 2125-2136. https://doi.org/10.1109/tasl.2011.2114881
[28] Lu, X., Tsao, Y., Matsuda, S., Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. Proceedings of the Annual Conference of the International Speech Communication Association, InterSpeech, Lyon, France, pp. 436-440.
[29] Du, J., Huo, Q. (2008). A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. Proceedings of the Annual Conference of the International Speech Communication Association, InterSpeech, Brisbane, Australia, pp. 569-572.
[30] Loizou, P.C. (2013). Speech Enhancement: Theory and Practice. 2nd Edition, CRC Press.
[31] Xu, Y., Du, J., Dai, L.R., Lee, C.H. (2015). A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio Speech and Language Processing, 23(1): 7-19. https://doi.org/10.1109/taslp.2014.2364452
[32] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J.R., Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. LNCS, Springer, 9237: 91-99. https://doi.org/10.1007/978-3-319-22482-4_11
[33] Pascual, S., Bonafonte, A., Serra, J. (2017). SEGAN: Speech enhancement generative adversarial network. Proceedings of INTERSPEECH, Stockholm, Sweden, pp. 3642-3646. https://doi.org/10.21437/interspeech.2017-1428
[34] Zhao, H., Zarar, S., Tashev, I., Lee, C.H. (2018). Convolutional-recurrent neural networks for speech enhancement. Proceedings of ICASSP, Calgary, AB, Canada, pp. 2401-2405. https://doi.org/10.1109/icassp.2018.8462155
[35] Dendani, B., Bahi, H., Sari, T. (2020). Speech enhancement based on deep AutoEncoder for remote Arabic speech recognition. International Conference on Image and Signal Processing, pp. 221-229. https://doi.org/10.1007/978-3-030-51935-3_24
[36] ITU-T (2019). Software tools for speech and audio

coding standardization, Telecommunication Standardization Sector (ITU-T), International Telecommunication Union. G.191 (01/2019).
[37] ITU-T (1992). Coding of speech at 16 kbit/s using low-delay code excited linear prediction, ITU-T Rec. G.728.
[38] Semmlow, J. (2012). The Fourier transform and power spectrum. Signals and Systems for Bioengineers, Elsevier, 131-165. https://doi.org/10.1016/b978-0-12-384982-3.00004-3
[39] Kingma, D.P., Ba, J.L. (2015). Adam: A method for stochastic optimization. Proceedings of 3rd International Conference on Learning Representations, San Diego, CA, USA.
[40] Frihia, H., Bahi, H. (2017). HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications. International Journal of Speech Technology, Springer US, 20(3): 563-573. https://doi.org/10.1007/s10772-017-9427-z
[41] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Woelfel, J., Wolf, P. (2004). Sphinx-4: A Flexible Open Source Framework for Speech Recognition, Mountain View, CA: Sun Microsystems, Inc.
[42] Almeman, K., Lee, M., Almiman, A.A. (2013). Multi dialect Arabic speech parallel corpora. 1st International Conference on Communications, Signal Processing and Their Applications, Sharjah, UAE. https://doi.org/10.1109/iccspa.2013.6487288
[43] Almeman, K. (2018). The Building and evaluation of a mobile parallel multi-dialect speech corpus for Arabic. Procedia Computer Science, Elsevier B.V., 142(2017): 166-173. https://doi.org/10.1016/j.procs.2018.10.472
[44] Hu, G., Wang, D.L. (2010). A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Transactions on Audio, Speech and Language Processing, 18(8): 2067-2079. https://doi.org/10.1109/tasl.2010.2041110
[45] Alalshekmubarak, A., Smith, L.S. (2014). On improving the classification capability of reservoir computing for Arabic speech recognition. In: Wermter, S., ICANN 2014. LNCS, 8681: 225-232. https://doi.org/10.1007/978-3-319-11179- 7_29
[46] Salamon, J., Jacoby, C., Bello, J.P. (2014). A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1041-1044. https://doi.org/10.1145/2647868.2655045
[47] Fu, S.W., Wang, T.W., Tsao, Y., Lu, X., Kawai, H. (2018). End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Transactions on Audio Speech and Language Processing, 26(9): 1570-1584. https://doi.org/10.1109/taslp.2018.2821903