Ministry of Higher Education and Scientific Research
وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University

جامعة باجي مختار –
عنابـــــــــة

Faculty of Technology

كلية التكنولوجيا

Department of Computer Science

قســـــم الاعلام الآلي

Thesis

Presented to obtain the diploma of

Doctorate

Specialty: Information and Communication Sciences and Technologies

Field: Computer Science

By:

DENDANI Bilal

Title:

Design and Implementation of a Ubiquitous Framework for Pronunciation Learning

Thesis defended on January 19th, 2022 in front of the jury:

N°	Last name and first name	Grade	Institution	Quality
01	Khadir Mohamed Tarek	Prof.	Badji Mokhtar Annaba University	Chairman
02	Bahi Halima	Prof.	Badji Mokhtar Annaba University	Supervisor
03	Sari Toufik	Prof.	Badji Mokhtar Annaba University	Co-supervisor
04	Mohamed Ben Ali Yamina	Prof.	Badji Mokhtar Annaba University	Examiner
05	Maazouzi Faiz	MCA.	Mohamed-Cherif Messaadia Souk Ahras University	Examiner

In memory of my mother “Zeyneb”

Acknowledgements

I would like to thank my advisor, Halima Bahi, for her encouragement and support over the past five years. During these years, my advisor has given me great freedom and followed me to pursue my research which I am extremely interested in. I am extremely grateful for her guidance, mentoring, and support. Her academic vision and humility made me appreciate her. I thank her for the opportunity that she gave me to be mentored by her during this period. Thank you for taking such an interest in my work and making time to meet with me regularly.

I am also very grateful to my co-advisor, Sari Toufik, who helped me numerous times with discussing ideas and suggestions related to computer vision which can be applied also to speech processing.

Five years passed in the blink of an eye. Thanks to my colleague's family in the LABGED lab and the computer science department who supported me.

My gratitude is extended to Prof. Mohamed Tarek Khadir, for accepting to be the chairman of my thesis committee. Many thanks also to Prof. Yamina Mohamed Ben Ali and Dr. Faiz Maazouzi for kindly accepting to examine and review this thesis.

Last but not least, I am thankful to my father and his wife, my grandmother, brothers, sisters, and all my family members. Big thanks to my wife who encouraged me at every stage of my thesis. Big thanks for her help to keep care of my three children Zeyneb, Maria, and Houdaifa, especially during the hardship of my thesis. I would like also to thank all my friends for always encouraging me.

Table of Contents

Publications List

Dendani, B., Bahi, H., Sari, T. (2021). Self-Supervised Speech Enhancement for Arabic Speech Recognition in Real-World Environments. *Traitement Du Signal*, *38*(2), 349–358. https://doi.org/10.18280/ts.380212

Dendani B., Bahi H., Sari T. (2020) Speech Enhancement Based on Deep AutoEncoder for Remote Arabic Speech Recognition. In Image and Signal Processing, ICISP 2020. Lecture Notes in Computer Science, vol 12119. Springer, Cham. https://doi.org/10.1007/978-3-030-51935-3_24

Dendani, B., Bahi, H., Sari, T. (2019). A ubiquitous application for Arabic speech recognition, In International Conference on Artificial Intelligence and Information Technology(ICA2IT'19), 281–284. https://dspace.univ-ouargla.dz/jspui/bitstream/123456789/20831/1/Bilal%20Dendani.pdf

Dendani, B., Bahi, H., Sari, T. (2018). A Network-based Speech Recognition Application of Ubiquitous Computing for the Arabic Language, Journées d'Etude sur l'Intelligence Artificielle et ses Applications (JEIAA'2018), Dec. 2018

الملخص:

زَادت الْأَجْهِزَة الْمَحْمُولَة وَالْقَابِلَة لِلْارْتِدَاء بِشَكْل كَبِير مِنْ اِسْتِخْدَام الْوَاجِهَات الَّتِي تَدْعَمْ الْكَلَام، كَمَا عَزَّزَت الْاِسْتِخْدَام الْوَاسِع النِّطَاق لِتَطْبِيقَات التَّعَلُّم فِي كُلّ مَكان، وَالَّتِي تَهْدِف إِلَى الْوُصُول إِلَيْهَا مِنْ أَيّ مَكان وَفِي أَيّ وَقْت. عَلَى وَجْه الْخُصُوصِ، شَهِدَت تَطْبِيقَات تَعَلَّم اللُّغَة بِمُسَاعَدَة الكمبيوتر نموا كبيرًا. هُنَا يُعَدُّ تَعَلُّم النُّطْق مُهِمَّة صَعْبَة فِي الْبِيئَات فِي كُلّ مَكان. فِي الْوَاقِـع، تكُون إِشَارَات الْكَلَام عُرْضَةً لِلتَّلَف مِنْ قِبَلِ عِدَّة مَصَادِر، مِثْل ضَوْضَاء الْخَلْفِيَّة أَوْ أَخْطَاء التشفير أَوْ اِضْطِرَاب الْقَنَاة. حِينَهَا يَجِب اِسْتِعَادَة الْخِطَاب الْأَصْلِيّ مِنْ النُّسْخَة التَّالِفَة لِتَقْيِيمه بِشَكْل مَوْثُوق. لِهَذَا الْغَرَض، مِنْ الْمُفْتَرَض أَنْ تَتَوَفَّر الْعَدِيد مِنْ عَينَات الْكَلَام فِي الْعَالَم الْحَقِيقِيّ. مِنْ نَاحِيَة أُخْرَى، فَإِنَّ مُهِمَّة تَقْيِيم النُّطْق هِي الْمُكَوّن الْأَسَاسِيّ لِأَيّ نِظَام تَعَلُّم النُّطْق بِمُسَاعَدَة الْكَمْبِيُوتِرِ (CAPL)، إِذْ يُوَفّر مُلَاحظَات مَوْثُوقَة لِلطُّلَّاب لِتَحْسِين تَدْرِيبهِمْ. مِثْل هَذِهِ التَّطْبِيقَات تَتَطَلَّب تَوَفُّر بَيَانَات الْكَلَام غَيْر الْأَصْلِيَّة المشروحة وَالْمُصَنَّفَة. وَمَعَ ذَلِكَ لَا تَتَوَفَّر مِثْلُ هَذِهِ الْبَيَانَات فِي مُعْظَم الأحيان، خَاصَّةً بِالنِّسْبَة لِلُغَات مُنْخَفِضَة الْمَوَارِد مِثْلُ الْعَرَبِيَّة. تَهْدِف هَذِهِ الْأُطْرُوحَة إِلَى تَطْبِيق نِظَام تَعَلُّم النُّطْق بِاللُّغَة الْعَرَبِيَّة فِي بِيئَة مُنْتَشِرَة فِي كُلَّ مَكَان، فِي ظِلّ نُدْرَة مَجْمُوعَات الْبَيَانَات الْمُتَخَصِّصَة. وبالتالي فَإِنَّ مُسَاهَمَة هَذِهِ الْأُطْرُوحَة ذَاتُ شِقَّيْنِ.

فِي حَالَة عَدَم وُجُود مَجْمُوعَة بَيَانَات مُخَصَّصَة، تَمّ اِعْتِمَاد النَّهْج غَيْر الْمُوَجَّه لِأَدَاء تَحْسِين الْكَلَام، يَتَكَوَّن مِنْ خَطْوَتَيْن. أَوَلا، يَتِمّ تَدْرِيب بَرْنَامَج التشفير التِّلْقَائِيّ الْعَمِيق (OAE) عَلَى أَزْوَاج بَيَانَات مُزْعِجَة / مُزْعِجَة لِإِنْتَاج كَلَام مُحسن. بَعْدَ ذَلِكَ يَتِمّ تَدْرِيب جِهَاز تَقْلِيل الضَّوْضَاء التِّلْقَائِيّ الْعَمِيق بِطَرِيقَة خَاضِعَة لِلْإِشْرَاف لِلْاِسْتِفَادَة مِنْ الْمَرْحَلَة السَّابِقَة، والَّتِي تُقَدّر الْإِصْدَارَات الْمُحسّنَة مِنْ الْإِصْدَارَات الْأَصْلِيَّة. أَظهَرَت النَّتَائِج الَّتِي تَمّ الْحُصُولِ عَلَيْهَا تَحَسُّنا فِي معدل خطأ الكلمة (WER) بِحَوَالِي 4.48٪ لِمَجْمُوعَة الْبَيَانَات الْعَرَبِيَّة الْمُتَنَقّلَة (Mobile Arabic corpus). عِلَاوَةً عَلَى ذَلِكَ، تَمّ تَحْقِيق تَحَسُّن كَبِير فِي جَوْدَة الْكَلَام وَوُضُوحِه بِمِقْدَار 0.835 و0.06 عَلَى التوالي.

تَهْدِف الْمُسَاهَمَة الثَّانِيَة إِلَى التَّغَلُّب عَلَى نُدْرَة التَّدْرِيب غَيْر الْأَصْلِيّ عَلَى النُّطْق بِمُسَاعَدَة الْحَاسوب (CAPT)، الْمُخَصَّص لِلنُّطْق بِاللُّغَة الْعَرَبِيَّة. مستوحاة مِنْ نَجَاح التَّعَلُّم الْعَمِيق، نَقْتَرِح الْكَشْف عَنْ النُّطق الْغَيْر صَحِيح بِطَرِيقَة غَيْر مُوَجَّهه بِاِسْتِخْدَام خوارزميتين لِلتَعَلُّم الْعَمِيق تَمّ تَدْرِيبهُمَا عَلَى النُّطْق الصَّحِيح فَقَطْ. أَثْبَتَت النَّتَائِج التَّجْرِيبِيَّة عَلَى مَجْمُوعَتَيْنِ عَرَبِيَّتَيْنِ قُدْرَة الْمَنْهَج الْمُقْتَرَح عَلَى التَّمْيِيز بَيْنَ النُّطْق الْجَيّد وَالسَّيِّء. كَمَا أَكَّدت التَّجَارِب الْإِضَافِيَة الَّتِي اِسْتَفَادَت مَنْ تِقْنِيَّات زِيَادَة الصَّوْت لِتَوْسِيع مَجْمُوعَة بَيَانَات التَّدْرِيب كَفَاءة الطَّرِيقَة الْمُقْتَرَحَة.

الكلمات المفتاحية: النُّطْق بِمُسَاعَدَة الْحَاسُوب، تقييم النطق، اللغة العربية، التعرف على الكلام، تحسين الكلام، التعلم غير الموجه، نهج الكشف عن التشوهات، التعلم العميق.

Abstract:

Handheld and wearable devices have exponentially increased the usage of speech-enabled interfaces and promoted the widespread use of ubiquitous learning applications that aim to be accessible from anywhere and at any time. In particular, computer-assisted language learning (CALL) applications witnessed high growth. Herein, pronunciation learning is a challenging task in ubiquitous environments. Indeed, speech signals are prone to be corrupted by several sources, such as background noises, coding errors, or channel disturbance. The original speech should be recovered from the corrupted version to assess it reliably. For that purpose, many real-world speech samples are available. On the other hand, the pronunciation assessment task is the core component of any computer-assisted pronunciation learning (CAPL) system since it provides reliable feedback for students to improve their training. Such applications require the availability of annotated and rated nonnative speech data. However, most of the time, such corpora are not available, especially for low resource languages such as Arabic. This thesis aims to develop an Arabic pronunciation learning system in a ubiquitous environment under the scarcity of dedicated corpora. Thus, the contribution of this thesis is twofold.

In the absence of dedicated corpus, an unsupervised approach is adopted to perform the speech enhancement; it consists of two steps. First, an overcomplete deep autoencoder (OAE) is trained with noisy/noisy pairs to produce enhanced speech. Next, a denoising deep autoencoder is trained in a supervised way leveraging the previous stage. The obtained results showed an improvement of the word error rate (WER) of about 4.48% for a mobile Arabic corpus. Moreover, a significant improvement was achieved for speech quality and intelligibility by 0.835 and 0.06, respectively.

The second contribution aims to overcome the scarcity of nonnative computer-assisted pronunciation training (CAPT) dedicated Arabic speech corpora. Inspired by the success of deep learning, we propose to detect abnormal pronunciation in an unsupervised manner using two deep learning algorithms trained on solely correct pronunciations. Experimental results on two Arabic corpora proved the potential of the proposed approach to distinguish between good and bad pronunciations. Additional experiments leveraging audio augmentation techniques to expand the training dataset confirmed the efficiency of the proposed method.

Keywords: CAPT, Pronunciation assessment, Arabic language, speech recognition, speech enhancement, unsupervised learning, anomaly detection approach, deep learning.

Résumé :

Les appareils portables ont augmenté de façon exponentielle l'utilisation des interfaces vocales et ont favorisé l’expansion des applications d'apprentissage ubiquitaires. En particulier, les applications d'apprentissage des langues assisté par ordinateur (CALL) ont connu une forte croissance. Ici, l'apprentissage de la prononciation est une tâche difficile dans des environnements ubiquitaires. En effet, les signaux vocaux sont susceptibles d'être corrompus par plusieurs sources, telles que des bruits de fond, des erreurs de codage ou des perturbations de canal. La parole prononcée doit être récupéré à partir de la version corrompue pour l'évaluer de manière fiable. D’autre part, la tâche d’évaluation de la prononciation est l’élément central de tout système d’apprentissage de la prononciation assisté par ordinateur (CAPL), car elle fournit une rétroaction fiable aux étudiants pour améliorer leur formation. De telles applications nécessitent la disponibilité de données vocales non natives annotées. Cependant, la plupart du temps, de tels corpus ne sont pas disponibles, en particulier pour les langues à faibles ressources comme l'Arabe. Cette thèse vise à mettre en œuvre un système d'apprentissage de la prononciation arabe dans un environnement ubiquitaire sous la rareté des corpus dédiés. Ainsi, l'apport de cette thèse est double.

En l'absence de corpus dédié, une approche non supervisée est adoptée pour effectuer l'amélioration de la parole ; elle se compose de deux étapes. Tout d'abord, un auto-encodeur profond (OAE) sur complet est entraîné avec des paires bruitées/bruitées pour produire une parole améliorée. Ensuite, un auto-encodeur profond de débruitage est entraîné de manière supervisée en tirant parti de l'étape précédente qui estime les versions améliorées des versions originales. Les résultats obtenus ont montré une amélioration du taux d'erreur sur les mots (WER) d'environ 4,48% pour un corpus arabe mobile. De plus, une amélioration significative a été obtenue pour la qualité de la parole et l'intelligibilité de 0,835 et 0,06, respectivement.

La deuxième contribution vise à pallier la rareté des corpus non natifs de parole dédiés au CAPT. Inspirés par le succès de l'apprentissage en profondeur, nous proposons de détecter les prononciations anormales de manière non supervisée à l'aide de deux algorithmes de l'apprentissage en profondeur, entraînés uniquement sur des prononciations correctes. Les résultats expérimentaux sur deux corpus arabes prouvent le potentiel de l'approche proposée pour distinguer les bonnes et les mauvaises prononciations. Des expériences supplémentaires utilisant des techniques d'augmentation audio pour étendre l'ensemble de données d'entraînement ont confirmé l'efficacité de la méthode proposée.

Mots clés : Prononciation assistée par ordinateur, Évaluation de la prononciation, langue arabe, reconnaissance vocale, amélioration de la parole, apprentissage non supervisé, approche de détection d'anomalies, apprentissage profond.

List of Figures

List of Tables

Glossary

AM	Acoustic Model
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BD-LSTM	Bi-Directional Long Short-Term Memory
CA	Classical Arabic
CALL	Computer-Assisted Language Learning
CAPT	Computer-Assisted Pronunciation Training
CDAE	Convolutional Deep Auto Encoder
CNN	Convolutional Neural Network
CRN	Convolution Recurrent Network
CRNN	Convolution Recurrent Neural Network
DAE	Deep Auto Encoder
DDAE	Denoising Deep Auto Encoder
DNN	Deep Neural Network
DTW	Dynamic Time Warping
DL	Deep Learning
ERN	Extended Recognition Network
GLL	Global Local Likelihood Score
FCN	Fully Convolution Network
FuSPA	Fuzzy logic-based System for Pronunciation Assessment
GAN	Generative Adversarial Network
GLL	Global Log Likelihood score

GMM	Gaussian Mixture Model
GOP	Goodness of Pronunciation
GP	Gaussian Posteriorgrams
GSM	Global Standard for Mobile
HMM	Hidden Markov Model
Kbps	Kilobits per second
L1	native language
L2	second language
LPC	Linear Predictive Coding
LPS	Log Power Spectrum
LSD	log-spectral-distortion
LSTM-RNN	Long Short-Term Memory-Recurrent Neural Network
MDD	Mispronunciation Detection and Diagnosis
MFCC	Mel-Frequency Cepstral Coefficient
MFP	Mel frequency power spectrum
MLP	Multilayer Perceptron
MMSE	Minimum Mean-Squared Error
MOS	Mean Opinion Score
MSA	Modern Standard Arabic
MSE	Mean-Squared Error
NLP	Natural Language Processing
NN	Neural Network
OAE	Over Complete Auto Encoder
PCM	Pulse-Code Modulation

PDA	Personal Digital Assistant
PESQ	Perceptual Evaluation of Speech Quality
ReLU	Rectified Linear Unit
RER	Reference-free Error Rate
SCT	Spoken Chinese Test
SEGAN	Speech Enhancement Generative Adversarial Network
SFFT	Short Fast Fourier Transform
SGD	Stochastic Gradient Descent
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
SVM	Support Vector Machine
TDS	Time Duration Score
T-F	Time frequency
UAE	Under Complete Auto Encoder
VoIP	Voice over IP
WER	Word Error Rate

Chapter 1: Introduction

1.1 Problem Description and Motivation

The world has become a global village; “global village” is a term, introduced by Herbert Marshall McLuhan, that refers to the widespread use of information and communication technologies leading to a highly interconnected world. Being fluent in multiple languages is a necessity in such a world and is a sine qua non-condition to be productive and effective in an interdisciplinary context. Indeed, the global village tends to discard written communications and promote speech as the main means of communication. Thus, learning new languages becomes a necessity if one wants to take advantage of this opportunity and interact with the rest of the world. In this context, the availability of mobile devices and the easy access to the Internet have favored the emergence of numerous applications relating to Computer-Assisted Language Learning (CALL); however, it should be noted that although speech appears to be the simplest and most natural means of communication, systems that deal with this modality are rare. So, often, pronunciation learning is not included in the language learning systems.

Lately, advances made in Automatic Speech Recognition (ASR) technology sparkled the re-emergence of the research in Computer-Assisted Pronunciation Training (CAPT). Herein, CAPT is mainly used as an assistive tool that helps learners practice speaking by providing automatic pronunciation evaluation and corrective feedback (Eskenazi, 2009; O’Brien et al., 2019). For a long time, ASR technology based on Hidden Markov Models (HMMs) was the key to automated pronunciation assessment. Herein, the incoming speech is decoded as a sequence of phonemes; the recognition stage produces a collection of scores at the phoneme’s level and the word’s level as well, representing the similarity between the expected speech (to pronounce) and the recognized one. Although, this approach requires the availability of a great amount of labeled data.

At the same time, extensive use of handheld and wearable devices exponentially increased the use of speech as communication means and favored the widespread use of ubiquitous systems that aim to be accessible anywhere and at all times. Therefore, the speech signal is prone to be corrupted by several sources, such as coding errors, background noises, or channel disturbance. To overcome these limitations, in absence of real-world speech corpora for Arabic, we propose an unsupervised speech enhancement method based on the use of Deep Neural Networks (DNNs). The experiments carried on two Arabic speech corpora aim to show the effectiveness of the proposed approach to enhance the speech signal that will be assessed.

Once the incoming speech is enhanced, the assessment stage decides whether the incoming pronunciation is correct or wrong. In the absence of a labeled and rated corpus for the Arabic, we suggest the use of an anomaly detection approach. The proposed model is solely trained on correct samples to detect wrong pronunciations during the test stage. We also leverage the properties of deep architectures to discover complex relationships among the given data, as the detection system was based on a Deep Learning (DL) model. Indeed, for the assessment stage, two DL architectures, the Deep Auto Encoder (DAE) and the Fully Convolution Network (FCN) were trained in an unsupervised way using solely correct pronunciations. Moreover, data augmentation techniques were adopted to artificially expand the limited available datasets.

In summary, this work deals with pronunciation evaluation in a ubiquitous environment. Under the constraint of the scarcity of resources for the Arabic language. In particular, the lack of corpus dedicated to both speech recognition in real-world environments and pronunciation assessment led us to adopt an unsupervised way to model the enhancement system and the anomaly detection approach to build the assessment system.

1.2 Main Contributions

This thesis aims to contribute to the development of a ubiquitous framework for Arabic pronunciation assessment. Figure 1.1 depicts conceptual elements of the ubiquitous pronunciation learning system (any device, anywhere, any time, and any context).

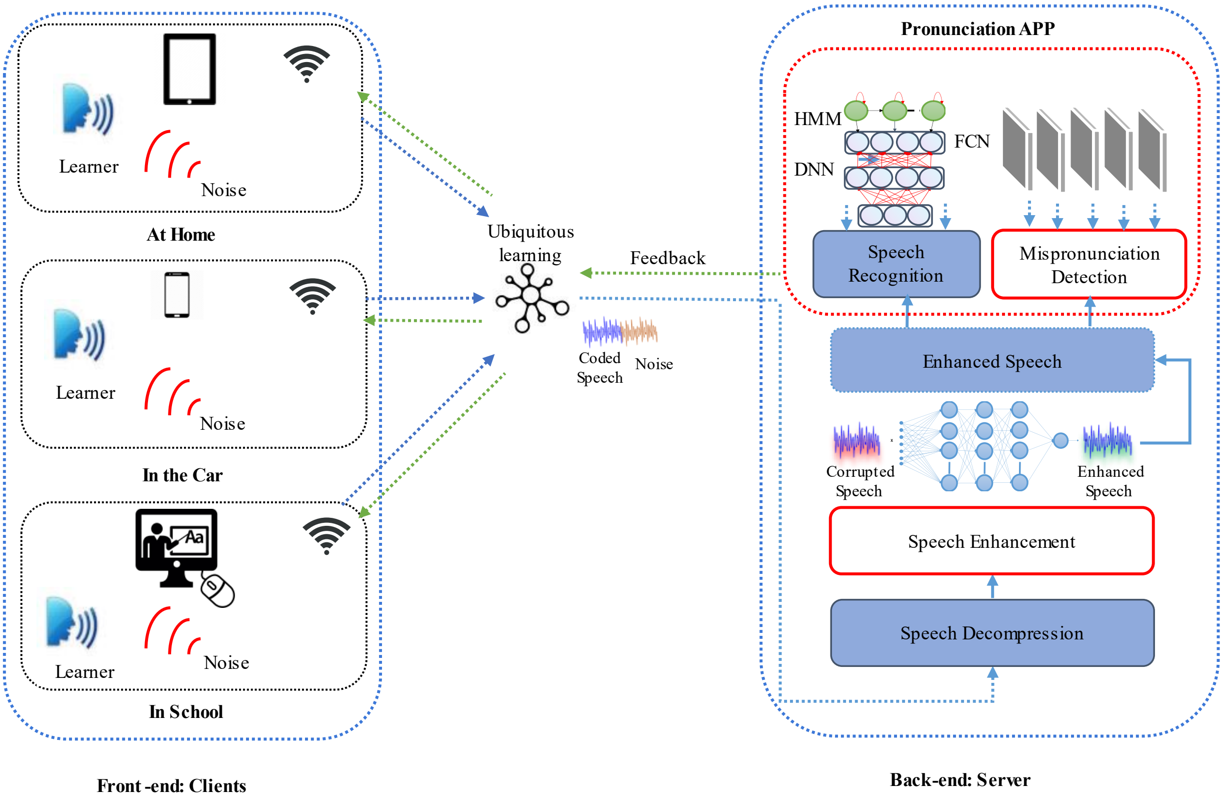


Figure 1.1 Overview of the ubiquitous proposed framework

As already said, the present thesis deals with the Arabic pronunciation assessment in real-world conditions. The main faced issue was the lack of resources for the Arabic language, including real-world speech signals and CAPT dedicated speech samples. Thus, the thesis brings two main contributions that can be summarized as follows:

- **A deep auto-encoder for unsupervised speech enhancement**. We proposed a two-step approach where an overcomplete deep autoencoder is trained in an unsupervised way using noisy/noisy pairs to produce the enhanced speech, then a denoising deep autoencoder is trained in a supervised way leveraging the previous step that produces clean versions of the speech. The obtained results show an improvement of the Word Error Rate (WER) of about 4.48% for a mobile Arabic corpus, which makes the proposed approach an effective alternative to the implementation of robust ubiquitous speech recognition systems. The enhancement system also 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.

- **Mispronunciation detection in noisy environments using deep neural networks.** We investigated the effectiveness of deep learning architectures trained in an unsupervised way to detect deviant pronunciations. Thus, two DL models (DAE and FCN) were trained solely on correct pronunciations and they are intended to detect deviant ones during the test stage as mispronunciations. The experimental results, on two Arabic corpora, proved the potential of the proposed approach to distinguish between good and bad pronunciations. Additional experiments leveraging audio augmentation techniques to expand the training dataset confirmed the efficiency of the proposed approach and allowed its improvement.

1.3 Thesis Outline

The thesis is structured into five chapters. The organization of each chapter is as follows:

Chapter 1 is concerned with the introduction and the aims of the research. It provides an overview of the research and outlines the structure of the thesis.

Chapter 2 ties together the background and the literature review of the research. It briefly presents the need for speech recognition and the main difficulties involved in the ubiquitous context. Chapter 2 is divided into three distinct sections. The first section deals with speech recognition, as it provides an overview of the ASR and its different steps. It introduces particularly hidden Markov models (HMM) and different ASR architectures. The second section addresses the speech coding aspect and presents the various categories of the codecs. The last section introduces the speech enhancement principles and provides a literature review of the existing methods. This section ends the chapter with a summary of the main findings related to DNN-based SE research.

Chapter 3 introduces the computer-assisted pronunciation teaching systems. It particularly reviews research approaches in pronunciation assessment. Additionally, it highlights some well-known corpora, and underlines limitations related to the lack of dedicated corpora, particularly for Arabic speech. We give a brief literature review of Arabic pronunciation assessment findings, after describing the language particularities. Finally, the chapter summarizes the main findings and different approaches pertinent to Arabic pronunciation learning.

Chapter 4 details the first contribution of this thesis: a self-supervised approach for speech enhancement. This approach is motivated by the lack of dedicated corpora for Arabic speech processing. First, the chapter presents a supervised speech enhancement model, then the main contribution is presented. This study consists of supervised and self-supervised speech enhancement methods for Arabic speech recognition in ubiquitous environments under challenging real-world conditions.

Chapter 5 proposes the second contribution related to the pronunciation assessment for CAPT systems. The proposition was also motivated by the lack of dedicated corpora for Arabic pronunciation assessment. The chapter presents the principal elements and the obtained results.

Chapter 6 concludes the thesis with a summary of the contributions and discusses possible future works for researchers and developers interested in teaching pronunciation in mobile environments.

Chapter 2:

Speech Recognition for Pronunciation Assessment in Real-World Environments

2.1 Introduction

Automatic speech recognition is the core component of state-of-the-art computer-assisted pronunciation teaching systems. Indeed, "Most of the pronunciation assessment methods are based on local features derived from automatic speech recognition" (Cheng et al., 2020, p.1). This chapter presents the different parts involved in automatic speech recognition in a ubiquitous environment. First, an overview of ASR is presented, I introduce the hidden Markov models (HMMs) that consist of the state-of-the-art models in speech recognition and pronunciation assessment as well. Then, I present several architectures involved in the context of mobile speech recognition. The several architectures are compared in the context of the suggested implementation, herein, a ubiquitous speech recognition system for pronunciation learning. Meanwhile, server-based models for mobile speech recognition are concerned with speech coding that refers to the speech signal representation in a digital form with few bits. Therefore, enabling the speech signal transmission while preserving the quality is required for further applications, such as the pronunciation assessment. Finally, ubiquitous real-world speech environments are often contaminated by background noise which deteriorates speech quality, intelligibility and decreases the ASR performance. Consequently, a speech enhancement stage is required in such scenarios. Figure 2.1 depicts the successive modules, at either the client-side or the server-side, to build a ubiquitous speech recognition system.

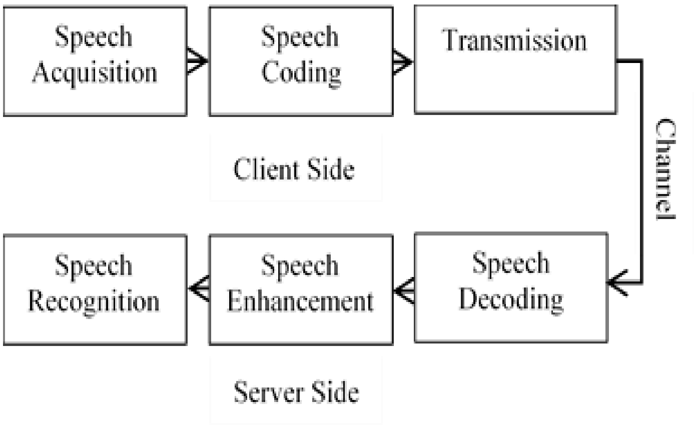


Figure 2.1 Block diagram for the proposed ubiquitous speech recognition system

2.2 Automatic Speech Recognition (ASR)

2.2.1 ASR basics

An ASR system converts the captured speech signals to a sequence of words. The typical architecture of a recognition system is shown in Figure 2.2. The incoming speech signal is acquired via a microphone then the obtained digital vector is transformed into a set of acoustic vectors. This step is called the feature extraction stage. The obtained vectors represent the observation denoted by Y = y_1, y_2,…, y_T where each y_i represents an acoustic vector. The second module of the recognition system allows the transcription of the acoustic vector series into a series of lexical units. The lexical units are words, namely, W = w1, w2,…, wL, where the sequence of words of length L is more likely to be generated by observation Y; this is the recognition stage. The decoder seeks to estimate the sequence of words W that maximizes the probability of observing the sequence Y.

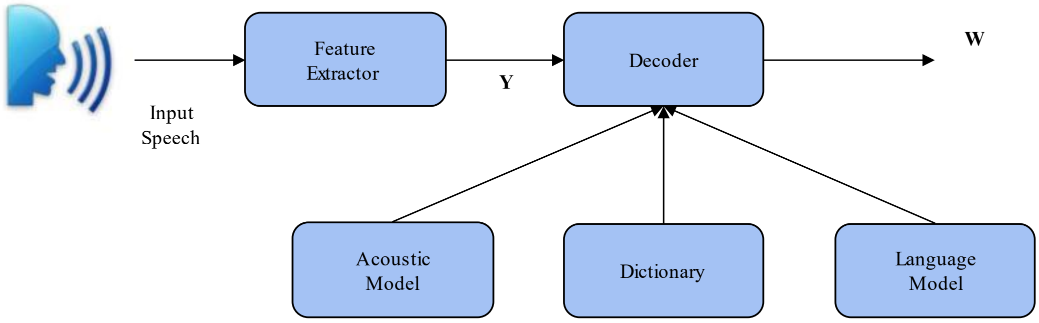


Figure 2.2 Block diagram for a speech recognition system

The state-of-the-art ASR systems are based on statistical acoustic models, the Hidden Markov Models (HMM) (Rabiner, 1989), used in the recognition stage.

2.2.2 ASR components

As depicted in figure 2.2, a typical ASR system performs speech recognition through the following steps:

- Capturing the speech signal.
- Extracting the features vector from the spoken utterances, this stage is known as the front-end part.
- Decoding and recognizing the speech using a speech engine, based on the acoustic models, phonetic dictionary or lexicon, and the language model. This stands for the back-end part where the algorithms of decoding are used.

According to (Schmitt et al., 2008, p. 66), only “2 % of all processing time [is dedicated for feature extraction] in case of medium-sized vocabulary and even less for large vocabulary recognition tasks”, whereas the ASR search takes the most computational resources.

2.2.2.1 Feature extraction

Feature extraction is the first important phase that serves to represent the pattern to recognize into a compact and representative form. It transforms the speech signal into a set of feature vectors to reduce the speech signal variability. Various feature extraction methods are used in speech recognition technology. The most widely used are the spectral-based features achieved using cepstral analysis. Particularly, the Mel Frequency Cepstral Coefficients (MFCCs) are among the most used features in speech recognition for several applications.

Another feature that can be obtained from the spectral analysis is the linear prediction coefficients (LPC), and their various transformations, such as the Linear Predictive Cepstral Coefficients (LPCC) as well as the coefficients resulting from filter bank analysis.

2.2.2.2 Hidden Markov models

The HMM model is a technique used in speech recognition technology to model a pattern as a sequence of states (called observations). Transitions between states are represented by edges (in a graph representation). The edges are labeled with values; a value represents the transition probability from a state to another. Meanwhile, each state is provided with a probability function characterized by a probability distribution function.

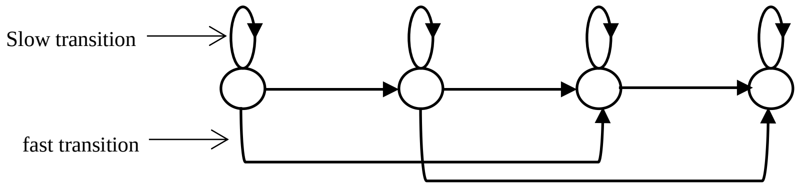


Figure 2.3 An HMM representing a word

When an HMM is applied to speech recognition, the states are interpreted as acoustic models, indicating what sounds are likely to be heard during their corresponding segment of speech; while the transitions provide temporal constraints, indicating how the states may follow each other in sequence (Bahi & Sellami, 2005).

The feature vectors are extracted from the speech signal to represent the observation data Y. The word sequence $\hat{W}$ is obtained through the Bayesian decision rule:

$$\hat{W} = \arg\max P(W|Y) = \arg\max P(W)^{*}\, P(Y|W) \qquad (2.1)$$

The P(W) is the prior probability of observing some specified word sequence and is given by the language model (LM).

P(Y|W) is the likelihood probability of observing the speech data Y given the word sequence W. It is determined by the HMM acoustic model (AM).

2.3 Mobile Speech Recognition

Today, various techniques of ASR can be used for the design, implementation, and deployment of speech recognition components over the networks or at the hand-held devices. Three different approaches are used, namely: Embedded Speech Recognition (ESR), Network Speech Recognition (NSR), and Distributed Speech Recognition (DSR). These techniques differ from each other in the distribution of the ASR components. The following is the description of each architecture with the pros and cons for each one.

2.3.1 Embedded Speech Recognition (ESR)

The embedded speech recognition systems are also known as client-based systems where both ASR components: feature extraction (front-end) and speech recognition (back-end) are embedded into the client-side, without the need for an external component at the server-side, as depicted in figure 2.4. The ESR is frequently the model adopted for hand-held devices such as personal digital assistants (PDA). The main advantage of the ESR architecture is that it is free from the data transmission quality, and hence no latency for data processing.

An important issue for the ESR architectures is the lack of resources when processing computationally demanding ASR applications. Therefore, remote-based models (NSR and DSR) make the resources more available.

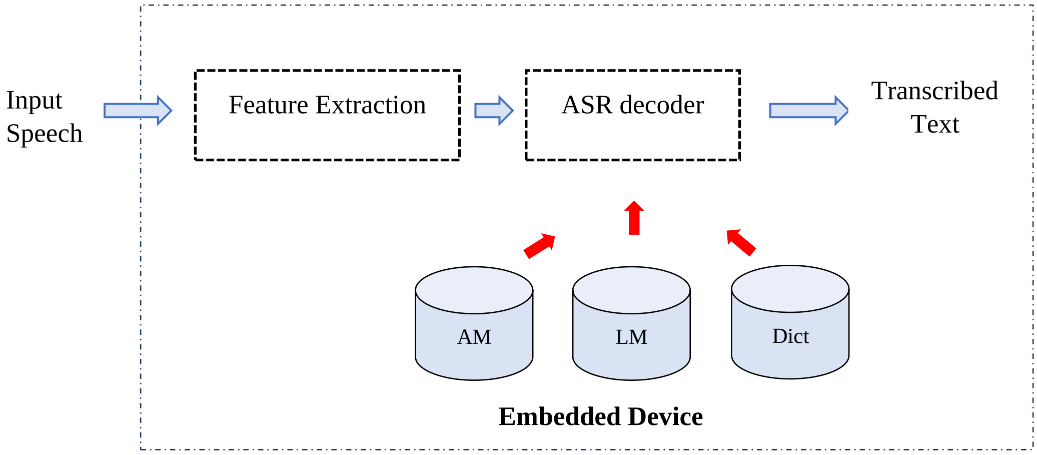


Figure 2.4 Client-based ASR (ESR)

2.3.2 Network Speech Recognition (NSR)

Network speech recognition is characterized by the location of both feature extraction and ASR search at the server-side, while the speech signal is captured at the client-side as shown in Figure 2.5. Sometimes, NSR is referred to as a cloud speech recognition system. The NSR architecture imposes no restriction relied on the resource's limitations. No need for increased resources on the client side because the central server is responsible for the feature extraction and decoding processes.

The main advantage of the NSR model is the simplicity to update the ASR systems on the server-side. Moreover, it enables the plug and play of the ASR system on the server-side without changes on the client devices. However, this approach has the disadvantage of degrading the recognition performance when using low-bit-rate codecs for speech encoding. This degradation becomes more severe when data transmission errors occur, and in the presence of noise background conditions (Schmitt et al., 2008).

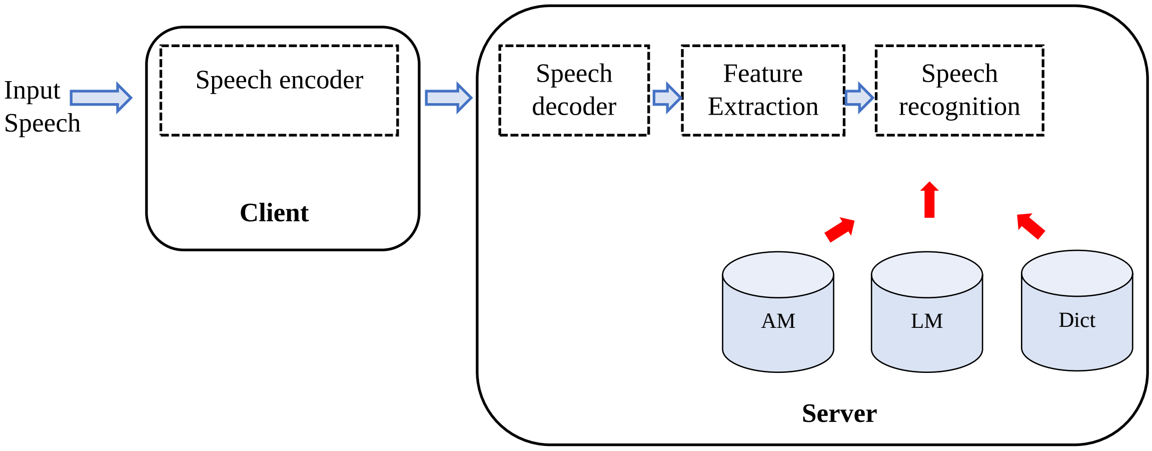


Figure 2.5 The block diagram for network speech recognition model (NSR)

2.3.3 Distributed Speech Recognition (DSR)

Distributed speech recognition is a client-server-based architecture that extracts the features at the client-side, whereas the ASR search is undertaken in the remote server, as shown in Figure 2.6. The DSR is a more recent approach, undergoes similar benefits as NSR, and there is no data loss by encoding speech when transmitting data likewise NSR (Schmitt et al., 2008). The advantage of DSR is to spread the charge across the client and the server.

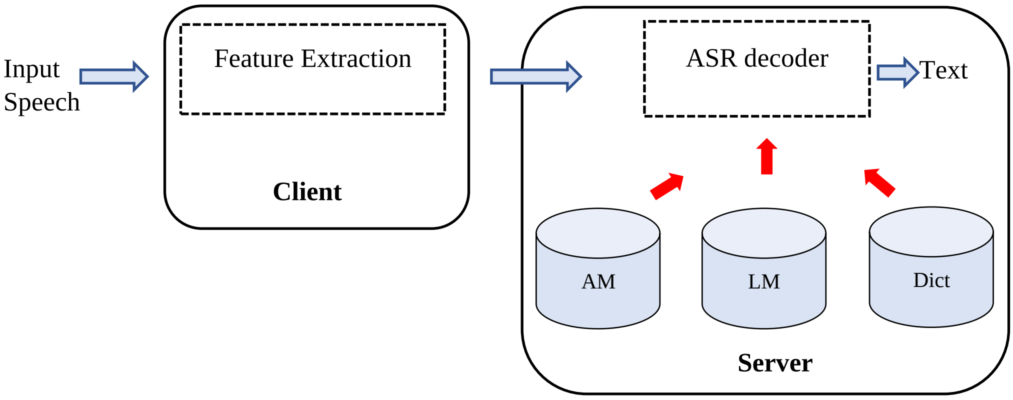


Figure 2.6 The block diagram for distributed speech recognition model (DSR)

Speech signal quality, noise, and channel robustness are important parameters for preferring DSR, while the wide deployment of high-quality speech coders makes NSR a better alternative.

In particular, for the application of pronunciation assessment, the incoming signal needs a huge number of computational resources to extract information about how the speech was uttered.

2.4 Speech Coding

Computer applications that deal with speech need the speech signal to be digitized. Speech coding aims to represent the continuous waveform into numerical form. The digitized signal may be compressed with or without loss of information; however, this transformation has to preserve the characteristics of the original signal according to further applications. An ideal speech coder represents the input speech in a few bits as possible without quality degradation. Nevertheless, there is a trade-off between the codec bit rate and the quality of the transmitted voice.

2.4.1 Speech coding

Speech coding or compression refers to the process of representing the speech signal in a digital form with few bits while preserving the quality and the intelligibility for further applications. The main goal of speech coding is to minimize the bit rate as possible and to maintain the quality, intelligibility of the transmitted speech, by removing the redundancies. Many applications are concerned, including Voice over IP (VoIP) networks and automatic speech recognition technology with various applications such as healthcare, multimedia information retrieval, and educational language learning. The present work is about pronunciation learning in a ubiquitous context, herein, the ASR accuracy highly depends on the quality of the transmitted coded/decoded speech signal. The speech quality is subjectively estimated using the mean opinion score (MOS) measure.

A wide range of speech codecs is available, ranging from uncompressed, lossy, and loss-less compressed codecs, such as PCM, G.711, FLAC, MELP, GSM, MP3, etc. Many investigations about speech compression and recent speech coding technologies are presented in (Chu, 2003; Sinder et al., 2015; Gibson, 2016). Others studied the effects of codecs on ASR performance (Ramana et al., 2012; Raghavan et al., 2017; Sun et al., 2013). Particularly, (Raghavan et al., 2017) show the distortion effects on the performance of the speech recognition system. After comparison between some codecs, the study concluded that the narrowband high bit rate codec’s G.711 provides the best performance with five acoustic modeling techniques.

2.4.2 Basics of speech coding

The conversion from continuous speech signal to digitized form implies three successive processes: sampling, quantization, and coding. Sampling is the process of converting the continuous signal to a discrete sequence of points by extracting every period T a value s; T is known as the sampling rate. Quantization refers to the conversion from the continuous amplitude signal to the discrete amplitude signal. The coding is the conversion of the discrete amplitude signal resulting from the quantization process to a bit-stream (set of bits).

In figure 2.7, the amplitude space is divided into 16 levels of quantization. Four bits can be used to represent the quantization results. The horizontal lines represent the quantization phase while vertical lines represent the sampling phase. The continuous signal is represented by the red line while the blue points represent the discrete signal samples.

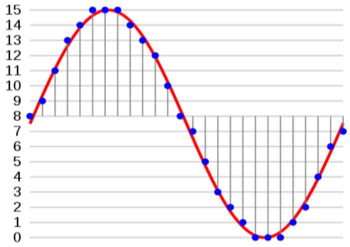


Figure 2.7 Sampling and quantization of a periodic signal

To select the best-fitted speech coding method, some factors could be considered. Mainly, three properties are considered: low bit rate, high speech quality, and high robustness.

The bit rate is the measure of speech storage for an audio encoded signal along with one unit of time; it is expressed by kilo-bits per second (kbps). A low bit rate means fewer bits for encoding input speech data and less bandwidth for audio speech transmission. The speech quality refers to the measure which indicates the closeness between the original signal and the reproduced signal. Moreover, the performance of the speech codec is evaluated based on the robustness against channel errors, packet loss, low delay, and low computational complexity.

Speech coders may be classified based on their bit rate attribute; they are divided into high and low bit rate algorithms. Table 2.1 describes these types.

Table 2.1 Classification of speech coders based on the bit rate

Class	Bit-Rate range (kbps)
High bit-rate	> 15
Medium bit-rate	2 – 15
Low bit-rate	2 – 5
Very Low bit-rate	< 2

There are trade-offs between the previous parameters where a good performance of one property implies a lower performance of the other one. For example, the internet codec (iLBC) (Anderson et al., 2004) provides the advantage of high robustness against the packet loss, whereas the value of the bit rate is increased.

2.4.3 Speech coding techniques: a comparative study

Various techniques and methods that perform speech coding are known; they mainly belong to three categories: waveform, parametric and hybrid techniques. The waveform speech coding methods aim to regenerate the speech signal at the decoder side as closely as possible to the original speech signal. The most known waveform speech coding systems are PCM (Pulse Code Modulation) and the ADPCM (Adaptive Differential PCM). Parametric speech coding or vocoder methods consider a model to generate the speech signal using some parameters. The technique cannot preserve the quality of the input waveform speech signal. The most known example of parametric coders is the LPC (Linear Prediction code). The combination between waveform and parametric codec results in the hybrid-based speech coding technique which takes the advantages provided by both codecs. It behaves as waveform speech coders on the decoder side to save the speech signal as closely as possible to the original speech signal. Moreover, it operates like the parametric speech technique on speech production during the encoding phase. It is noteworthy that CELP (Code Excited Linear Prediction) is the most known technique for hybrid-based codecs. The majority of modern speech codes are based on the CELP technique. The following section details some characteristics of these techniques.

2.4.3.1 Waveform codec

Waveform speech coding methods reconstruct the speech signal and preserve as possible the form of input speech signal regardless of the nature of input speech provided by the speaker. A waveform codec is characterized by a low complexity and a good speech quality using a high bit rate. Waveform-based speech compression techniques reduce the amount of information by decreasing redundancies. The bit rate ranges between 16 kbps and 64 kbps. If the data rate is below 16 kbps, the speech quality will degrade. The coder is sensible to channel errors with a high bit rate. The simplest form of waveform coding is the PCM which involves sampling and quantization.

Pulse Code Modulation (PCM)

PCM is a method for digitizing analog signals, considered as the standard form for other speech/audio formats. Many applications and usage include landline telephone, digital audio, compact discs. PCM is the simplest technique and a non-compressed method, used as a reference for comparison between speech/audio codecs (Spanias, 1994). The description of PCM is presented in (ITU-T G.711, 1988). Two PCM types exist μ-law and A-law. μ-law is standardized for usage in North America and Japan, and A-law for usage in Europe and other countries in the world. The size of the sample for μ-law and A-law is 8 bits. If the sampling rate is 8KHZ, the bit rate for the PCM is 64 kbps. ITU standardized G.711 for μ-law and A-law codecs (ITU-T G.711, 1988).

Adaptive Differential Pulse Code Modulation (ADPCM)

Adaptive Pulse Code Modulation (ADPCM) is a method resulting from the Pulse Code Modulation (PCM), where the ADPCM sample is coded with 5; 4; 3 or 2 bits rather than 8 bits. Consequently, the bit rate is reduced, and the storage is minimized. ADPCM is used in video conferencing applications. ITU standardized G.726 for ADPCM codec (ITU-T G.726, 1990). The G.726 bit rate varies from 16, 24, 32, or 40 kbps corresponding to the sample size of 2, 3, 4, or 5 bits, respectively. The bit rate of 40 kbps presents a better quality for speech/audio than other bitrates 32, 24, or 16 kbps. Indeed, the higher is the selected bit rate, the higher the speech/audio corresponding quality.

2.4.3.2 Parametric coder (vocoder)

Parametric-based coding methods are based on speech production where only parameters are sent from one side to another. At the receiver side, the speech is regenerated using the parameters of the model. Parametric codecs are characterized by a high compression rate that exceeds the compression rate of waveform methods. The resulting bit rate is lower, and the quality is not comparable to that of waveform methods. The original speech signal is not preserved, except the spectral and some statistical properties of the signal when encoding it. The bit rate varies in the range of 2 to 5 kbps. Linear Prediction Coding (LPC) is the most known example of vocoders, with a bite rate ranging from 1.2 kbps to 4.8kbps. The codec was firstly presented by Atal in 1971 (Atal & Hanauer, 1971) to remove redundancy in the signal. The lower the bit rate, the lower the speech/audio quality, however, the codec provides an intelligible speech, a high compression rate, and a robotic sound (Sun et al., 2013). It is used in speech analysis and synthesis.

2.4.3.3 Hybrid coder

Hybrid coders are produced by combining waveform methods and vocoders. The parametric coders are used for encoding speech, and the waveforms are used in decoding speech. One of the known hybrid coding techniques is the CELP coder (Code-Excited Linear Prediction) which most modern codecs are based on. Examples of CELP-based speech codecs include AMR, SPEEX, G.728, G.729, and cellular telephony GSM codecs. The CELP algorithm is considered the basis of other algorithms.

Code-Excited Linear Prediction (CELP)

The CELP hybrid method combines the advantages of waveform and parametric coding techniques. The coder has been proposed by Schroeder and Atal (Schroeder & Atal, 1985). They have demonstrated that the coder provides a high speech quality at a lower bit rate of 4.8 kbps. The CELP coding system is based on vector quantization (VQ) of LPC coefficients. It uses the analysis by synthesis technique which means combining parameters for the objective of preserving the analysis signal as the reconstructed signal (Huang et al., 2001). It predicts the residual of the current frame using the periodicity provided by the residual of voiced speech (Huang et al., 2001).

2.4.4 Speech codecs classification

Speech codecs can be classified according to the attributes: speech quality, degree of complexity, bit rate, and bandwidth. These codecs are divided into high and low bit rates following the bit rate attribute. The high bit rate occurs when the value is higher than 15 kbps. Under 2 kbps bit rate, the system is categorized as a very low bit rate. Medium bitrate ranges between 5 and 15 kbps. For lower bit-rate, values are distributed between 2 and 5 kbps.

In another hand, speech/audio compression techniques are classified according to the bandwidth into the narrowband (NB), wideband (WB), and full band (FB). Narrowband speech coding methods operate from 0 to 4 kHz, some applications include digital telephony conversations where the bandwidth ranges from 300 Hz to 3.4 kHz. Wideband speech compression provides a better speech quality and expands the bandwidth between 0 and 7 kHz. Today, a wideband-based speech codec finds its use in considerable applications such as VoIP.

In the super wideband speech (SWS), the frequency is limited between 50 Hz and 14 kHz. SWB is adopted in video streaming. The full-band (FB) speech compression operates at 20 Hz to 20 kHz, the same frequencies as human voices. Table 2.2 summarizes different speech coding methods with respect of the bandwidth.

Table 2.2 Summary of NB, WB, SWB and FB for speech/audio coding (Sun et al., 2013b)

Mode	**Signal bandwidth (HZ)**	**Sampling rate (KHZ)**	**Bit-Rate (kb/s)**	**Examples**
Narrowband (NB)	300-3400	8	2.4-64	G.711, G.729, AMR
Wideband (WB)	50-7000	16	6.6-96	G.711.1, G.722
Super-wideband (SWB)	50-14000	32	24-48	G.722.1
Full band (FB)	20-20000	48	32-128	G.719

2.4.5 Comparison between speech codecs

Speech coding systems can be categorized into two main methods, uncompressed and compressed techniques. Uncompressed audios are like PCM or wave audio forms where no compression is applied for audio/speech. However, the compressed technique uses some compression algorithms to compact the audio data. They are divided into lossy and lossless algorithms. The lossy mode discards information. Therefore, the original speech signal differs from the reproduced one with a deteriorated quality of speech, whilst the lossless preserves the original speech signal to be as close as possible to the reconstructed speech data.

Speech codecs are divided according to the used algorithms, such as Pulse Code Modulation (PCM), Adaptive Differential Pulse Code Modulation (ADPCM), Linear Predictive Coding (LPC), and others are based on the Code Excited Linear Predictive Coding (CELP). Table 2.3 provides a comparison between the most known codecs based on the attributes (bit rate, BW, algorithm, and source). Google, the pioneer of cloud-based speech recognition systems supports a large list of encoding audio systems including uncompressed, lossless, and lossy compressed systems. It supports LINEAR 16 (RAW or WAV), FLAC, MULAW, AMR, AMR-WB, Ogg Opus, Speex to represent the speech audio data to be transmitted (Google Cloud Speech API).

Table 2.3 Comparison between audio/speech codecs

Speech/Audio coding mode	**Codec**	**Codec Type**	**Bit rate kbps**	**Algorithm**	**BW**	**Source and Standardization**	**Features**
Uncompressed	PCM	Waveform	128	/	NB	SOX[1]	Provides good performance, used when no restriction about internet bandwidth.
Lossless Compression	FLAC (Coalson, 2001)	/	935	Linear prediction		Xiph.org Foundation[2]	Fastest and most widely supported loss less audio codec.
	G.711 (ITU-T G.711, 1988)	Waveform	64	Log PCM	NB	ITU-T	Better speech quality, versions are μ-law and A-law
	G.722 (ITU-T G.722, 2012)	Waveform	64, 56, 48	Subband ADPCM	WB	ITU-T	
	G.726 (ITU-T G.726, 1990)	Waveform	40, 32, 24, 16	ADPCM		ITU-T	Used in submarine cables
	G.728 (ITU-T G.728, 1992)	Hybrid	16	CELP	NB	ITU-T	Used for voice streaming, teleconferencing
	G.729 (ITU-T G.729, 1996)	Hybrid	8	ACELP	NB	ITU-T	Used in streaming through internet

[1] http://sox.sourceforge.net/

[2] https://xiph.org/

Lossy Compression	MELP (McCree and Barnwell, 1995)	parametric	2.4	/	NB	USDOD	
	GSM-FR (ETSI GSM-FR, 1998)	Hybrid	13	Regular Pulse Excitation-Long Term Prediction (RPE-LTP)		ETSI	Global standard for mobile telecommunications (GSM)
	GSM-HR (ETSI GSM-HR, 2000)	Hybrid	5.6	Vector-Sum Excited Linear Prediction (VSELP)		ETSI	Global Standard for Mobile telecommunications (GSM)
	AMR-NB (ETSI AMR, 2001)	Hybrid	4.75,5.15,5.9,6.7,7.4,7.95,10.2,12.2	Algebraic Code Excited Linear Prediction (ACELP)	NB	ETSI	3rd generation mobile telephony
	AMR-WB (ETSI AMR, 2001)	Hybrid	6.6,8.85,12.65,14.25,15.85,18.25,19.85,23.05,23.85	Algebraic Code Excited Linear Prediction (ACELP)	WB	ETSI	3rd generation mobile telephony
	Speex (Valin, 2016)	Hybrid	2.15 – 24.6	CELP	WB	Xiph.org Foundation	Optimized for speech and a low latency communication
	Ogg Vorbis (Moffitt, 2001)	Hybrid	48 – 500	Modified discrete cosine transform MDCT	FB	Xiph.org Foundation	Non-proprietary, patent free and alternative to MP3
	Opus (Valin et al., 2012)	Hybrid	6 – 510	SILK and CELT	NB, WB, SWB, FB	Xiph.org Foundation	Storage use and streaming applications

MP3 (Brandenburg, 1999)	Hybrid	8 – 320	MDCT	FB	Moving picture Experts Group	
AAC (Brandenburg, 1999)	Hybrid	16 – 320	MDCT	FB	Moving picture Experts Group	
ILBC (Anderson et al., 2004)	Hybrid	13.33 – 15.2	BI-LPC	NB	Global IP Solutions	Applied in audio streaming and VOIP

2.5 Speech Enhancement (SE)

Speech enhancement is an active and fascinating area of research that aims to improve the perceptual speech quality, and intelligibility of a corrupted noisy speech signal (Loizou, 2013a). The process of SE or speech denoising consists of removing noise from the speech signals. It targets various applications, such as automatic speech recognition, hearing aids, and speech communication. Several approaches deal with different types of noises. These approaches have been investigated over the past few years. SE methods belong to two main groups: traditional or signal processing methods, and data-driven-based methods. Both methods are based on a single channel or multi-channel input (Benesty et al., 2005).

2.5.1 SE objective metrics

Speech enhancement aims to improve speech quality and intelligibility for better human perception. The performance of SE algorithms can be evaluated either with subjective evaluation or objective metrics (Loizou, 2013b). In subjective evaluation, experts or human raters are asked to evaluate the speech quality by comparing the original clean speech with the enhanced one. The objective evaluation is performed by measuring the distance between the original speech and the processed one.

The most commonly used metrics for SE evaluation are the intelligibility and the quality of speech signals. To evaluate these two metrics, perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) and short-time objective intelligibility (STOI) (Taal et al., 2011) have been proposed. The PESQ score reflects the value of the mean opinion score (MOS) ranging between -0.5 and 4.5. On the other hand, the intelligibility score (STOI), ranging from 0 to 1, reflects the correlation between the original speech and the speech to evaluate. Both scores are obtained by comparing the processed speech to the reference original one. The higher the values, the better speech quality, and intelligibility.

When the SE preprocessing is related to the ASR applications, the word error rate (WER) metric is used to evaluate the performance of the ASR system. Herein, the main goal is to reduce the WER to improve the noise-robustness, by comparing the transcription results of the speech recognizer to the reference transcription. The observed errors belong to three categories: insertion (I), deletions (D), and substitutions (S). the WER score is computed using the ratio between all the errors and the total number of spoken words (N).

$$WER = \frac{I + D + S}{N} \qquad (2.2)$$

Other objective metrics can evaluate the processed speech following other aspects, such as the frequency or the time domain. The log-spectral distortion (LSD) estimates the log spectrogram difference between the enhanced and the clean original speech. The segmental signal-to-noise ratio (SSNR in dB) measures the difference in terms of the time domain and computes the rate between the processed speech and the noisy one. Other measures include speech distortion (SD) and noise reduction (NR) (Lu et al., 2013). CSIG is the MOS predictor of the signal distortion, and CBAK reflects the MOS predictor of the background noise intrusiveness (Pascual et al., 2017).

2.5.2 State-of-the-art in speech enhancement

Various conventional speech enhancement (SE) approaches called spectral restoration are proposed. Distinguished examples include spectral subtraction, minimum mean square error (MMSE)-based spectral amplitude estimator, Wiener filter, and non-negative matrix factorization (NMF). These traditional approaches are based on statistical models and are often applied when the noise is stationary. Recently, deep neural networks acoustic models replaced the Gaussian Mixture Models (GMM) acoustic models, and have been improved the performance of ASR systems (Hinton et al., 2012). Meanwhile, the success of deep learning techniques has extended towards SE and noise suppression. The following section discusses some techniques for each category.

2.5.2.1 Traditional approaches

2.5.2.1.1 Spectral subtraction

One of the first proposed approaches for SE is the spectral-based approach used to suppress stationary noise from the speech signal (Boll, 1979). Methods based on spectral subtraction algorithms aim to suppress and reduce the spectral acoustic effects of an additive noise from the speech signal, this is done by estimating the noise spectrum speech pauses, and it is subtracted from the noisy speech spectrum to estimate the clean speech. As the noise varies randomly, the spectral subtraction leads to the presence of processing distortion. To overcome the resulted distortion, some algorithms were developed.

2.5.2.1.2 Minimum Mean Square Error (MMSE)

In signal processing, the minimum mean square error (MMSE) is an estimation method that minimizes the mean square error (MSE) between the initial and the reconstructed signals (Ephraim and Malah, 1984). It is a common measure of estimating quality for the fitted values of a dependent variable.

2.5.2.1.3 Wiener filtering

This technique assumes that if noise is present in the system, then it is considered to be an additive white Gaussian noise (AWGN) (Lim and Oppenheim, 1979). The computation of the Wiener filter requires the assumption that the signal and noise processes are second-order stationary (in the random process sense). The main drawback of the Wiener filtering is that it requires a priori knowledge of the power spectra of the noise and the original signal.

Almost all of these traditional methods present a common limitation, as they are relevant to the additive stationary noise or the statistical properties of speech and noise signals. Consequently, they fail to handle the non-stationary noise of real-world environments where acoustic conditions are unexpected. Although these traditional techniques are effective for SE in stationary noise conditions, their ability to model the non-linear relationship between noisy and clean speech signals is very limited. Meanwhile, deep learning-based techniques demonstrated superior ability in non-linear relationships compared to the aforementioned approaches.

2.5.2.2 Data-driven approaches using deep learning techniques

A new era of SE has emerged when deep learning techniques have been introduced. The ability of deep learning models to approximate complicated functions, and to provide a strong regression operation makes them advantageous to be explored for many applications such as modeling the relationship between clean and noisy speech features. DL models have been used to estimate the clean speech from the noisy speech using spectral mapping domain (Xu et al., 2014) or via time-frequency mapping (Y. Wang et al., 2014). It is worth pointing that the deep neural network algorithms have become the state of the art for the SE task.

The DNNs are trained on a huge number of speech data pairs (noisy, clean) to learn the mapping between noise and clean pairs in a supervised way. The representation of the speech fed to these network models follows two feature types: frequency domain and time domain.

Many architectures are proposed for SE, through the state-of-the-art. Various studies are conducted in the frequency domain, while others are based on time-domain representation (end-to-end temporal mapping).

2.5.2.2.1 Frequency domain works

DL methods operate typically on Short Time Fourier Transform (STFT) to predict the clean speech signal from the noisy speech using the frequency domain. The STFT has been applied as a way to extract features from the speech and the noisy signals using data-driven DL techniques. To our knowledge, deep learning was first introduced to speech enhancement by Lu et al. in two conference papers (Lu et al., 2012; 2013). The first paper (Lu et al., 2012) used a deep autoencoder architecture, trained on clean speech pairs in an unsupervised way to recover the clean speech from the noisy version. Both input and the output are the Mel power spectrum, and the DAE learns to map the clean Mel spectrum input to the output one. In the second work (Lu et al., 2013), the researchers expand the training of the DAE to noisy-clean speech pairs in a supervised fashion. The model is trained based on the feature of the Mel frequency power spectrum (MFP) to map the noisy MFP input to the clean MFP output. The trained DAE is then used to enhance speech signals from noisy input. The performance results achieved by the DAE outperformed that of traditional SE using the MMSE. While the previous DAE learns to encode only statistical clean information, the later DAE learns to denoise speech from the noisy version.

A subsequent regression DNN-based SE approach in the frequency domain was proposed to estimate the mapping function between noisy and clean utterances (Xu et al., 2014, 2015). The DNN architecture was trained to map the log power spectrum of the noisy speech to that of the clean speech as depicted in Figure 2.8.

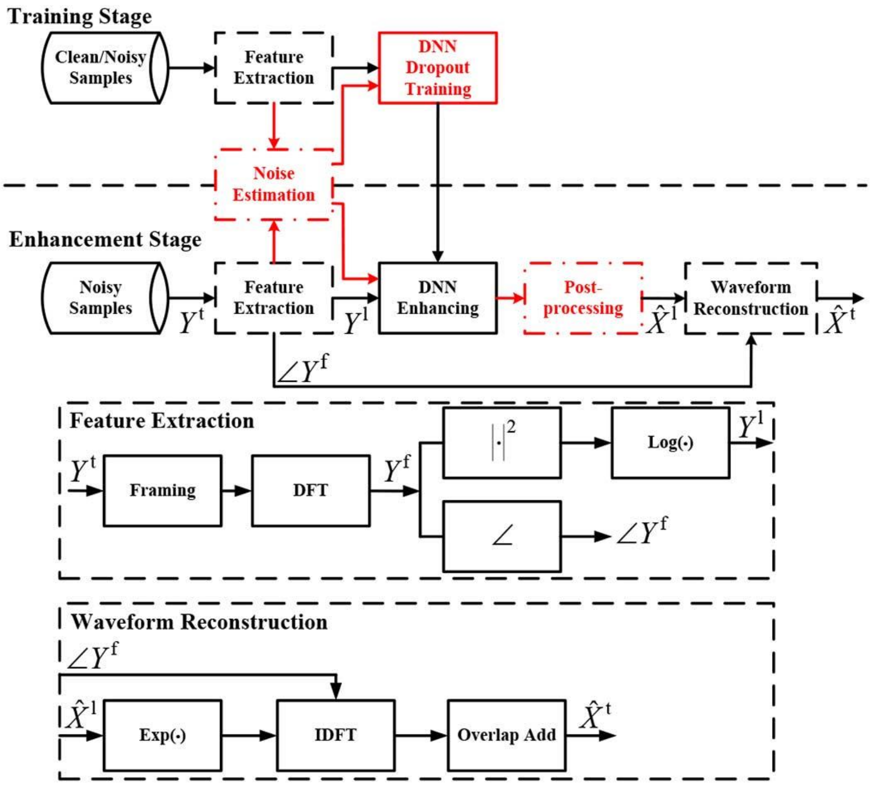


Figure 2.8 Speech enhancement using DNN based on spectral mapping (after, Xu et al., 2015)

The DNN model was trained on a large data set of 100 hours with multiple noise conditions. The obtained results show that the performance of SE using DNN outperformed the traditional based MMSE approach, and provided better results in presence of non-stationary noise. Moreover, a feed-forward neural network model was investigated in (Kumar & Florencio, 2016) for SE in a real-world environment. The input of the network was the log-power spectra of the noisy speech, and the output was the log-power spectra of the clean speech.

Although DNN-based SE models achieved better performance than the traditional SE approaches, these fully connected-based models are computationally heavy and use many parameters to learn the regression function. In addition, the DNN-based models may not effectively represent the local temporal-spectral structure of the speech signals. Thus, the convolution neural network (CNN) models were designed to overcome these limitations and better represent speech in 2D structure. The CNN architecture processes data in local regions and reduces the model complexity when compared to the fully connected models. CNN algorithms are particularly suitable for image analysis and classification (Krizhevsky et al., 2012; LeCun et al., 1998; Shen et al., 2017). A typical CNN has convolutional layers interspersed with pooling layers, followed by fully connected layers as in a multilayer neural network (Lawrence et al., 1997; Shen et al., 2017).

Recently, the CNN model highly improved the speech recognition performance compared to the DNN architecture (Abdel-Hamid et al., 2014). Subsequently, it was investigated for the task of SE, in the frequency domain (S. W. Fu et al., 2016; 2017). This architecture demonstrated the ability to model the local temporal and spectral structures of the speech signals. In (S. W. Fu et al., 2016), researchers investigated the CNN model to restore clean speech from a noisy version using the SNR-aware algorithms. The SE task was based on the spectral mapping between the noisy log power spectrum (LPS) input and the clean LPS output. The performance achieved using CNN outperformed that of the DNN.

On the other hand, some recent studies have focused on the SE task based on the complex spectrograms (Williamson et al., 2016, S. W. Fu et al., 2017). It should be noted that the phase of information has been taken to improve the performance of SE (Williamson et al., 2016). In (S. W. Fu et al., 2017), researchers estimated the complex spectrogram of clean speech from that of the noisy speech using CNN as depicted in figure 2.9. The results achieved using complex spectrogram estimation outperformed that using magnitude spectrum with the DNN structure.

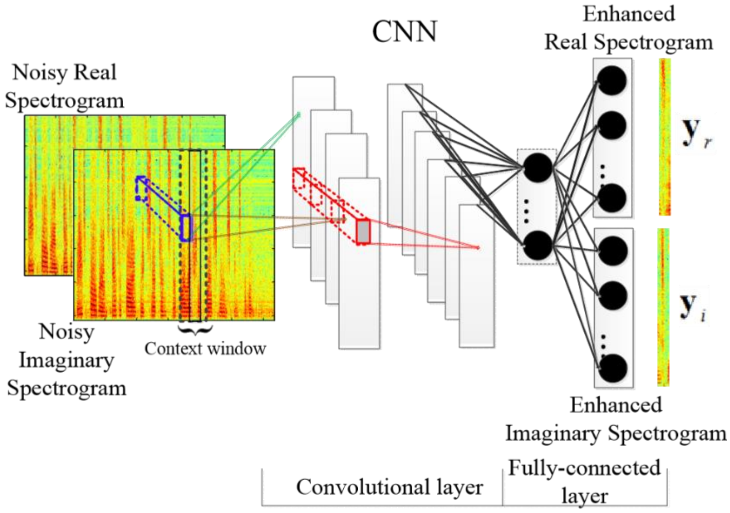


Figure 2.9 Speech enhancement using CNN based complex spectrogram (after, S. W. Fu et al., 2017)

Along the same line of spectral representation, various studies have disentangled the speech and the noisy signals from the noisy speech signals using other types of models. In (Maas et al., 2012; Weninger et al., 2015), the researchers used a recurrent neural network (RNN) based on the Long Short-Term Memory (LSTM) for speech enhancement. The RNN structure was applied to improve the robustness against the noise for the ASR system and achieved the best average WER 13.76% (Weninger et al., 2015). In (Park & Lee, 2016), the researchers proposed a fully convolution network (FCN) (a CNN without dense layers) for SE using the LPS of the signals as features. The performance results obtained using the FCN model can be similar or outperform the DNN and RNN structures. Moreover, it yields much fewer parameters compared to the DNN and RNN. Therefore, it is suitable for embedded systems. Another architecture, the deep convolution encoder-decoder network, has also been used for the enhancement of the coded speech in presence of background noise (Z. Zhao et al., 2018). This CNN-based topology is used for direct mapping in both cepstral and time domains. It is built using three-layer types: convolutional layers, max-pooling layers, and up-sampling layers. The performance results achieved with the proposed approach outperformed that of the baseline G.711 postprocessing in terms of PESQ for G.711, G.726, and AMR-WB speech codecs. In addition, it achieved the best LSD compared to the fully connected neural network (FCNN) architectures.

More recently, hybrid DL techniques have been used for SE. For instance, researchers in (H. Zhao et al., 2018) stacked convolutions and recurrent layers to build a convolution recurrent neural network (CRNN), termed as EHNET. The performance results of the CRNN architecture (EHNET) outperformed the DNN and RNN models in terms of the five used metrics (SNR, LSD, MSE, WER, and the PESQ). Meanwhile, Tan & Wang (2018) have proposed a SE approach based on convolution recurrent network (CRN), using magnitude spectrograms for real-time applications. Figure 2.10 depicted the structure of the CRN. The results of the proposed approach outperformed the results of LSTM based line structures in terms of PESQ and STOI metrics and have fewer trainable parameters.

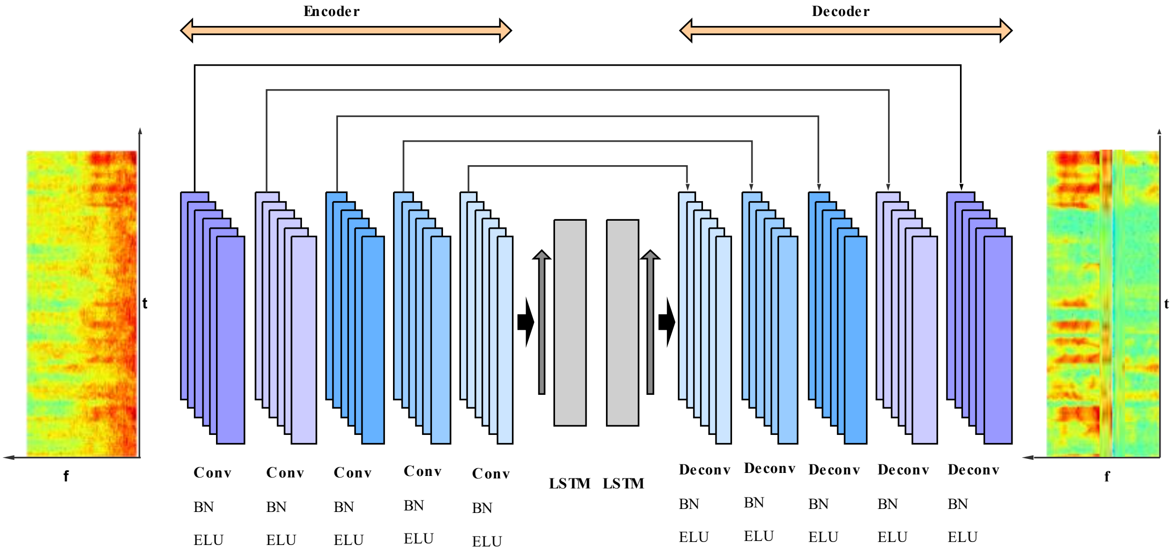


Figure 2.10 The architecture of the CRN (after, Tan & Wang, 2018)

Although the vast majority of studies have been conducted on the time-frequency (T-F) domain approaches for both speech separation and enhancement tasks (D. Wang & Chen, 2018), few have focused specifically on the time-domain approaches. Meanwhile, it is worth noting that the end-to-end speech separation approaches as well as CNN and GANs architectures represent the recent development in the field.

2.5.2.2.2 End-to-end time-domain works

Recently, deep learning techniques have been proposed for SE using temporal mapping approaches, in an end-to-end fashion. The input to the deep neural network models as well as the output is raw waveform representations, rather than the T-F format. The frequency-based approaches reviewed in the previous section, in particular, the fully connected-based architectures may not extract well the local information in the speech signal, to produce high-frequency components (S. W. Fu et al., 2017). Therefore, the researchers in (S. W. Fu et al., 2017) have proposed a fully convolution network (FCN) to model simultaneously the high and the low-frequency components of the raw waveform. The results of the FCN to reconstruct the waveforms show that the speech components are well preserved, while noise is removed effectively. In addition, the FCN model outperformed both models, DNN and CNN based on LPS and waveform inputs, respectively. In their subsequent study (S. W. Fu et al., 2018), they have proposed an utterance-based SE approach from raw waveform using the FCN architecture. The proposed approach processes speech even though with a variable length of inputs.

Along the same line of processing speech signals in the time domain, researchers in (Pascual et al., 2017) have developed a SE method based on the generative adversarial network (GAN). The GAN structure is an FCN architecture based on the AE structure. The encoding part contains the convolution layers whereas the decoding is based on the deconvolution layers. The performance results achieved using the GAN proposed approach termed as SE GAN(SEGAN) outperformed the classical Wiener method in terms of CSIG, CBAK, COVL, and Segmental SNR except for the PESQ metric. A subsequent study used CNN to enhance coded speech in the time domain (Z. Zhao et al., 2018) and provided remarkable performance results.

2.5.3 Comparison of different SE approaches in terms of feature type, model, and other characteristics

Table 2.4 summarizes the most important findings in the state-of-the-art of SE. It shows the comparison of DNN-based SE approaches in terms of feature type, noise type, DL architecture, and other characteristics such as the language and metrics for evaluation. A key component in these deep learning algorithms 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.

This SE section attempted to provide a summary of the literature relating to speech enhancement tasks using various deep learning techniques and following the supervised fashion. Although the majority of DL techniques applied for the SE task were supervised, there is a considerable need for unsupervised SE or self-supervised SE where clean data set are almost absent for low resource languages, and in real-world environments. In the fourth chapter, we propose a speech enhancement two-step approach before the recognition task. The proposed approach suggests the usage of the DAE architecture. An overcomplete DAE has been proposed for the first step, which was trained in an unsupervised fashion. In the second stage, a denoising DAE is trained in a supervised method, leveraging the clean speech recovered from the previous stage.

Table 2.4 Summary of different deep learning algorithms for speech enhancement

Reference	Feature Type	Language	Noise	Input / Output	DL algorithm	Measures
(Lu et al., 2012)	MFP	Japanese	White, Car, Factory, Babble	Clean/ Clean	DAE	Phone recognition accuracy
(Lu et al., 2013)	MFP	Japanese	Factory, Car	Noisy/ Clean	DDAE	NR, SD, PESQ
(Xu et al., 2014)	LPS	English	AWGN, Babbles, Car, Restaurant, Street	Noisy/ Clean	DNN	LSD, SegSNR, PESQ
(Xu et al., 2015)	LPS	English	Car, Crowd, Traffic	Noisy/ Clean	DNN	PESQ
(Weninger et al., 2015)	Magnitude spectrum	English	Home (children, TV, Radio)	Noisy/ Clean	LSTM-RNN	WER, SDR
(Kumar & Florencio, 2016)	LPS	English	Office environment (stationary and non-stationary)	Noisy/ Clean	DNN	PESQ, STOI, SD, NR
(S. W. Fu et al., 2016)	LPS	Mandarin	Babble, Car, Jackhammer, Pink, Street, WGN, Engine	Noisy / Clean	CNN	MSE, SegSNR
(Park & Lee, 2016)	LPS	English	Babble noise	Noisy / Clean	FCN, DNN, RNN	STOI, PESQ, SDR
(S. W. Fu et al., 2017)	Frame-wise waveform	English	Bable, Car, Jackhammer, Pink, Street, White Gaussian (WGN), engine, and baby cry noises.	Noisy/ Clean	FCN	PESQ, STOI

(Pascual et al., 2017)	Waveform	English	Babble, Domestic, Office, Public, Transportation, Nature, Street	Noisy/ Clean	GAN	PESQ, SegSNR CSIG, CBACK
(H. Zhao et al., 2018)	End-to-end based model	English	More than 25 types of noises	Noisy/ Clean	RNN, CNN	PESQ, SNR, WER, LSD, MSE
(Tan & Wang, 2018)	Magnitude spectrogram	English	10000 noises, Babble, Cafeteria	Noisy/ Clean	CRN	PESQ, STOI
(S. W. Fu et al., 2018)	Waveform based utterance-level	English, Mandarin	Babble, Car, Jackhammer, Pink, and Street	Noisy/Clean	FCN	PESQ, STOI
(Z. Zhao et al., 2018)	LPS, Time domain	English, German	Cafeteria, Car, Traffic Road, Coding	Noisy/ Clean	CNN	PESQ
(Dendani et al., 2020)	LPS	Arabic	G.711 Coding	Noisy/ Clean	DDAE	Accuracy

Chapter 3: Pronunciation Assessment Algorithms

3.1 Introduction

Computer-Assisted Language Learning systems enable students to learn languages on their own using interactive and individual lessons. In previous years, CALL systems were primarily based on Natural Language Processing (NLP) including components based on grammar and vocabulary. Fortunately, advances in automatic speech recognition have contributed to the development of computer assisted pronunciation teaching systems by enabling automatic pronunciation assessment. This assessment is often provided by feedback in the form of a measure or a score. Different measures have been proposed to quantitatively assess the quality of the learner's pronunciation or to measure speech proficiency.

CAPTs systems have been specifically designed to assess the pronunciation quality on one hand. In another hand, the second task: mispronunciation detection and diagnosis (MDD) consist of pinpointing out where occurs mispronunciation in an utterance and providing feedback to the language learner, this induces two main components of a CAPT system: an automatic speech recognition module and an evaluation module. Indeed, to assign a pronunciation score to the learner; a recognition step is required. The first step in an automatic speech recognition system is to extract the characteristics of the acoustic signal from speech. Then the evaluation stage begins.

Several studies have been focused on automatic pronunciation assessment and mispronunciation detection, which cover a variety of primary languages (L1) and second languages (L2). Many studies, targeted languages such as English, Chinese, Dutch, French, and Japanese. Meanwhile, Arabic is in the top five languages to learn for several considerations, and it remains a challenging low-resource language for both automatic pronunciation evaluation and MDD.

In this chapter, first, I present pronunciation assessment principles and review the approaches relevant to automatic pronunciation evaluation, mispronunciation detection, and diagnosis (MDD). Next, in the same context, an overview of DL algorithms for pronunciation assessment and MDD is given. Afterward, the performance metrics used in pronunciation assessment and a description of non-native speech corpora are presented in sections 3.4 and 3.5, respectively. Moreover, while the Arabic language has been popular, and many people around the world aim to learn Arabic pronunciation, its related pronunciation assessment remains challenging. Thus, I review techniques used in Arabic pronunciation assessment after presenting the particularities of spoken Arabic. Finally, a comparison between findings relevant to Arabic pronunciation assessment is addressed to highlight different opportunities.

3.2 Automatic Pronunciation Assessment

The pronunciation assessment process aims to automatically provide pronunciation evaluation that a human rater would produce to the students. For that purpose, speech recognition technology is mainly used.

Different blocks are involved in a pronunciation assessment process, ranging from the feature extraction module to the feedback correction. Figure 3.1 depicts an ASR-based pronunciation assessment block diagram; it shows two kinds of feedbacks: the pronunciation scoring is represented by the blue arrow and the error detection and diagnosis followed by the corrective feedback task, using the green arrows.

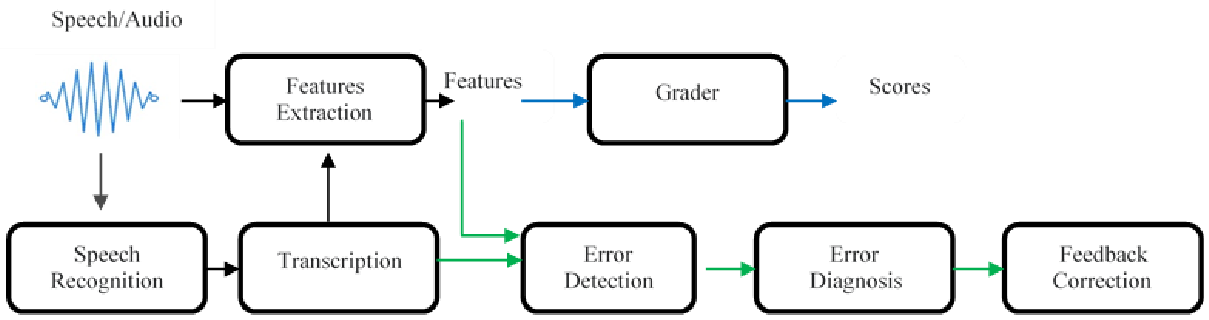


Figure 3.1 Block diagram of an ASR-based pronunciation assessment process and its different tasks

As already said, the speech recognizer is a key component in a CAPT system. A typical ASR-based CAPT system involves several stages. As depicted in figure 3.2, the CAPT system involves an ASR module to assess the learner's pronunciation or to give corrective feedback after pinpointing pronunciation errors. This typical architecture shows that the two main CAPT tasks are the pronunciation assessment and the mispronunciation detection followed by a corrective feedback process.

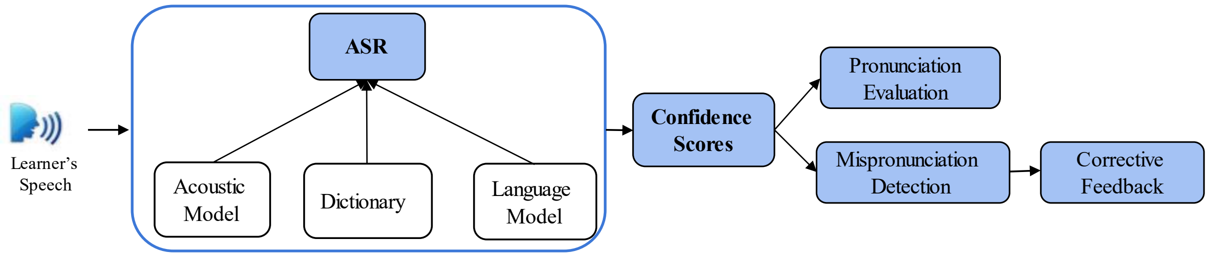


Figure 3.2 ASR-based CAPT system architecture

When the learner pronounces a given word (or another part of speech), acoustic observations are extracted from the incoming signal and are represented as a collection of acoustic vectors. Afterward, the system force aligns this representation with the model of the correct pronunciation (native-like). In the force-alignment stage, the speech recognizer computes the probability $p(W|O)$, where O is the observation represented by the extracted features from the incoming signal, and W is the model of the word to pronounce. In the following, we review the existing scores issued from this stage, with both HMM-based and deep learning approaches.

According to the review of research approaches in CAPT by (Chen & Li, 2017), the pronunciation assessment approaches were grouped under four main methods: likelihood-based scoring, classifier-based scoring, extended recognition network (ERN), and the unsupervised error discovery. A comparison between these approaches is described in table 3.1.

Table 3.1 Comparison of different pronunciation assessment approaches (Chen & Li, 2017)

Framework	**ASR-based**	**L1 Independence**	**L2 Independence**	**Error Detection**	**Error Diagnosis**
Likelihood-based Scoring (GOP)	✓	✓	✓	✓	
Classifier-based Scoring		maybe	maybe	✓	✓
Extended Recognition Network (ERN)	✓			✓	✓
Unsupervised Error Discovery		✓	✓	✓	✓

3.2.1 Confidence-based methods

Confidence scores or loglikelihood-based scores are conventional measures that focused on extracting features from HMM-based ASR (Bernstein et al., 1990; Franco et al., 1997; Neumeyer et al., 1996; Witt and Young, 2000). The ASR acoustic models have been a CAPT key component for a while.

One of the first attempts to evaluate pronunciation began when Bernstein et al. (1990) evaluated the pronunciation of Japanese students, reading English aloud, using an HMM-based ASR system at the sentence level. The ASR part has been prompted from a fixed text. Nevertheless, Neumeyer et al. (1996) suggested newly text-independent pronunciation assessment algorithms which were very close to the human expert ratings, with an arbitrary text. Research in (Franco et al., 1997; Neumeyer et al., 1996) explored different types of machine scores: HMM-based log-likelihood scores, phone duration scores, phone classification, and time-based scores. In particular, the logarithm of the likelihood of the speech data, computed by the Viterbi algorithm, is a good measure of the similarity between native speech and non-native speech (Neumeyer et al., 2000). For each sentence, the phone segmentation is obtained, along with the corresponding log-likelihood of each segment. Let t_i denote the start time of the ith phonetic segment, the total log-likelihood of this segment can be computed by the equation (3.1):

$$\mathrm{LL_i} = \sum_{\mathrm{t=t_i}}^{t_{i+1}-1} \log\left(\mathrm{p}(S_t|S_{t-1})\mathrm{p}(X_t|S_t)\right) \qquad (3.1)$$

Where X_t is the observed spectral vector and S_t the HMM state at time t, respectively, $\mathrm{p}(S_t|S_{t-1})$ is the HMM transition probability, and $\mathrm{p}(X_t|S_t)$ is the so-called output distribution of state S_t (Neumeyer et al., 2000).

Franco et al. (1997, 2000, 2010) proposed a phone posterior probability score that extended the HMM-based score. The phone posterior probability outperformed the likelihood and the normalized duration scores and provided the best correlation with the human scores. Moreover, the aforementioned scores have been combined, and correlate better with the human ratings. The acoustic-based models used for computing these machine scores, adapted to non-native speech data can outperform models trained only on native speech (Franco et al., 2010; Moustroufas & Digalakis, 2007). Moustroufas and Digalakis (2007) developed a system using native and non-native acoustic models that yield much better performance. The system evaluated the spontaneous English pronunciation of Greek students without knowing any target text at the sentence level.

The abovementioned confidence scores are compared to a threshold value to detect mispronunciations. Franco et al. (1999) proposed two mispronunciation methods at the phone level. The first was based on the posterior probability score where the model was trained on native speech. The second approach considered the log-likelihood ratio (LLR) which was trained on non-native speech and outperformed the posterior-based approach. Both computed scores were compared to a determined phone threshold to detect mispronunciations.

The most investigated score for CAPT purpose is the Goodness Of Pronunciation (GOP) score (Witt, 1999; Witt et al., 2000), which is derived from the log-likelihood score. The GOP algorithm calculates the ratio of the log-likelihood that is once the phoneme is spoken corresponds to the phoneme that really should be pronounced. When receiving the learner's speech, two recognition modes are used: the forced alignment mode forces the recognition of the speech to its known transcription, and the free recognition. The GOP score for a specific phoneme realization is determined by taking the difference between the log probability of the forced alignment, and the log probability of the free recognition. When a GOP score is calculated, a threshold value must be applied to reject phonemes that have been mispronounced. The final value depends on the proficiency required level.

While likelihood-based scores can assess pronunciation quality and detect errors as described in table 3.1, an extra error diagnosis could further provide feedback on the learner’s pronunciation, and pinpoint the type of error. Therefore, the following approaches are proposed.

3.2.2 Classifier-based methods

Probability-based strategies can recognize the pronunciation quality, yet these scoring computations are not enough to distinguish the nature of the error and the correct location of that error. Therefore, classification-based mispronunciation detection methods are used for this purpose and target confusion pairs of phonemes.

The algorithm described in (Ge et al. 2009) uses the log-posterior probabilities extracted when applying the force alignment with HMMs to classify syllable quality using Support Vector Machine (SVM) wide-margin separators. The classification over a large number of syllables produces a final score on a speaker's pronunciation proficiency. This score correlates with the Putonghua Shuiping Kaoshi ‘PSK’ corpus of scores which represents a corpus of

Chinese speakers of different dialects. Meanwhile, researchers in (Strik et al., 2009) studied classifier-based approaches to detect erroneous pronunciation made by Dutch L2 learners. Four classifiers (two acoustic phonetic-based classifiers, LDA-MFCC and the GOP) were investigated to detect confusion pairs.

Furthermore, Necibi et al. (2015) proposed a computer-assisted pronunciation teaching tool for young Algerian pupils to learn standard Arabic pronunciation. The system aims to decide whether the incoming pronunciation is “correct” or “incorrect” as well as to be able to separate pupils who have difficulties in pronunciation from those who have normal pronunciations.

First, the speech signal is captured, and a collection of acoustic features are computed (a set of MFCCs acoustic vectors). Then, the acoustic representation was transmitted to the speech recognition engine (ASR module) that compared it to the possible pronunciations of a given word stored in the database. This stage provides two scores; the global log-likelihood (GLL) and the time duration score (TDS). Finally, these two scores served as inputs to a decision tree classifier that accepts or rejects pronunciations (Figure 3.3).

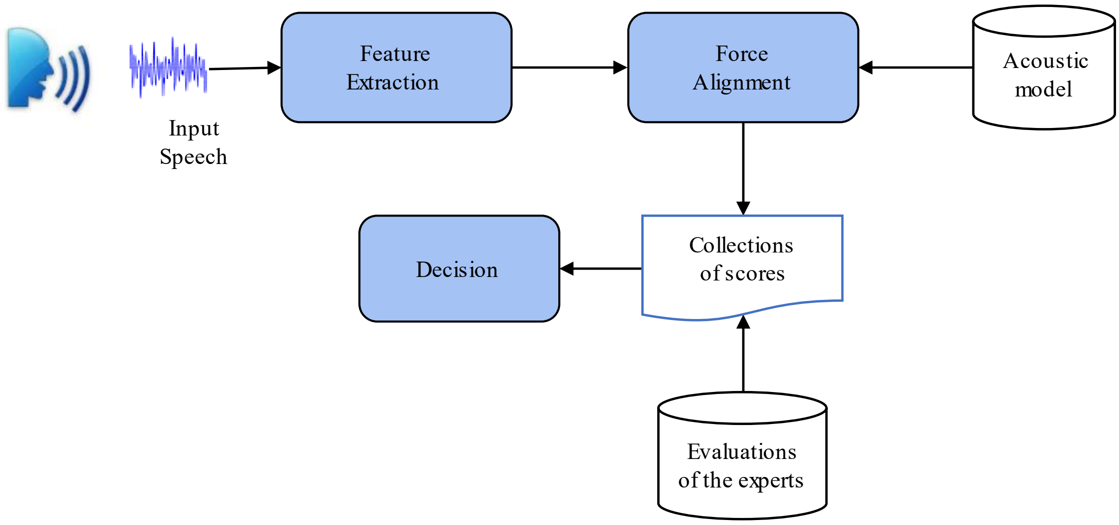


Figure 3.3 Overview of the decision based-classification system (after, Necibi et al., 2015)

While the classifier-based approaches can pinpoint and diagnose different types of errors, the mismatch between the trained acoustic model and the non-native learner’s speech can decrease the performance. Thus, the ERN approach can overcome this issue.

3.2.3 Extended Recognition Network (ERN)

Harrison et al. (2009) developed a powerful CAPT framework based on ASR to detect mispronunciations at the phone level. The framework called the extended recognition network (ERN) was proposed for Chinese learners of English and includes common phonetic mistakes of learners. A typical ERN-based MDD system includes an ASR module that transcribes the input speech of the learner. Afterward, a forced alignment step between the resulting transcription and the canonical pronunciations is performed to produce the final feedback. Figure 3.4 depicted a typical CAPT framework based on ERN.

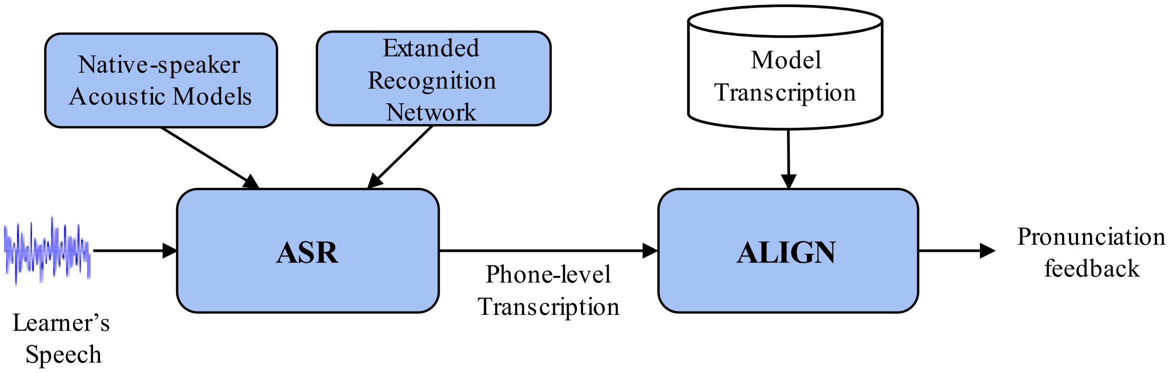

Native-speaker Acoustic Models
Extanded Recognition Network
Model Transcription
ASR
ALIGN
Pronunciation feedback
Phone-level Transcription
Learner's Speech

Figure 3.4 The different components of the CAPT system based on the ERN (after, Harrison et al., 2009)

Practically, an ERN is a finite state transducer (FST) representing phonological rules based on canonical pronunciations. Herein, the recognition network includes standard language pronunciation extended with the common mispronunciations of learners.

The ERN provides an automatic approach to identify the locations and types of phonetic errors in the pronunciation of second language learners. While the ERN detects and diagnoses errors, this approach depends on L1 and L2 as well. Meanwhile, the collection of phonetic errors is time-consuming. Therefore, the unsupervised error discovery approach was proposed to overcome these limitations.

3.2.4 Unsupervised error discovery approaches

Recently, unsupervised error discovery approaches have been adopted, to overcome the scarcity in non-native speech corpora that are needed to develop CAPT systems. Wang and Lee (2015) proposed an unsupervised approach to discover automatic error patterns directly from data, using phoneme posteriograms. Meanwhile, based on these findings (Lee & Glass, 2015; Lee, 2016), Lee proposed a novel L1-independent mispronunciation framework that does not require non-native data, using acoustic similarity between learners' speech segments. The proposed system focused on discovering the individual learner's phonemic errors, in the first stage. In the second stage, a procedure of decoding pronunciation errors is involved, based on discovered errors. Subsequently, the proposed framework has been improved and experimented with English (L1) learners of Mandarin (L2) (Lee et al., 2016).

While the adoption of unsupervised error discovery approaches provides high error coverage, their performances are relevant to the acoustic models. Therefore, researchers overestimated the benefits of investigating deep learning in pronunciation assessment and MDD to improve the performance at the acoustic model level.

3.3 Deep Learning for Pronunciation Assessment and MDD

Deep neural network acoustic models have drastically improved the performances of ASR-based systems and succeeded to model speech signals (Hinton et al., 2012) through different DL techniques. As ASR is a key component of CAPT, the benefit of deep learning has also impacted the CAPT field for both pronunciation evaluation and MDD. Qian et al. (2012) introduced a mispronunciation detection system based on a deep belief network (DBN)-HMM acoustic model in (L2) English language. The results showed significant improvement in pronunciation error rate compared to the GMM-HMM models. This initial work was followed by an improvement of MDD using DNN acoustic model for both L2 English and L2 Mandarin Chinese (Hu et al., 2015). Further, a neural network (NN) based logistic regression classifier was fed with the obtained GOP scores to improve the performance of the MDD. W. Li et al. (2016) proposed a framework to improve the MDD based on speech attributes that were fed to the DNN model. The proposed system outperformed the GOP-based methods and provides results comparable to the classifier-based approaches. Furthermore, K. Li et al. (2017) used a multi-distribution of DNNs with different input features to detect and diagnose mispronunciations in L2 English.

On the other hand, the task of automatic pronunciation evaluation has been improved using DNNs which outperformed the GMM-based acoustic models (Cheng et al., 2015; Fu et al., 2020; Hu et al. 2013; Tao et al., 2016; X. Chen & Cheng, 2014). Hu et al. (2013) proposed a DNN-based pronunciation scoring system to estimate the GOP score using averaged frame-level posteriors. The GOP score generated from the DNN outperformed the GOP from conventional approaches GMM-HMM. Moreover, the GOP based on DNN achieved the best correlation with the human ratings at both word and sentence levels. Earlier studies have examined the context-dependent DNN-HMM to automatic Spoken Chinese Test (SCT) and to assess the English of children learners (Metallinou & Cheng, 2014; X. Chen & Cheng, 2014). Two network activation functions were considered, the Rectified Linear Unit (ReLU) and the sigmoid. Both activation functions provided a very approximate performance for the Mandarin speech recognition system using both native and nonnative speech data. Afterward, DNN-HMM acoustic models were investigated for the pronunciation assessment of the young English learners and the adult English and Chinese learners (Cheng et al., 2015). The performance results achieved a WER improvement of DNN over GMM with 20.4%, 29.3%, for adult English and child English learners, respectively. The character error rate (CER) was reduced by 14.3% compared to the GMM for adult Chinese. Researchers (Tao et al., 2016) reported that the DNN-HMM significantly outperformed the GMM-HMM in performance recognition and spoken assessment results. The performance of the scoring system using deep learning architectures correlated with human experts for nonnative English spontaneous speech. Meanwhile, using an out-of-domain scoring corpus for the assessment task, degraded the ASR performance when the ASR trained on another non-native spontaneous speech.

Recent studies explored various DL structures for pronunciation assessment. For instance, Lee (2016) investigated CNN to detect mispronunciations from the speech input. Another example is an end-to-end solution based on neural network models for predicting automatic speech scores that has been studied in (Chen et al., 2018). This deep learning-based model used the attention mechanism for bidirectional long short-term memory (BD-LSTM). In fact, the end-to-end BD-LSTM based model outperformed the CNN-based one and the handcrafted conventional techniques for the assessment of online English readiness. Gretter et al (2019) trained a feedforward neural network to predict scores based on extracted features from speech and automatic transcription of spoken sentences. Oh et al. (2020) proposed an automatic spoken proficiency assessment system for non-native speakers reading Korean utterances (Oh et al., 2017).

A new machine score was proposed in (J. Fu et al., 2020), named Reference-free Error Rate (RER), to evaluate the English proficiency of Japanese learners without using the reference sentence. The automatic assessment system combined GMM-HMM and DNN-HMM acoustic models which were trained on native American English and non-native speech data from

Japanese readings. The results achieved a better correlation with the human ratings, and showed the effectiveness of the RER score. The best correlation was achieved when combining RER to other machine scores such as the log-likelihood score.

Low resource languages such as Arabic are still challenging and need more efforts to use data-driven models for automatic pronunciation assessment tasks compared to the English language where data-driven-based models are investigated in more findings.

3.4 Performance Measures

Various measures have been proposed to evaluate the pronunciation assessment quality. Figure 3.5 depicts the hierarchical structure of these metrics. The following scores (mainly at the phoneme level) are identified to measure the effectiveness of a pronunciation assessment system as follows:

1) True acceptation (TA): represents sequences that were pronounced correctly and judged as correct by the automatic system;
2) True rejection (TR): represents sequences that were mispronounced and judged as incorrect;
3) False acceptation (FA): represents sequences that were mispronounced and judged as correct;
4) False rejection (FR): represents sequences that were pronounced correctly and judged as mispronounced.

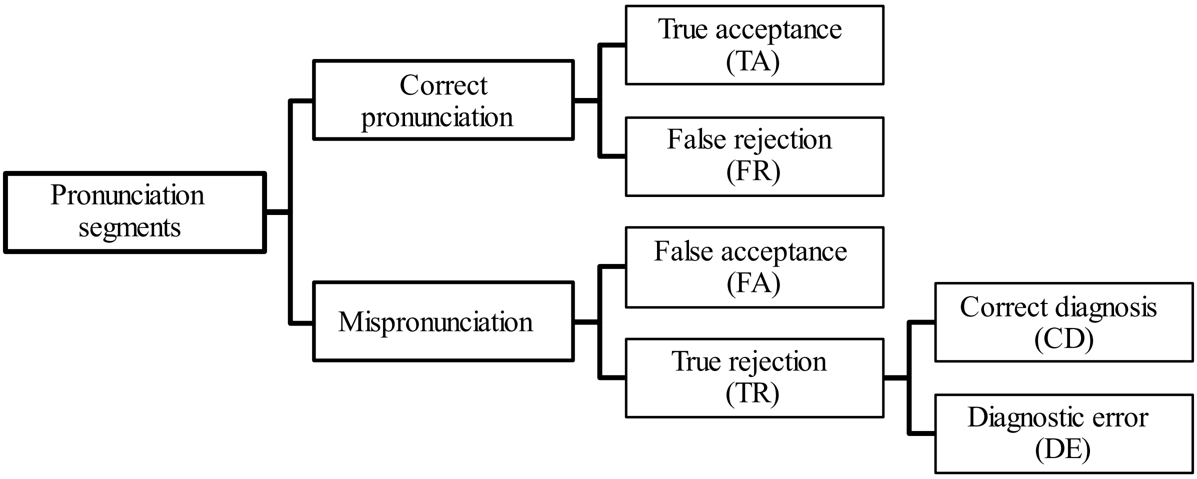


Figure 3.5 Hierarchical structure of the performance metrics for MDD (after, Wang and Lee, 2015)

It is clear according to figure 3.5 that, for correct pronunciation: true acceptance (TA) and false rejection (FR) are the system judgment parameters, whereas false acceptance (FA) and true rejection (TR) are the possible evaluations for mispronunciation. Based on these four scores, the following ones can be computed:

1. **Accuracy:** the accuracy describes the closeness of the automatic decision to the human expert one, it is computed as follows:

$$Accuracy = \frac{TA + TR}{TA + TR + FA + FR} \qquad (3.2)$$

2. **False Rejection Rate (FRR)**: the percentage of the total number of correct pronunciations that are identified by the system as mispronounced. FRR is computed from the ratio between the correct phonemes, identified as mispronounced (FR), and the total number of correct phonemes (TA+FR).

$$FRR = \frac{FR}{TA + FR} \qquad (3.3)$$

3. **False Acceptance Rate (FAR):** the percentage of the total number of mispronounced segments that are correctly accepted by the system. FAR calculated from the ratio between incorrect pronunciations that are accepted by the system as correct (FA), and the total number of mispronounced segments.

$$FAR = \frac{FA}{TR + FA} \qquad (3.4)$$

4. **Diagnostic Error Rate (DER):** the percentage of the incorrect diagnosis (DE), from the total number of correct rejected pronunciations (TR).

$$DER = \frac{DE}{TR} \qquad (3.5)$$

3.5 Nonnative Corpora for Pronunciation Assessment

In the era of big data, a huge amount of speech data is available, however, their use for the development of ASR applications is not easy without transcriptions. In particular, the development of speech corpora for CAPT applications is not an easy task, involving a long time and great resources for phonetic (and tonal) transcription, proficiency rating, and requires annotated mispronunciations in the target language with respect to L1. This might explain their scarcity.

3.5.1 Nonnative English corpora

As a great portion of the world population speaks English (20% of earth population) and the other one attempt to learn it, English is the first recommended language to interact with other people of most countries. Consequently, a wide range of the nonnative available corpora has English as the target language (see table 3.3).

As an example of such corpora, L2-ARCTIC (Zhao et al., 2018) (accessible at https://psi.engr.tamu.edu/l2-arctic-corpus/.) is a free access corpus. L2-ARCTIC was designed to support three tasks: voice conversion, accent conversion, and mispronunciation detection. It includes recordings of Korean, Mandarin, Arabic, Spanish and Hindi speakers learning English. For each L1, two speakers (one male and one female) are recruited from Iowa State University (ISU) students. The proficiency level of their English was measured using TOEFL scores. Recordings were done in a quiet room where speakers were asked to read ARCTIC prompts (for approximately one hour) from Carnegie Mellon University (CMU). Orthographic and phonetic transcriptions are done as well as manual annotation of a selected subset of utterances (150) indicating the types of mispronunciation errors. In (Zhao et al., 2018) difficulties encountered when learning English by several speakers are investigated throughout the corpus recordings. L2-ARCTIC was used to depict a set of common substitution, deletion, and addition errors. For example, the English phone “DH” is replaced by “D” for L1-Hindi and by “Z” for

L1-Arabic. The study of phone additions shows that each of the five L1 backgrounds favored the apparition of a different phone to make words easily pronounceable.

3.5.2 Nonnative Chinese corpora

If English is the most studied language in the world, it is the third most spoken language after Mandarin and Spanish. Besides that, China is an emerging nation with millions of projects around the world. Consequently, one of the most important (in terms of variety and duration) corpora used in CALL context is for learning Chinese Mandarin by English learners (see table 3.3). iCALL (Chen et al., 215) corpus is presented as a nonnative Mandarin corpus for developing computer-assisted language learning applications; it includes the recordings of 305 speakers who are from Europe (including Germanic, Romance, and Slavic origin). The speakers' ages are ranging from 18 to 25 and the gender ratio was balanced. All speakers are beginner learners of Mandarin and rely heavily on the Pinyin phonetic representations (instead of Chinese characters, Pinyin is the Romanization of the Chinese characters based on their pronunciation) to read the prompts. The nonnative speech recordings were recorded in quiet office rooms, sampled at 16 kHz. The fluency scoring protocol was developed by two native Mandarin speakers, every utterance has a proficiency score provided by an expert, ranging from 1 to 4, with 4 being the highest level. iCALL has been used for lexical tone error detection, lexical tone recognition, and automatic fluency assessment (Chen et al., 2015).

3.5.3 Other corpora

Similarly, Spanish is one of the most spoken languages worldwide. One of the freely available nonnative Spanish corpora is the Spanish Learner Oral Corpus (SLOC) (Campillos Llanos, 2014), available at http://www.lllf.uam.es/ING/SLOC.html. Each recording has been synchronized with its orthographic transcription. Moreover, files include data with the proficiency level of the Spanish speaker.

In another hand, many other languages might be considered as low-resourced, in particular Arabic. While Arabic is considered as one of the recommended languages to learn, nonnative speech corpora dedicated to learning Arabic pronunciation are quite inexistent.

Table 3.3 reports datasets that could be used for the development of CAPT systems; few of them were designed for that purpose (such as ISLE, ERJ, iCALL, or L2-ARCTIC) and others might deviate from that goal.

Table 3.2 Abbreviations of languages used in Table 3.3

Arabic	A	Dutch	Dut	Hindi	H	Korean	K	Russian	R
Cantonese	C	English	E	Italian	I	Mandarin	M	Spanish	S
Czech	Cze	French	F	Indonesian	In	Portuguese	P		
Danish	D	German	G	Japanese	J	Polish	Pol		

Table 3.3 Overview of nonnative databases for pronunciation training

Corpus	Target language	Native language	#spkr	#utt	Duration	Trans.	Prof. Rating
ATR-Gruhn (Kim et al., 2016)	E	C, G, F, J, In	96	15,000		No	No
C-AuDiT (Fitt, 1995)	E	F, G, I, S	56	18,424		No	No
CU-CHLOE (Menzel et al., 2000)	E	C M	211	77,437	104,5h	P	No
ERJ (Gruhn et al., 2004)	E	J	200	68,000		No	Yes
ISLE (Honig et al., 2009)	E	G, I	46	11,484	18h	No	No
iCALL (Chen et al., 2015)	M	E F G I P R S et al. (24 in total)	305	90,841	142h	Yes	Yes
L2-ARCTIC (Zhao et al., 20018)	E	K, H, A, M, S	10	11,026	11,2h	Yes	Yes
NTU (Minematsu et al., 2004)	M	36 in total	278	8340		Yes	No
SLOC (Campillos Llanos, 2014)	S	P, I, F, E, Dut, G, Pol, C, J	40		13h36mn	Yes	Yes

3.6 Arabic Pronunciation Assessment

3.6.1 Language particularities

The Arabic language belongs to the Semitic language family; “Semitic languages are marked by a limited vocalic system and a rich consonantal system.” (Watson, 2002). Spoken Arabic has only six vowels, three short vowels /a, u, i/, and their long counterparts /a:, u:, i:/. The short vowels are represented by diacritics (َ- ُ- ِ) in the written form and are essential (for nonnative speakers) to correctly pronounce Arabic words. Typically, when writing the Arabic language, words are written without diacritics (or few ones), hence the reader vowelizes the words according to the context or based on his prior knowledge. Table 3.4 gives the list of the 28 Arabic consonants and their corresponding sound in IPA alphabet, and table 3.5 presents the Arabic sounds grouped by their places of articulation, and manner of articulation.

Table 3.4 List of Arabic letters and their IPA symbol counterpart

Arabic letter	ء	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي
IPA symbol	ʔ	b	t	θ	dʒ	ħ	x	d	ð	r	z	s	ʃ	sˤ	tˤ	dˤ	ðˤ	ʕ	ɣ	f	q	k	L	m	n	H	w	j

Table 3.5 Classical Arabic consonant chart

	Plosives	**Nasal**	**Fricatives**	**Affricate**	**Approximant**	**Trill**	**Glides**
Bilabial	ب	م					و
Labio-Dental			ف				
InterDental	ض-ط-ت - د	ن	ظـ ذ ـ ث				
Dental/Alveolar			ص - ز- س		ل	ر	
Alveo-Platal			ش	ج			
Palatal							ي
Velar	ك		خ غ				
Uvular	ق						
Pharyngeal			ح ع				
Glottal	ء		ه				

Owing to the geographical extent of the Arab world, many dialects exist. However, independently of the different dialects, Arabic can be classified into two major variants: Classical Arabic (CA) and Modern Standard Arabic (MSA). CA is an ancient literary form of Arabic, which is the most formal type and is the language of the Holy Quran. MSA is the current standard form of Arabic and is used in current research work, particularly in automatic speech recognition. “Although there’s no huge difference between today’s Arabic (MSA) and that spoken by the early Arabs (CA), due to the fact that Arabic is one of the most stable languages throughout history” (Khelifa et al., 2017)

3.6.2 State-of-the-Art in Arabic pronunciation assessment

While pronunciation assessment on English has been popular, automatic pronunciation assessment in low resources languages such as the Arabic is still challenging. Research in Arabic pronunciation assessment remains in its infancy; however, many research efforts have been carried for that purpose in recent years. Generally speaking, CAPT systems provide two types of evaluation: pronunciation scoring and mispronunciation detection and diagnosis. Subsequently, findings related to the Arabic language can be divided into automatic pronunciation evaluation and mispronunciation detection and diagnosis.

3.6.2.1 Pronunciation scoring

Conventional speech assessment approaches focused on extracting features from HMM-based ASR. First, Abdou et al. (2006) developed the HAFSS system for teaching Arabic pronunciations to non-native speakers. HAFSS used phoneme duration and confidence scores to detect errors and to give feedback on mispronounced letters. The confidence score is computed based on the likelihood ratio. Another pronunciation scoring tool based on HMM log-likelihood probability was developed by (Khan et al., 2013) to learn and teach Arabic for Malaysian teachers. The proposed system was trained on native and non-native utterances and outperformed the system trained only on native utterances.

Lee and Glass (2013) proposed a pronunciation scoring approach to compare students' and teachers' pronunciations via the dynamic time warping (DTW). The proposed approach aligns the student's speech and the teacher's speech similarities on Levantine Arabic language, in terms of Mel-frequency cepstral coefficients (MFCCs), Gaussian posteriorgrams (GPs), and English phoneme state posteriorgrams. Subsequently, Necibi and Bahi (2015) investigated a statistical-based method to detect difficulties in the reading skills of pupils at an earlier stage.

Recently, Bahi and Necibi (2020) proposed a fuzzy logic-based system for pronunciation assessment (FuSPA). FuSPA aims to enhance the existing thresholding method and to overcome limitations related to the experts' rating disparities. As depicted in figure 3.6, the fuzzy-based system was built based on a speech recognizer, fuzzy evaluation engine, and feedback components. The speech recognizer was trained on native speech to construct an HMM-based acoustic model. Three pronunciation quality levels were considered: good, poor, and acceptable level. The fuzzy engine evaluates the pronunciations based on the fuzzy rules stored in the knowledge base as previous assessments. Two machine scores were combined in a fuzzy manner, the time duration score (TDS) and the global log-likelihood (GLL) for automatic pronunciation assessment.

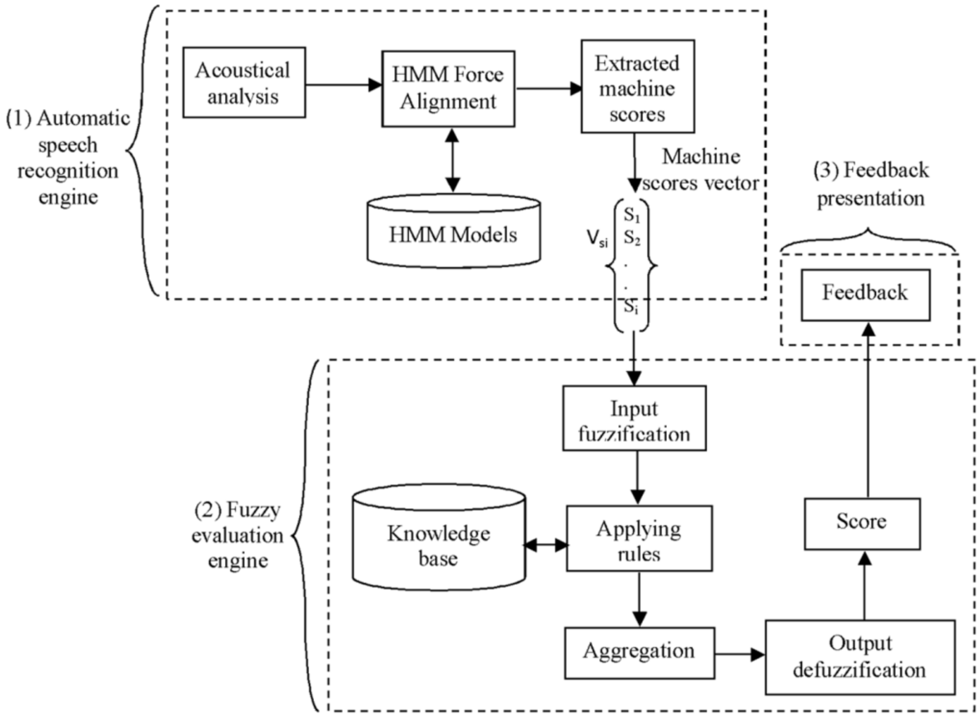


Figure 3.6 FuSPA components (after, Bahi and Necibi, 2020)

The FuSPA system achieved better performances when the input pronunciation meets the reference one. Moreover, it alerts teachers on difficult pronunciation or in case of mispronunciations.

3.6.2.2 Pronunciation errors detection and diagnosis

CAPT applications focus on two challenging tasks: mispronunciation detection and diagnosis (MDD). Pronunciation error detection is the process that precedes the diagnosis in language learning applications. While MDD in English and Chinese are popular, pronunciation error detection and diagnosis in Arabic remains challenging, due to the lack of L2 speech data. Below are presented the papers that dealt with Arabic detection of pronunciation errors.

Al Hindi et al. (2014) applied the Goodness of Pronunciation (GOP) score to detect errors in the pronunciation of non-native Arabic speakers at the phoneme level. The study focused on five phonemes: Tha’a (/θ/ث), Ha’a (ح/ħ/), Sad (/ş/ص), Dad (/ď/ض), and Dha’a (/đ/ظ). The selected phonemes are the difficult ones to pronounce by non-native Arabic learners. Maqsood et al. (2016) trained an SVM classifier for each of the five phonemes, using acoustic-phonetic features (APF). The Arabic mispronunciation detection was performed over the five abovementioned Arabic phonemes, to detect pronunciation mistakes for Pakistani (L2) Arabic learners. The proposed system outperformed the GOP-based classifier for Arabic mispronunciation developed in (Al Hindi et al., 2014). Hammami et al. (2020) trained a probabilistic classifier based on MFCC features to detect and identify speech sound errors at Arabic words containing the letter ‘r’ among Arabic native children.

Deep learning techniques have been used recently in Arabic MDD. Nazir et al. (2019) developed two approaches for mispronunciation detection that outperformed the state-of-the-art methods on the 28 Arabic phonemes. In the first approach, a deep convolutional neural network (CNN) extracted features from its different layers. Subsequently, the classification algorithms (KNN, SVM, and NN) detected mispronunciations using the extracted features. The second approach was based on the transfer learning method where both feature extraction and classification are performed by the CNN to detect whether a phoneme is mispronounced or not. Herein, the feature extraction is based on transfer learning using the pre-trained deep CNN AlexNet (Krizhevsky et al., 2012).

The two methods were compared with the baseline methods, the handcrafted feature-based method (Figure 3.7), and outperformed it. The proposition also outperformed the two proposed Arabic mispronunciation detection at the same five phoneme levels (Al Hindi et al., 2014; Maqsood et al., 2016).

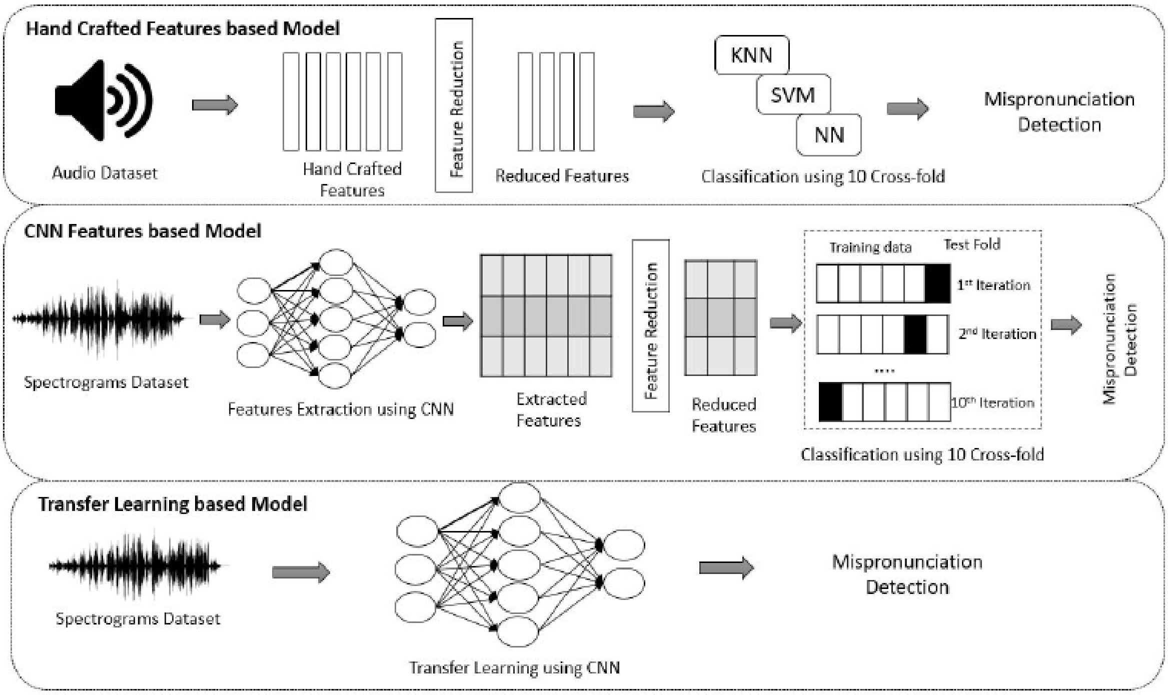

Hand Crafted Features based Model
Audio Dataset
Hand Crafted Features
Feature Reduction
Reduced Features
KNN
SVM
NN
Classification using 10 Cross-fold
Mispronunciation Detection
CNN Features based Model
Spectrograms Dataset
Features Extraction using CNN
Extracted Features
Feature Reduction
Reduced Features
Training data
Test Fold
1st Iteration
2nd Iteration
....
10th Iteration
Classification using 10 Cross-fold
Mispronunciation Detection
Transfer Learning based Model
Spectrograms Dataset
Transfer Learning using CNN
Mispronunciation Detection

Figure 3.7 CNN-based approach for Pronunciation assessment (Nazir et al., 2019)

3.6.2.3 Feedback

After the mispronunciation detection process, a CAPT system with diagnostic functionalities should be capable of providing feedback to (L2) learners to help them improve their pronunciation. Only a few studies investigate feedback in the Arabic language such as (Abdo et al., 2006; Alsabaan & Ramsay, 2014). First, Abdo et al. (2006) developed the HAFFS CAPT system that teaches and provides feedback to Arabic non-native speakers. The feedback is provided in several forms. Second, a computational tool developed by (Alsabaan & Ramsay, 2014) provides different forms of feedback to non-native Arabic learners. Particularly, their work aimed to determine which form of feedback is the most effective. To do that the system "analyses the differences between the user's pronunciation and that of a native speaker by using the grammar of minimal pairs". A minimal pair is a pair of words with one phonemic difference only, such as in "kalb" (dog) and "qalb" (heart). In the study of (Alsabaan & Ramsay, 2014), each incoming utterance is considered to belong to a set of words that sound similar. The tool provides feedback in three different sub-tools. First, as an animation of the vocal tract; the learner is given a graphical representation of both the way the sounds are articulated and the way the sound is produced correctly. Synthesized speech is the second source of feedback to learners. They can play their voice, listen to a synthesis version of what they said, and listen to a correct synthesis version. Written instruction is the third source of feedback to the learner. A written description of how the learner can pronounce the intended phoneme is displayed. The following figure describes an illustration of the feedback process.

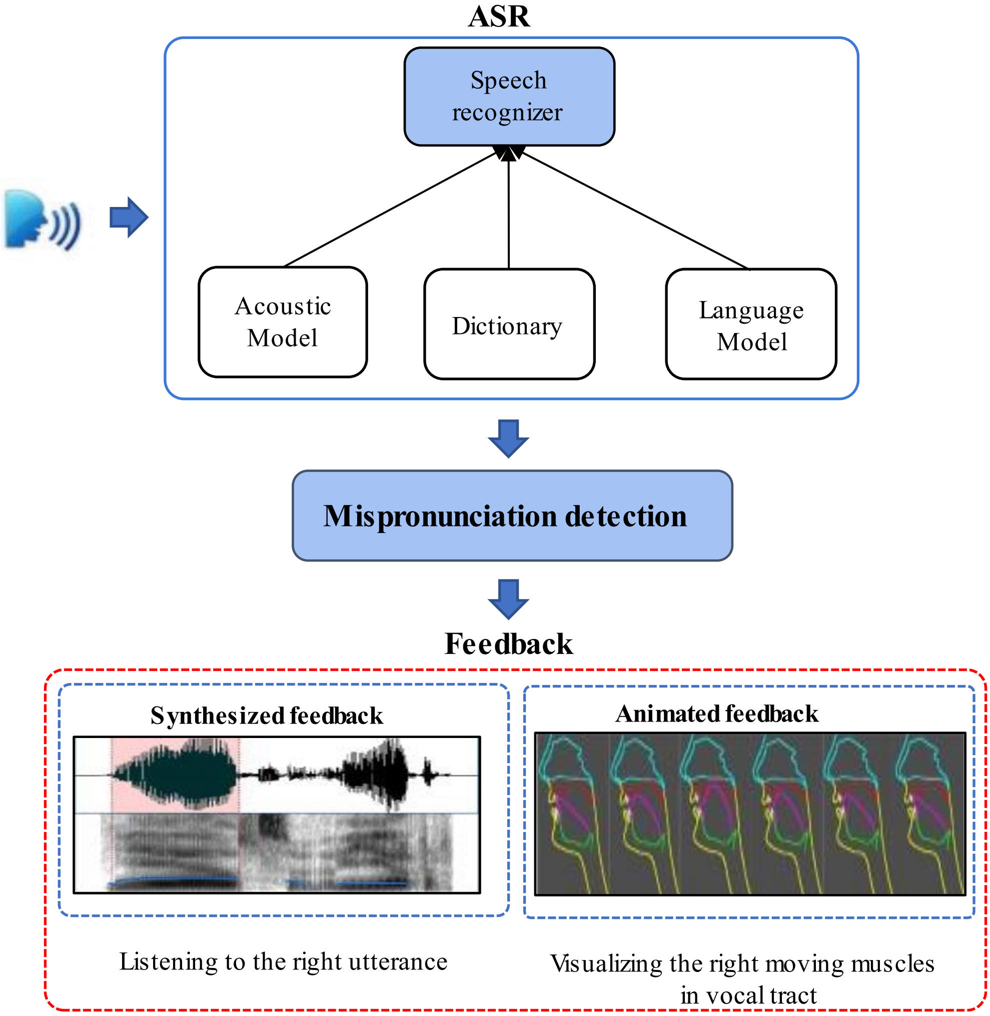

ASR
Speech recognizer
Acoustic Model
Dictionary
Language Model
Mispronunciation detection
Feedback
Synthesized feedback
Animated feedback
Listening to the right utterance
Visualizing the right moving muscles in vocal tract

Figure 3.8 Illustration of the feedback in (Alsabaan, 2015)

3.6.2.4 Summary of algorithms in Arabic pronunciation assessment

Table 3.6 summarizes different approaches from the literature review of the Arabic pronunciation assessment studies.

Table 3.6 A summary of different approaches for Arabic pronunciation assessment

Reference	**Task**	**Corpus**	**Algorithm**	**Accuracy (%)**	**Observation**
(Abdou et al., 2006)	Error detection and feedback on Quranic recitation	Holy Quran recitation	Phoneme duration classification.	62.4	The study was based on the phoneme duration score.
(Cheng et al., 2009)	Oral proficiency assessment	Six sentences, 246 hours of speech from native (116 hours) and non-native Arabic speakers (130 hours)	HMM		
(Khan et al., 2013)	Pronunciation assessment	110 sentences pronounced by 20 native and 10 non-native speakers.	HMM	89.69	Log-likelihood score
(Lee & Glass, 2013)	Pronunciation scoring	Levantine Arabic, 100 sentences, 21 non-native and 04 native speakers.	dynamic time warping	/	Alignment-based features (MFCC, and posteriograms) between teachers' and learners' utterances.
(Al Hindi et al., 2014)	Detecting pronunciation errors on five phonemes (/θ/ث), (ح/ħ/), (/ş/ص), (/ďʼ/ض), and (/đ/ظ))	King Saudi University (KSU) Arabic Speech Database, 32 native and 16 nonnative speakers.	GOP classifier	87 to 100	Log-likelihood based-score
(Necibi & Bahi, 2015)	Pronunciation assessment	Recorded data set of speakers pronouncing 16 utterances (9 speakers)	The statistical approach based on the student test (t test).	97.31	Log-likelihood and phoneme duration-based scores.

(Necibi et al., 2015)	Arabic Pronunciation Evaluation	Collected dataset of 15 speakers pronouncing a list of 100 MSA words	Decision Tree	96.55	Global average log-likelihood (GLL) score, Time Duration of the Speech (TDS)
(Maqsood et al., 2016)	Arabic Mispronunciation detection for five phonemes (ث, ح, ص, ض, ظ)	Recorded data set of 100 Pakistani speakers, 60 native and 40 nonnatives.	SVM	97.5	Acoustic Phonetic features (APF)
(Nazir et al., 2019)	Mispronunciation detection in Arabic phonemes.	Arabic data set recorded from 400 non-native Pakistani speakers.	transfer learning-based model	92	CNN features-based technique and the handcrafted features.
(Nazir et al., 2019)	Mispronunciation detection on the five Arabic phonemes (ث, ح, ص, ض, ظ)	Arabic data set recorded from 400 non-native Pakistani speakers.	Transfer learning	99.23	
(Hammami et al., 2020)	Speech sounds error detection for words containing the letter ‘r’	Arabic speech sound error data set contains 900 utterances from 60 native Arab children (30 boys and 30 girls), 4-12 years aged	Probabilistic classifier	71.75, 77.20, 74.06	Posterior probability
(Bahi & Necibi, 2020)	Automatic Pronunciation Assessment	Recorded data set of 09 children pronouncing 16 utterances.	Fuzzy logic	76	A fuzzy combination between two machine scores, the TDS, and the GLL.
(Al-Marri et al., 2018)	Qur’anic recitation error detection, feedback and correction, for ten letters (ث, ح, خ, ع, غ, ذ, ص, ض, ط, ظ)	1000 wave files, collected from 100 speakers pronouncing 10 sets of letters. The total duration is about 83 hours.	DNN-HMM	92.84	

3.7 Conclusion

In this chapter, we introduced the pronunciation assessment principles and reviewed the related approaches. In particular, I have reviewed methods relevant to automatic pronunciation scoring, mispronunciation detection, and diagnosis by presenting different works for each approach. Then, I introduce a summary comparison of Arabic pronunciation assessment works.

According to this review, one of the challenges facing low resource languages such as the Arabic is the lack of non-native suitable speech corpora. These corpora are crucial for the development and test of different hypotheses in CAPT. Thus, I proposed an unsupervised mispronunciation method to overcome this challenge. The proposed method is based on one-class objective training and uses deep learning methods for pronunciation assessment, as detailed in chapter 5.

Chapter 4: Speech Enhancement for Ubiquitous Arabic Speech Recognition

4.1 Introduction

In this chapter, I present the supervised and self-supervised speech enhancement propositions for Arabic speech recognition in ubiquitous environments under challenging real-world conditions. I particularly detailed the method, based on two steps to perform Arabic speech enhancement in the absence of a dedicated speech corpus. The final target of the enhancement stage is speech recognition for further CAPT purposes. The two steps self-supervised speech enhancement approach was implemented by an over-complete DAE model, followed by an under-complete DAE, which brings new perspectives to unsupervised learning. Precisely, the results proved that the WER was reduced, and both PESQ and STOI were improved. This chapter detailed the application of speech enhancement technology to improve the performance of the speech recognition system under noisy conditions.

As already said, one of the motivations for using the architecture of NSR instead of DSR and ESR is the simplicity to update the ASR components at the server-side. NSR is identified by the location of both feature extraction and ASR at the server-side while the speech signal is captured at the client-side as shown in Figure 4.1. Network and cloud-based speech recognition systems can be used in developing regions where the terminals are low-resources cellular phones. Moreover, there is no need for increased resources on the client-side because the central server handled the feature extraction and the decoding process. Figure 4.1 presents a ubiquitous network speech recognition system for the Arabic language.

One of the disadvantages caused by this approach is the performance degradation of the recognition process due to the use of low-bit-rate codecs for encoding speech, which becomes more severe when data transmission errors occur, and in the case of the noise background.

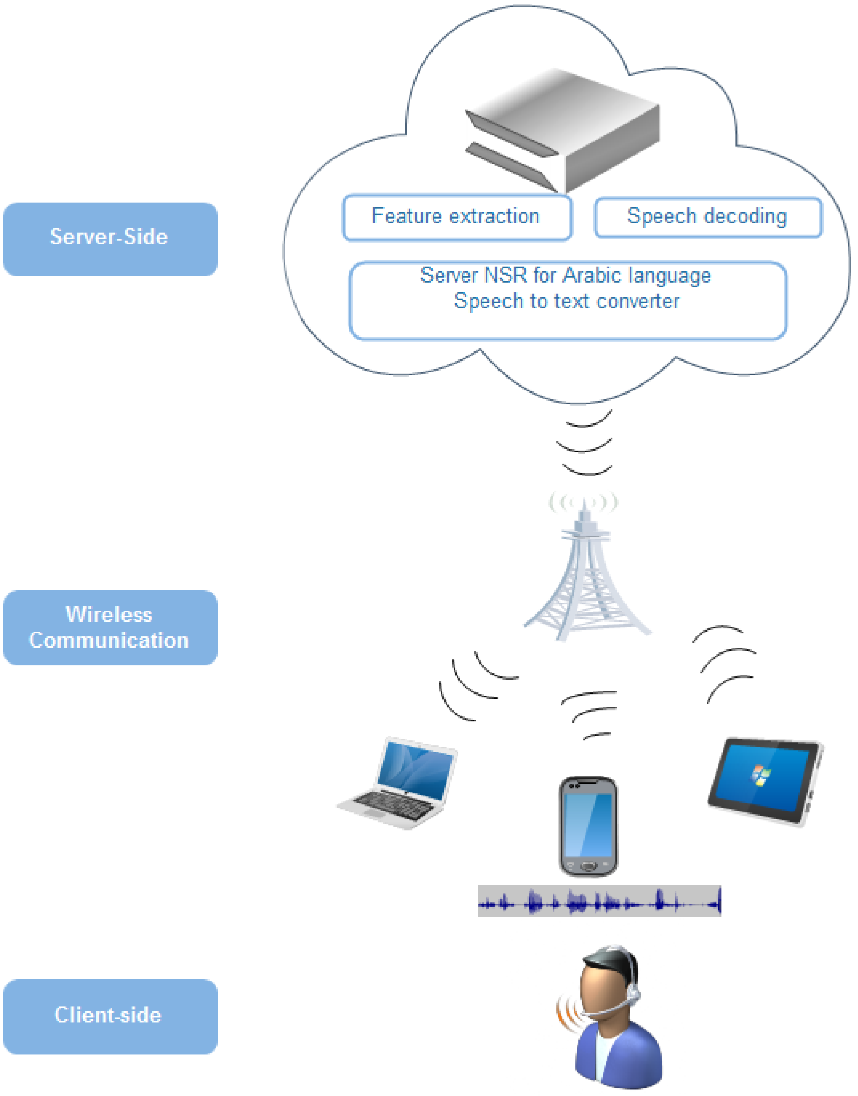


Figure 4.1 A ubiquitous network-based ASR system for Arabic (Dendani et al., 2019)

Another issue of this mode is how to serve requests that come simultaneously from clients. To overcome these limitations, I considered the two speech codecs G.711and G.728 as the speech compression algorithms for supervised and unsupervised proposed methods, respectively. In addition, I performed the speech enhancement task to remedy the potential noises caused by the ubiquitous context.

4.2 A Ubiquitous Speech Recognition System for the Arabic Language

Building efficient and robust ubiquitous speech recognition systems is a challenging complex task that requires the implementation of several parts. Figure 4.2 depicts the different modules and their location at the client or the server-side. These modules are detailed below as speech coding, speech enhancement, and speech recognition. The speech enhancement is used as the preprocessor for automatic speech recognition.

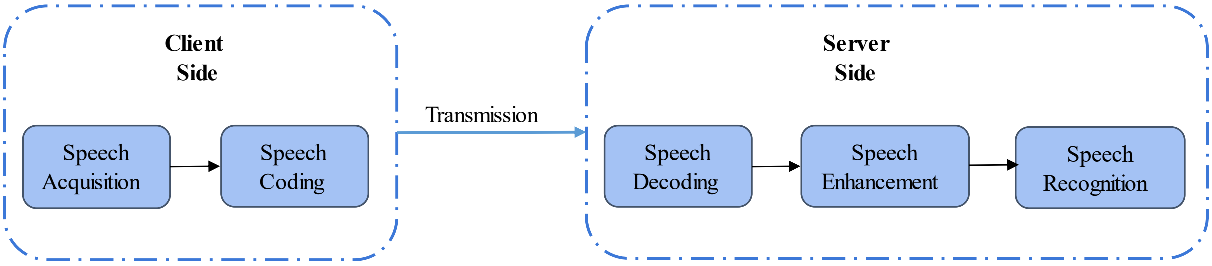


Figure 4.2 Ubiquitous Arabic speech recognition block diagram

4.2.1 Speech acquisition and coding

The uttered speech is captured at the client-side in the ubiquitous environments, as depicted in figure 4.1. Meanwhile, various real-world noise conditions altered the speech signal. Afterward, the speech coding step is performed at the client-side, before the transmission using the software tool library G.191 standardized by ITU-T (ITU-T, 2019). The G.728 ITU-T standard speech codec (ITU-T G.728, 1992) is based on the Low-Delay Code Excited Linear Prediction (LD-CELP) compression principles and provides a bit rate at 16 kbps. The G.728 speech codec is chosen based on its low bit rate since the low bitrate allows, eventually, the use of the remaining bandwidth for video transmission (this is of much interest for mobile applications involving other modalities than speech, such as CAPT). Moreover, the encoded speech signal is transmitted in real-life conditions and at different SNR levels towards the server-side for subsequent enhancement and recognition.

4.2.2 Speech enhancement proposed approaches

Two approaches are proposed, according to the availability of clean speech signals. A supervised SE approach assumes that clean speech data are available for training the DNN model. A set of training pairs of noisy and clean speech signals are engaged to minimize the loss function between the noisy input frames and the original clean frames. The second proposed SE approach is self-supervised that deals with ubiquitous real-world environments in the absence of labeled clean data. The following sections details each method.

4.2.2.1 A supervised SE approach based on DNN using LPS inputs and outputs

The present DNN-based SE approach assumes that clean speech signals are available for the training dataset. The DNN model is expected to enhance the speech degraded by the G.711 codec and the transmission. The inputs of the DNN are the five consecutive frames of LPS from the noisy speech data, and the output is one frame of LPS from the clean one. The tunning step concerns the inclusion of the left/right context of the target window. The corresponding spectrum power coefficients are extracted from each window containing 512 samples. The DNN adopts a non-linear mapping function between noisy features and clean ones.

Figure 4.3 depicts the DNN architecture where two frames from the left and two from the right of the targeted window are analyzed. The DNN contains five successive hidden layers of 2048 units per hidden layer. The input of DNN consists of five consecutive frames of LPS of noisy speech, and the output is a single frame of LPS of the clean speech.

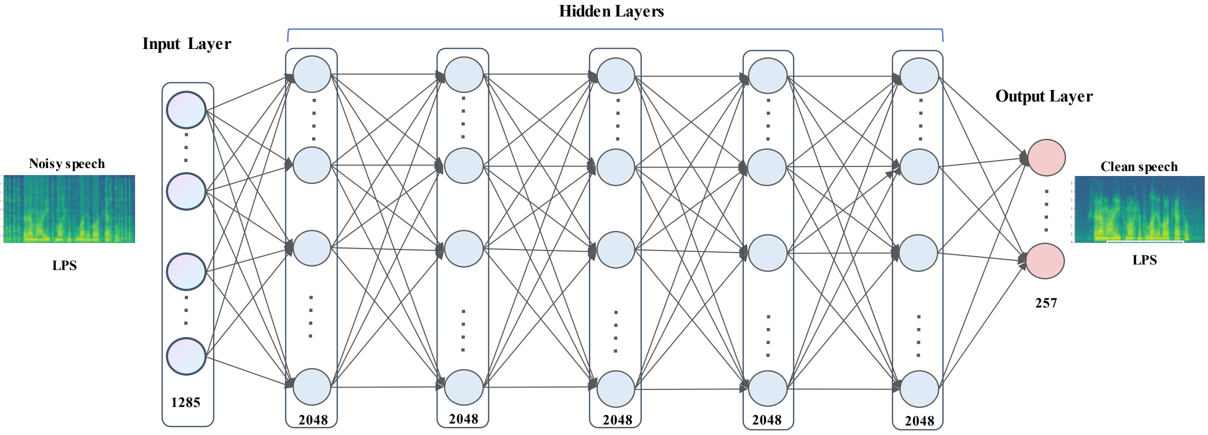


Figure 4.3 A supervised SE-based deep neural network with five hidden layers. the input of DNN corresponds to five consecutive frames, and the output stands for the middle one (the target to enhance) (Dendani et al., 2020)

4.2.2.2 Self-supervised two steps SE approach based on deep auto encoders architectures

The two-steps approach is a self-supervised SE method that assumes the clean labeled speech data are not accessible. Therefore, it is entirely suitable for the conditions of ubiquitous real-world environments, where various noisy speech signals are available, and not the clean ones.

According to the steps involved in figure 4.2, the encoded speech signal is sent from the client to the server-side. The server received the transmitted speech signal and decoded it before postprocessing. Afterward, the speech enhancement stage starts before the recognition one. The received signal is windowed into frames of 512 samples. The log power spectrum (LPS) is estimated from each temporal frame to obtain the frequency representation (Semmlow, 2012). The Short Fast Fourier Transform (SFFT) produces a symmetric vector, thus only 257 values were kept for each frame.

The obtained vector is submitted to a denoising deep autoencoder (DDAE) to produce the enhanced version; we called it the UAE (for Under Complete Auto Encoder). Both input and the output layers to the UAE model are log power spectrum coefficient (LPS), represented by 257 neurons. A two-steps approach was proposed to build the UAE model, as depicted in figure 4.4. Different phases are involved to enhance noisy speech signals.

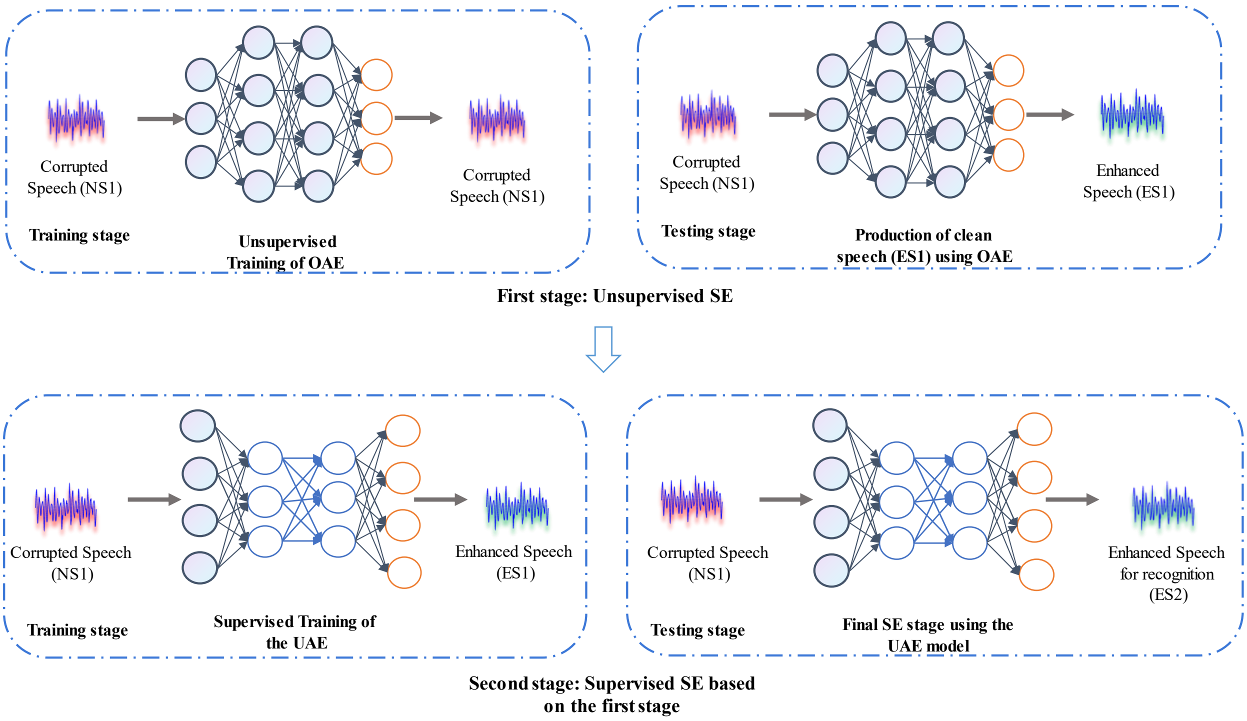

Corrupted Speech (NS1)
Training stage
Unsupervised Training of OAE
Corrupted Speech (NS1)
Corrupted Speech (NS1)
Testing stage
Production of clean speech (ES1) using OAE
Enhanced Speech (ES1)
First stage: Unsupervised SE
Corrupted Speech (NS1)
Training stage
Supervised Training of the UAE
Enhanced Speech (ES1)
Corrupted Speech (NS1)
Testing stage
Final SE stage using the UAE model
Enhanced Speech for recognition (ES2)
Second stage: Supervised SE based on the first stage

Figure 4.4 The two-step-based DL proposed approach for speech enhancement

In the first stage, the overcomplete autoencoder (OAE) is trained using Adam optimizer in an unsupervised manner with a fixed learning rate of 0.0001 (Kingma & Ba, 2015). An overcomplete AE is an AE “in which the hidden code has dimension greater than the input” (Goodfellow, 2016, p. 507) as depicted in figure 4.4 (First stage: Unsupervised SE). During the training stage of the OAE, both the input and the output are 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 (Dendani et al., 2021). 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 the clean speech data. These noisy/clean pairs of the speech signals 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-steps DAE approach is evaluated indirectly in terms of the WER score obtained after the recognition stage, given that the speech recognition task is the end-user application. Moreover, the SE unsupervised-based model is evaluated according to the perceptual evaluation of speech quality (PESQ) and intelligibility (STOI) metrics.

4.2.3 Speech recognition

Once the speech enhancement was performed, the resulting speech was fed to a speech recognizer for recognition. The Hidden Markov Models (HMMs) were used to model speech and to generate the acoustic models (Frihia & Bahi, 2017; Rabiner, 1989) using the Sphinxtrain tool (Walker et al., 2004). The acoustic models were built from the clean modern standard Arabic corpus described in (Almeman et al., 2013). During the test stage, the real-world noisy mobile utterances were decoded using PocketSphinx and the WER performance score computed using SphinxTrain tools. Moreover, the HMM-based speech recognizer is considered as a black box without tuning for the experiments related to the SE.

4.3 Experimental Setup

Several experiments were conducted to assess the proposed model. The performances of the speech recognition system were investigated under noisy conditions after SE using the two-steps proposed model. On the other side, the obtained model was compared to other models. Experiments were carried on the two corpora described below.

4.3.1 Datasets

The Arabic mobile parallel multi-dialect speech corpus is a free corpus that includes four Arabic dialects: Modern Standard Arabic (MSA), Levantine, Gulf, and Egyptian, consists of 67132 wave files uttered by 52 speakers and sampled at 48 kHz with 16 bits of precision (Almeman, 2018). For our experiments, we explored the MSA subset that contains 15492 utterances from 12 speakers. The data were collected in four different environments, inside home, in a moving car, in a public place, and a quiet place. The recordings from public areas and streets varied between the high noise and the medium noise. Moreover, noises that occur in the background can be classified into non-human (door closing, cutlery sounds, car horns, road traffic) noise, and human noise (crying, shouting, speaking). Besides that, many additional factors can affect mobile call quality, such as network signal quality, recording quality, and the distance between the mobile and the mouth, etc. The speech from the data set in (Almeman, 2018) is called NS1.

Besides the inherent noises, the recordings from (Almeman, 2018) were corrupted using some noises selected from 100 non-speech sounds (G. Hu & Wang, 2010). Different noise types were considered, ranging from the stationary car noise to the non-stationary noises (crowd and door moving) at different SNR levels. For the training set, 75% of the corrupted speech signals were used at SNR -5 dB, 0 dB, and 15 dB. For the testing set, the remaining 25 % corrupted speech signals (different from those used in the training set) were considered at -5dB, 0 dB, and 15dB SNR levels. The NS1 speech corrupted with the noises from (G. Hu & Wang, 2010) is called NS2.

A second dataset was used in this study, it is the Arabic speech corpus for isolated words from the department of Management Information Systems at Saudi King Faisal University (Alalshekmubarak & Smith, 2014). It consists of 9992 recorded utterances of 50 speakers pronouncing 20 words. Experiments related to this corpus are detailed in the next chapter, as it was used to assess our second proposition related to pronunciation assessment in real-world environments.

4.3.2 Model's training

All the proposed models were implemented using the TensorFlow library (Abadi et al., 2016). I used TensorFlow 1.4 and TensorFlow 2.3 for implementation of the two steps unsupervised SE and FCN-based models, respectively. All models were trained using Adam optimizer (Kingma & Ba, 2015), at a learning rate of 0.0001. The loss function is the mean-squared error between the estimated log-magnitudes and the log-magnitudes for the original clean signals. For both proposed models (supervised and unsupervised), the number of considered epochs is 30 epochs, used for training and to get the lowest validation loss. The parameters of the models obtained when the validation loss is the lowest were subsequently used for denoising the speech.

4.3.3 Objective evaluation metrics

While the main focus of this study is to improve the performance of the ubiquitous ASR system by reducing the word error rate, the effectiveness of the proposed SE approaches was also evaluated via the quality and objective intelligibility of the speech signal. The Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001) is a metric used to evaluate the perceptual quality of given speech signals by comparing the original clean signal with the degraded one. The result is a prediction in the range of [-0.5, 4.5]. On the other hand, the objective intelligibility of speech is measured using the Short-Time Objective Intelligibility (STOI) metric (Taal et al., 2011). The STOI reflects the correlation between the original speech and the speech to evaluate, in the range of 0 and 1. For both metrics, PESQ and STOI, the higher value is the best.

4.4. Experimental Results

The first experiments aim to validate the SE proposed methods, which serve as pre-process for the ubiquitous ASR application in real-world environments. These experiments were carried out on the real-world mobile speech corpus, providing the following results.

4.4.1 Effect of coding and transmission

The effect of speech coding is explored by evaluating the transcoded speech (encoding and decoding of speech). We report the WER score on real-world mobile speech signals. The transcoded speech signal is received at the server-side (NS1), then the HMM-based recognition took place without prior enhancement of the incoming speech signals. As depicted in figure 4.5, the reported WER is compared to the performances of the ASR system obtained at the client-side after speech recognition (before coding and transmission steps).

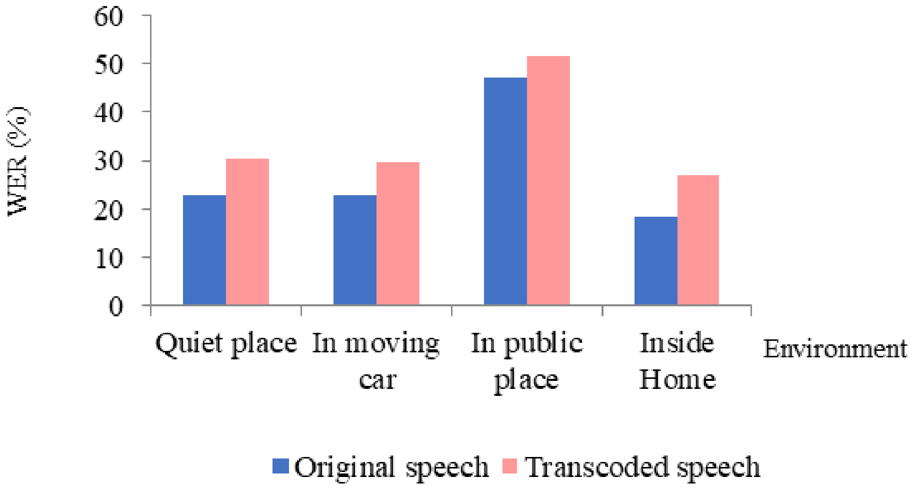


Figure 4.5 The ASR performance for original and transcoded speech (Dendani et al., 2021)

Figure 4.5 summarizes the results of the WER for the four environments. As we can see, the noisy background causes degradation of the WER. Indeed, the worst WER is seen in “public places” where the noise is unknown and from multiple sources, while the best WER is seen in the “inside home” environment. It is worth mentioning that coding and transmission degrade yet the WER of the speech signals received at the server-side.

4.4.2 DAE SE algorithm preprocessor for speech recognition

As discussed previously, the architecture of the DDAE SE model has 257 units for the input layer as well as for the output layer. The size of the hidden layer is 200 neurons, this model consists of the UAE (200). As depicted in figure 4.4, the OAE serves to produce clean speech data for training the UAE. The input layer and the output layer have the same size of 1024 neurons for the OAE, it is the OAE (1024). Table 4.1 reports the WER score after SE with UAE (200) model and compares the performance to the traditional SE well-known method MMSE. Moreover, we compare these obtained results to the performances achieved before applying the SE for the four environments.

Table 4.1 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. MMSE method does not perform well when non-stationary noises are present or at multiple sources of noises.

Table 4.1 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

4.4.3 DAEs fine-tuning strategy

When using the autoencoder model, choosing the right degree of compression is often a hyper-parameter that requires tuning to get the optimal results. Therefore, once the positive effect of the proposed SE approach is testified, further experiments are performed. We fine-tune the DAEs structures using various depths and a different number of neurons per layer.

The experiments reported in tables 4.2, and 4.3, investigated different configurations (depth and number of units) of stacking OAE and UAE structures. While multiple OAE configurations are tested in table 4.2, the UAE model that follows the OAE architecture has one hidden layer with 200 neurons of size. Table 4.3 examined the UAE of two hidden layers of depth with 200 neurons for each layer.

Regarding different models' sizes examined in table 4.2 and 4.3, it is noteworthy that all the configurations show an improvement of the WER score, confirming the effectiveness of the proposed SE model. It might be justified as the projection of the signal characteristics from the OAE in a higher dimensionality space that allows the isolation of noises' features.

From table 4.3, we can confirm the potential of the suggested SE approach as the WER results are improved for the ubiquitous speech recognition system compared to the WER computed before performing SE.

As seen in both tables 4.2 and 4.3, the reduced WER achieved using different SE models is due to the advantage of using fully connected-based models to model multiple complex real-world noisy environments. In particular, using an OAE for unsupervised pre-training provides a solution to generate clean data and allows the training of the classical DDAE.

Table 4.2 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 4.3 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

Table 4.4 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

To choose the best deep learning model that deals with SE under unsupervised real-world speech data. Another needed focus is the balance between the number of total parameters and the WER value. Indeed, optimizing computational resources and the time required for training the model is one of the paradigm challenges. Table 4.4 reports the number of parameters and the performance for each configuration.

From Table 4.4, the model#2 achieves a good trade-off between the number of parameters and the WER.

4.4.4 Comparison of unsupervised two steps SE approach with other SE methods

For a fair comparison, all deep neural networks are trained and tested with the same real-world mobile database, using different SNR levels described above. The performance metrics used to evaluate the various models were the three mentioned measures PESQ, STOI, and WER. For the following experiments, Model#2 stands for the proposed model. The other three additional DL models were used for comparison. A denoising deep autoencoder model (DDAE) called UAE2, trained in a supervised manner and has the same architecture as the UAE of Model#2, described in table 4.4. The UAE2 was trained on NS2 as input and NS1 as output. The two other models are based on a fully convolutional neural network that replaced the UAE in the second stage of Model#2. As depicted in figure 4.4, the OAE of the first stage is called OAE1. On the other hand, the second step is replaced by either FCN2 with two hidden layers or FCN4 with four hidden layers. I named the two structures OAE1-FCN2 and OAE1-FCN4, respectively.

It is worthy to notice that the FCN architecture used for comparison is the one described in (S. W. Fu, Wang, et al., 2018), with different depth (two and four) 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 4.5 shows the average results obtained from each network for the three measures. First, we note that the four SE models outperform the noisy version concerning the average of PESQ, STOI, and WER scores. Next, the average performance values of Model#2 achieved improvements over other models in terms of the PESQ, STOI, and WER. We assumed that, for computing the PESQ and the STOI, we considered utterances in the quiet place as the reference ones.

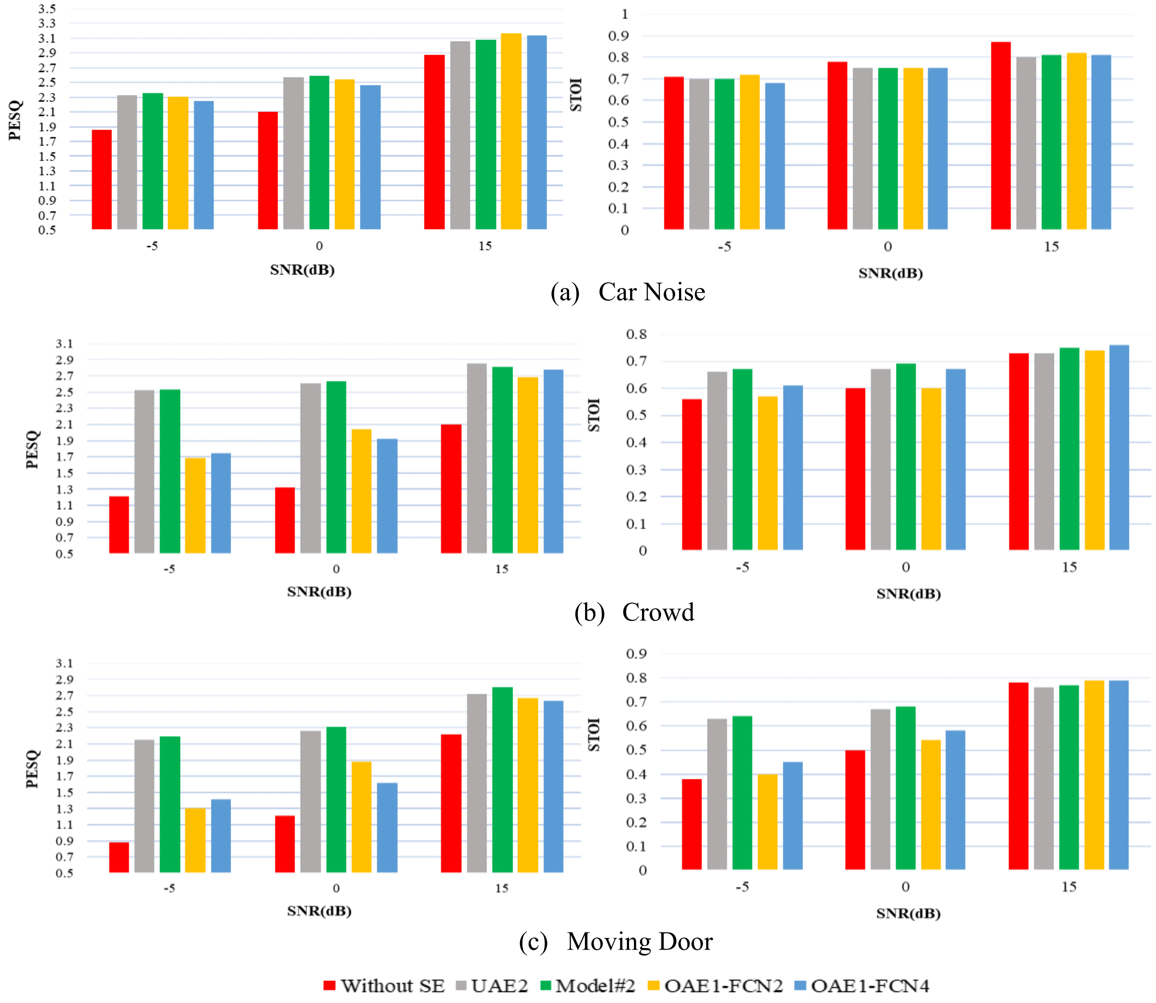
Table 4.5 Average performances of the different models based on the PESQ, STOI, and WER

Metrics	Model	Car Noise	Crowd Noise	Moving Door	Average
PESQ	Without SE	2.276	1.543	1.436	1.752
	UAE_2	2.653	2.66	2.376	2.563
	Model#2	2.673	2.656	2.433	**2.59**
	$\mathrm{OAE}_1-\mathrm{FCN}_2$	2.673	2.133	1.95	2.252
	$\mathrm{OAE}_1-\mathrm{FCN}_4$	2.616	2.146	1.886	2.22
STOI	Without SE	0.786	0.63	0.553	0.66
	UAE_2	0.75	0.686	0.686	0.71
	Model#2	0.753	0.703	0.696	**0.72**
	$\mathrm{OAE}_1-\mathrm{FCN}_2$	0.763	0.636	0.576	0.66
	$\mathrm{OAE}_1-\mathrm{FCN}_4$	0.746	0.68	0.606	0.68
WER (%)	Without SE	49.546	94.98	78.236	74.254
	UAE_2	42.06	75.313	59.99	59.121
	Model#2	42.396	74.04	60.046	**58.827**
	$\mathrm{OAE}_1-\mathrm{FCN}_2$	58.51	95.29	78.376	77.392
	$\mathrm{OAE}_1-\mathrm{FCN}_4$	61.72	93.26	78.896	77.958

As depicted in figure 4.6 and table 4.5, the proposed approach achieved the best PESQ value compared to other models, at different SNR levels, except at 15dB for car and crowd noises. Meanwhile, UAE2 provided better performances than the FCN-based models and achieved a comparable result to our proposed model in terms of PESQ and STOI.

We also note that the proposed model yields an improvement of up to 0.835 on PESQ, 0.06 on STOI, and 15.43% on WER relative to complex noise environments at different SNR levels.

For a detailed comparison, figure 4.6 reports the PESQ and the STOI of the various models for car noise (stationary noise), crowd noise (non-stationary noise), and a moving door non-stationary noise.



(a) Car Noise

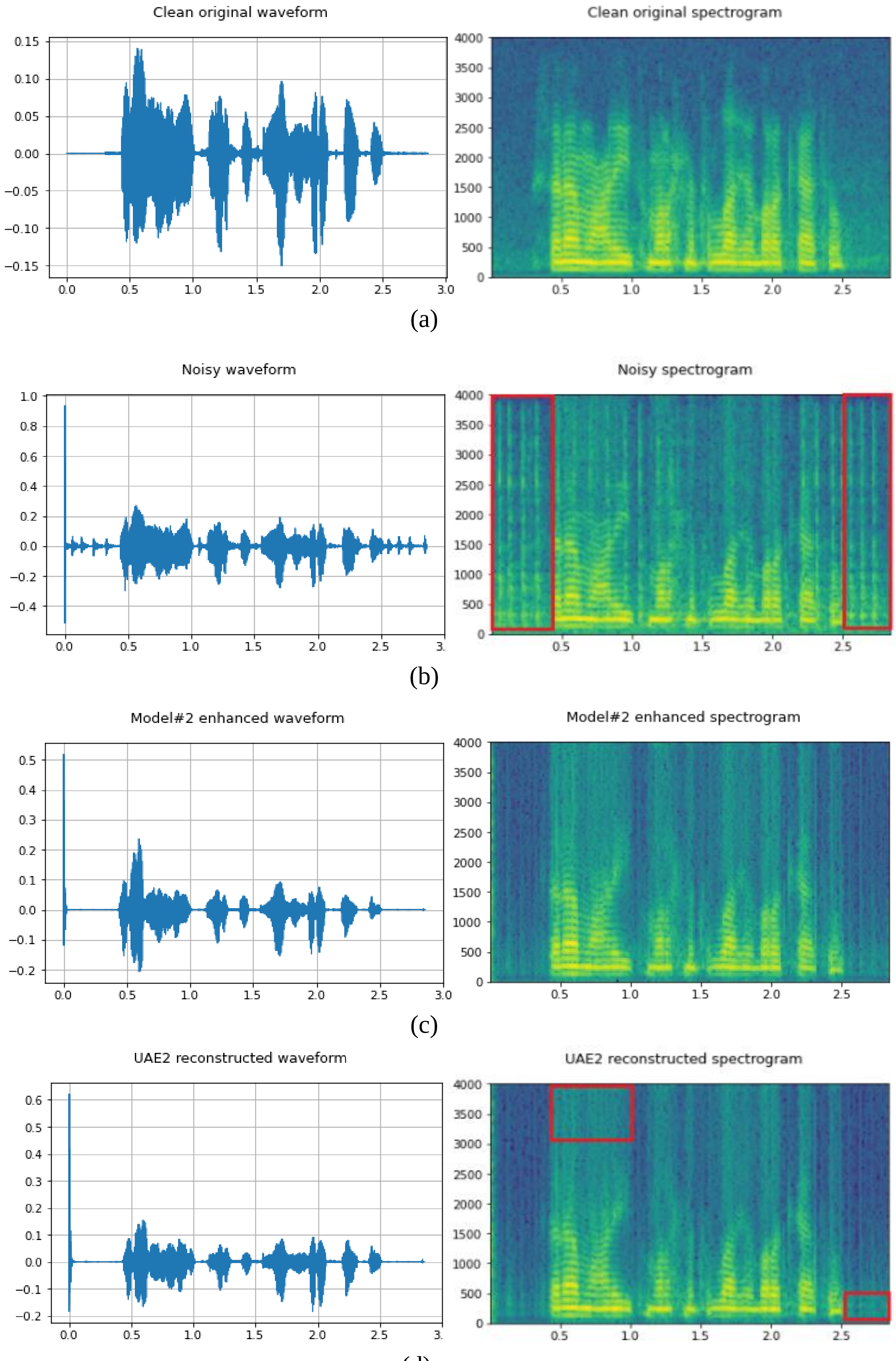
(b) Crowd

(c) Moving Door

Figure 4.6 PESQ and STOI at different SNR levels for various SE methods with multiple noises types

4.5 Spectrogram and Waveform Analysis

We present a visual comparison of spectrograms and waveforms for SE results achieved using Model#2, UAE2, and FCN-based models. Figure 4.7 shows the spectrograms and waveforms of the clean utterance, its noisy version, and the corresponding enhanced utterance using the above models. We can observe that Model#2 produces a denoised version of the spectrogram that is very close to the original clean spectrogram. As depicted in Figure 4.7, 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 the rectangular red regions.

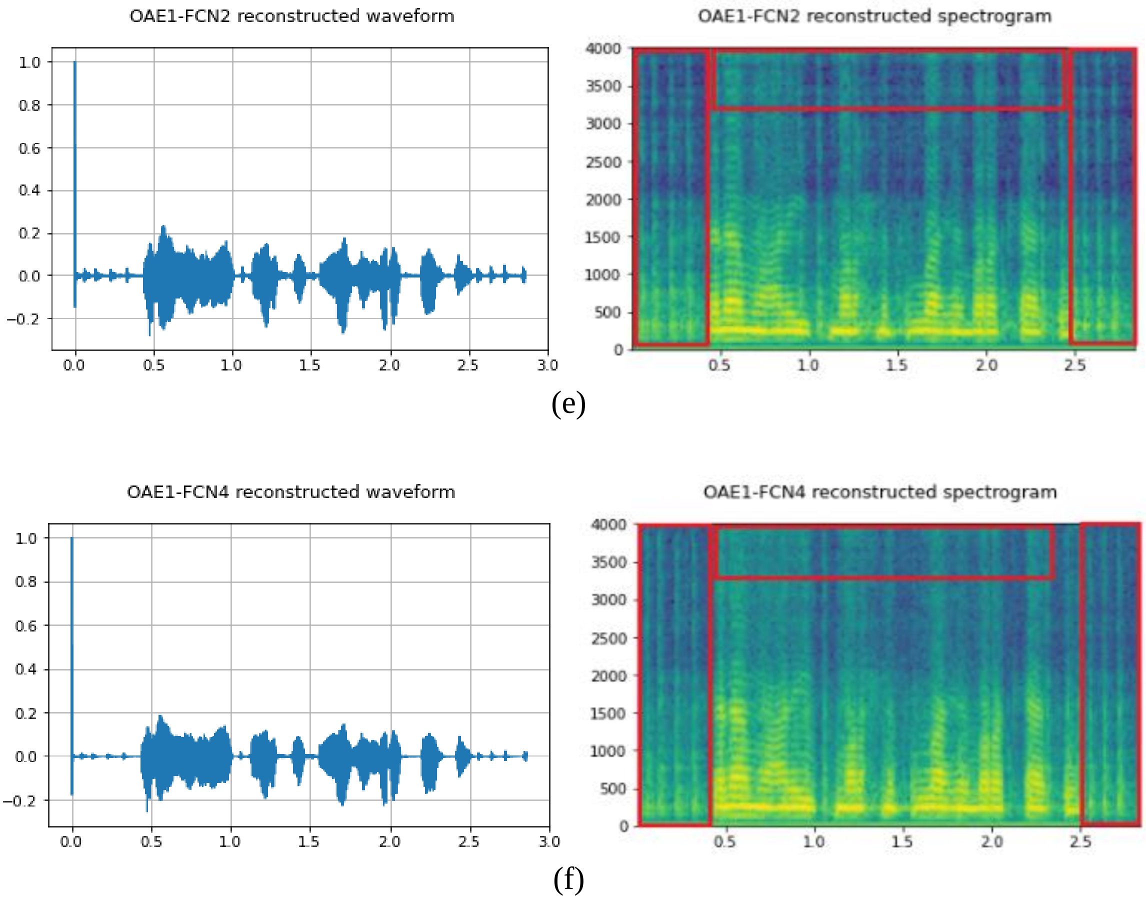

Clean original waveform
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Clean original spectrogram
4000 3500 3000 2500 2000 1500 1000 500 0
0.5 1.0 1.5 2.0 2.5
Noisy waveform
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4
0.0 0.5 1.0 1.5 2.0 2.5 3.
Noisy spectrogram
4000 3500 3000 2500 2000 1500 1000 500 0
0.5 1.0 1.5 2.0 2.5
Model#2 enhanced waveform
0.5 0.4 0.3 0.2 0.1 0.0 −0.1 −0.2
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Model#2 enhanced spectrogram
4000 3500 3000 2500 2000 1500 1000 500 0
0.5 1.0 1.5 2.0 2.5
UAE2 reconstructed waveform
0.6 0.5 0.4 0.3 0.2 0.1 0.0 −0.1 −0.2
0.0 0.5 1.0 1.5 2.0 2.5 3.
UAE2 reconstructed spectrogram
4000 3500 3000 2500 2000 1500 1000 500 0
0.5 1.0 1.5 2.0 2.5

(a)

(b)

(c)

(d)



(e)

(f)

Figure 4.7 Reconstructed waveforms and spectrograms using different models along with the clean and the noisy versions

It is worth noting that the two steps approach (Model#2) and the UAE2 alone, which are autoencoder-based models, outperformed the other architectures, including the FCN. In particular, using an OAE for unsupervised pre-training provides a solution to generate clean data and allows the supervised training of the classical DDAE.

4.6 Conclusion

This chapter deals with Arabic speech enhancement in a ubiquitous environment that aims to improve the WER score in the end-user ASR applications. For that purpose, a speech enhancement approach is suggested. SE is of paramount interest, despite, it is not an easy task due to the lack of real-life labeled data (clean/noisy pairs). I proposed a two-step approach where an overcomplete deep autoencoder is trained in an unsupervised way to produce the enhanced speech. Next, a denoising deep autoencoder is employed to reconstruct the final enhanced speech signal to be recognized. The achieved results show an improvement of the WER of about 4.48% for the mobile MSA corpus and make the proposed approach an effective alternative to the implementation of ubiquitous robust speech recognition systems. Meanwhile, the proposed model outperforms other models, improving speech quality (PESQ) and intelligibility (STOI) by 0.835 and 0.06, respectively, based on the mobile MSA.

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 challenging topics that need more focus in future works. Meanwhile, the indicator that measures the front-end algorithm and the accuracy of the back-end recognition are not positively correlated. Consequently, the improvement of the front-end may not have a positive effect on the back-end recognition. For instance, the improvement in SE approaches on some objective metrics does not necessarily mean the performance in the WER. We expect that the back-end recognition results will feedback the front-end, which would make the system more robust.

Chapter 5: Mispronunciation Detection in Noisy Environments

5.1 Introduction

In this chapter, we tackle the mispronunciation detection task in a ubiquitous CAPT system, considering the issue of the lack of non-native speech corpora. In this context, several noisy sources challenge the development of the mispronunciation detection task; this is particularly true for the Arabic language. First, to handle the noise problems, the incoming speech to assess is enhanced using the beforementioned DDAE. The second issue is related to the lack of annotated non-native speech that suits the assessment task. To overcome this issue, the automatic pronunciation assessment is considered as a classification problem. The proposed solution explores the deep neural networks trained in an unsupervised manner, using solely correct pronunciations in an end-to-end method (input and output are waveforms). Since the model is trained using the correct pronunciation utterances, the deviant pronunciation outputs are detected and considered as mispronounced input samples. The end-to-end based waveforms approaches are different from the other reconstruction methods based on magnitude spectrum, such as the log power spectrum (LPS). Moreover, the audio augmentation techniques are adopted to improve the models' performances, given the limited amount of training/testing data sets. Experiments on the two corpora: The isolated Arabic speech corpus and the Algerian pupil's data set (Bahi & Necibi, 2020), show how the deep learning-based techniques are effective to assess the learner's pronunciation. We selected the Algerian pupil's corpus since it is the only free corpus that includes the pronunciation assessment.

As depicted in figure 5.1, to detect the mispronunciations in the incoming speech, the input and the output to the deep learning models are raw waveforms. In this proposition, the end-to-end DL models are trained on enhanced correct speech utterances to recover the accurate enhanced versions in an unsupervised mode. It is worth noting that the enhanced version of the speech is obtained after applying the SE proposed task.

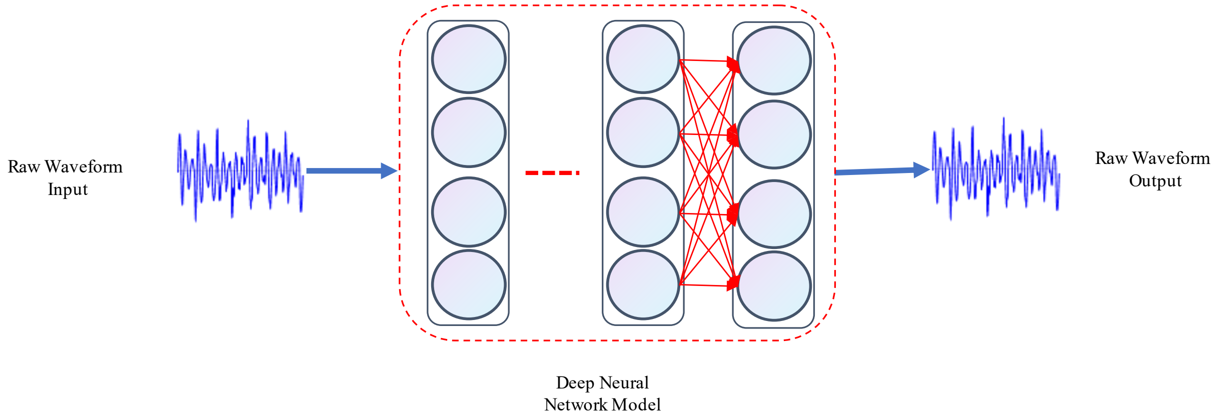


Figure 5.1 End-to-end speech reconstruction based on deep neural networks

5.2 Mispronunciation Detection in Noisy Environments

The proposed system consists of two stages: the speech enhancement and the mispronunciation detection stage. Figure 5.2 depicts the architecture of the proposed MDD. Firstly, the DAE architecture learns features from corrupted data to recover the clean utterances against different noisy environments. Secondly, the mispronunciation detection is performed based on the enhanced speech data obtained from the first stage.

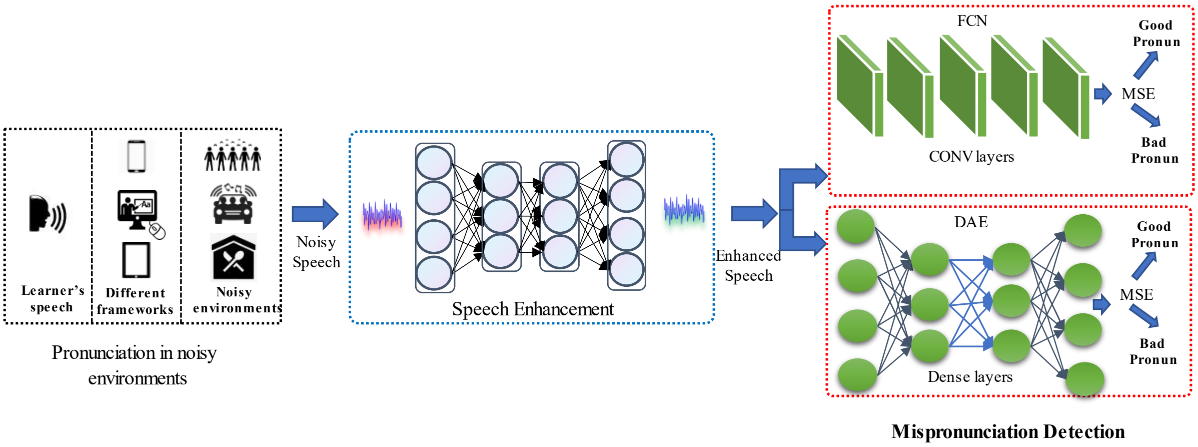


Figure 5.2 The proposed MDD system architecture

5.2.1 Speech enhancement

The denoising DAE is an autoencoder that is trained to predict the enhanced speech data from corrupted utterances. The DAE architecture is based on two sub-networks: the encoder and the decoder networks. In the encoder part: the DAE compresses the corrupted data as possible to remove the noisy information. In the decoder: the clean data is reconstructed from the compressed corrupted data. The results of speech reconstruction are sent to the next stage for MDD.

5.2.2 One-class training for pronunciation assessment

As depicted in figure 5.2, the enhanced speech data recovered from the corrupted data at the noisy environments are fed to the MDD stage. The aim here is to provide Arabic learners with an automated assessment system. To overcome limitations related to the scarcity of ground truth samples, including both “correct” and “incorrect” pronunciations, I developed a system based on a deep learning algorithm trained in an unsupervised way. Deep learning algorithms refer to the use of deep neural networks (DNN) for machine learning purposes. DNN algorithms offer many advantages over alternative classification approaches, such as the ability to detect complex non-linear relationships between the inputs and the outputs (Goodfellow, 2016). Currently, various deep learning architectures are employed. Here, the DAE and the FCN are explored for automatic pronunciation assessment.

5.2.2.1 Deep Auto Encoder model

A Deep Auto-Encoders (DAE) is a DNN architecture intended to learn the submitted pattern’s essence and restitute it in their output layer. This architecture is trained to reconstruct the input X on the output layer Y through one (or more) hidden layer(s), using encoder and decoder parts. Encoders reduce the dimensionality of the input data to represent them in a new space (encoding), while decoders reconstruct the data from the encoding by minimizing the reconstruction error. Figure 5.3 depicts the basic structure of the DAE architecture. The network structure is formulated based on two functions. The encoder function $h = f(x_i)$, and the decoder function $\hat{x}_i = g(h)$ which reconstructs the input vector x_i. The learning process of the DAE allows the structure to capture the most salient representation of the training data by minimizing the loss function (Goodfellow, 2016). The loss function $L(x_i, g(f(x_i)))$ used here is the MSE, which penalizes the dissimilarity between $g\big(f(x_i)\big)$ and the input vector x_i.

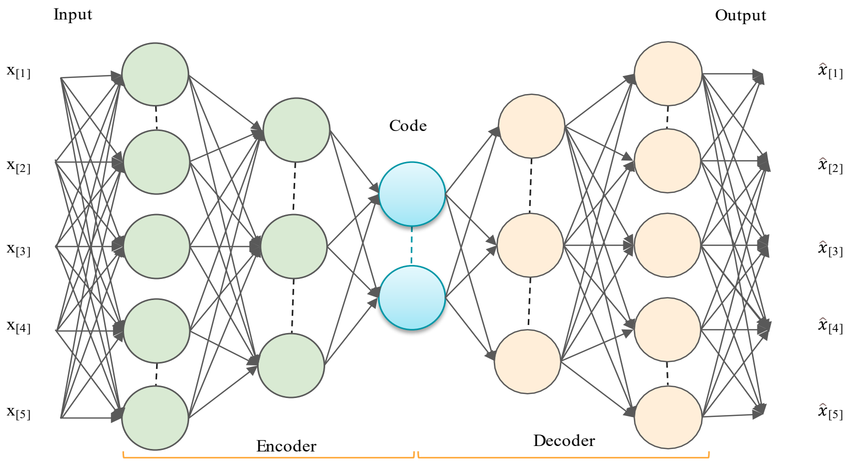
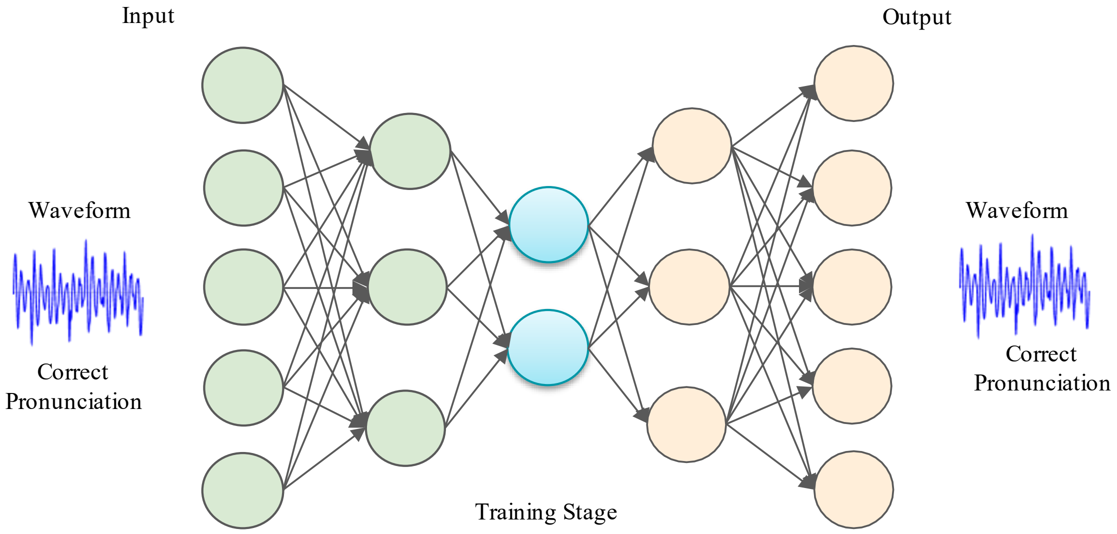
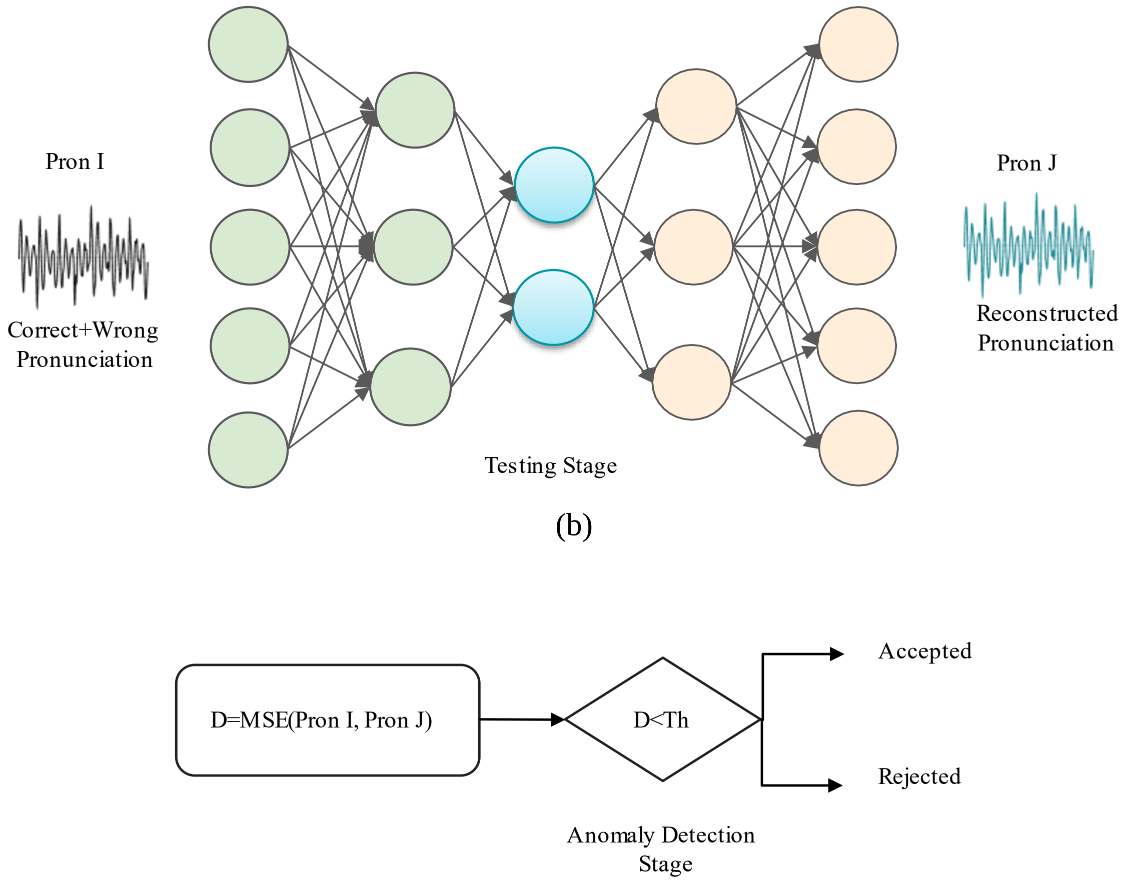


Figure 5.3 A Deep-Auto-Encoder (DAE) composed of five hidden layers

Autoencoders are a particular type of neural networks where the input and the output have the same dimensionality. This architecture is suitable for unsupervised learning since it does not need explicit labels for training. Moreover, AEs try to extract useful features from the correct input speech data. They can be powerful feature detector, as they generate new data that looks very similar to the training data. In the current work, we trained the DAE using only samples labeled as “Accepted” (See training stage figure 5.4a). Consequently, it is expected that, during the test stage, samples reconstructed from their encoding, that have large reconstruction errors would be considered as mispronounced samples (Rejected) (See testing stage: figure 5.4b).



(a)



(b)

(c)

Figure 5.4 Overview of the proposed system

For the classification stage, the mean square error (MSE) criterion was used as a threshold to classify samples (Figure 5.4c).

5.2.2.2 The Fully Convolutional Neural Network (FCNN)

While the fully connected models usually have many parameters, convolutional ones use a fewer number of parameters. CNNs algorithms have gained increasing popularity and are particularly suitable for pattern analysis and feature extraction. In addition, they deal with the local temporal and spectral structure of the speech. To explore DL techniques for mispronunciation detection, we compare the DAE with the CNN architecture. We suggest the usage of the Fully Convolutional Neural Network (FCN), which considers removing both the fully connected layer and the max-pooling layers from the CNN structure. The FCN model learns features from the correct pronunciations in an unsupervised manner. It is clear that the convolution layers extract features and generate the “Accepted” patterns at the output to learn the statistical representation of the correct pronunciations. The FCN model consists of five consecutive cascaded convolution layers that reconstruct the input at the target. The FCN model does not include any pooling layers, as the purpose is to reconstruct an output of the same dimensions as the input. Figure 5.5 depicts the FCN architecture.

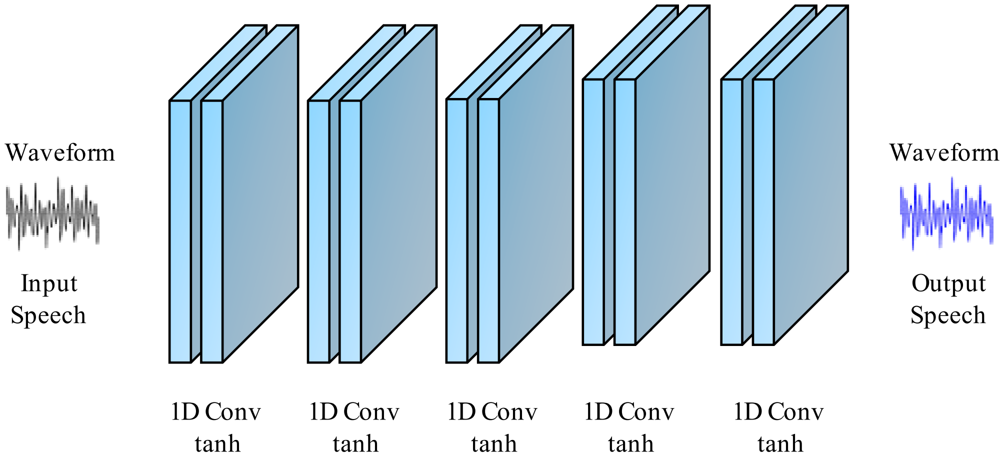


Figure 5.5 An FCN architecture composed of 05 stacked convolutional layers

Three steps are involved for pronunciation assessment, using the FCN model as well as the DAE model.

5.3 Training and Classification

The training dataset includes solely pronunciations rated as “Accepted” by the expert. During the training stage, the input as the output of the DAE, or the FCN model, is fed with accepted pronunciations (Pr. I). Both models are trained in an unsupervised way and exclusively with correct pronunciations and are expected to learn their statistical representation.

As the system is trained on solely correct pronunciations, it is expected to predict the correct version of the input. In the test stage, given one speech pronunciation to assess (Pr. I), the system provides an estimated representation (Pr. J), standing for (the Pr. I) “enhanced” version. The computed distance between (Pr. I) and (Pr. J) stands for the deviation between (Pr. I) and its correct version.

As shown in figure 5.4, the testing stage provides the enhanced version (in terms of pronunciation) for the input speech. The reconstruction error is computed between both signals (input and estimated speech). The decision on the incoming pronunciation is performed by comparing the measured distance and a defined threshold (Th.) (anomaly detection stage). To measure the distance, we computed the Mean Square Error (MSE), the average of squared differences between prediction and actual observation.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \widehat{y_j}\right)^2 \qquad (5.1)$$

Since the errors are squared before they are averaged, the MSE gives a relatively high weight to substantial errors. By penalizing significant errors, the MSE value increases with the variance of the frequency distribution for error magnitudes. Taking MSE as a prediction error to calculate the threshold allows mitigating legitimate detection of deviations in the correct pronunciations.

5.4 Experimental Setup

This section presents the detailed analysis carried out to prove the effectiveness of the proposed assessment approach. For that purpose, two corpora were used. The first corpus includes solely correct pronunciations. A subset of them was used for the training stage, while the remaining (correct) pronunciations and the wrong artificially generated pronunciations were used for the testing stage. The second one is CAPT-dedicated; a subset of the correct pronunciations was artificially expanded and used for the training stage, whereas the remaining ones were used for the test stage.

We have considered two main architectures of DNNs for the experiments, namely a traditional DAE and a fully convolutional network. The various investigated models are reported in table 5.1.

Table 5.1 The various investigated models

Model	Architecture	Observation
DAE - 4	32 – 16 – 16 - 32	Four fully connected layers
DAE - 5	64-32- 16 – 32-64	Five fully connected layers
DAE - 6	64-32- 16 - 16 – 32 - 64	Six fully connected layers
FCN – 4	32 -16 – 16- 32	Four convolution layers
FCN – 5	64-32 – 16- 32-64	Five convolution layers
FCN – 6	64-32 -16 – 16- 32-64	Six convolution layers

All the models used the same activation function, the hyperbolic tangent (tanh), and have the same size for the input and the output.

5.4.1 Datasets

Two datasets are utilized for the different experimentations. First, we considered the Arabic speech corpus for isolated words (Alalshekmubarak & Smith, 2014), which has been developed by the Department of Management Information Systems of King Faisal University. It contains correct 9992 recorded utterances of 50 speakers pronouncing 20 words through mobile devices. The list of the pronounced words is translated as: {Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Activation, Transfer, Balance, Payment, Yes, No, Funding, Data, Account, End}.

Audio data augmentation was used to overcome limitations related to the training dataset size. Several techniques were used to artificially expand the size of a dataset by creating modified versions of signals in the dataset (Kharitonov et al., 2020; Ko et al., 2015). This was done by applying domain-specific techniques to examples from the training data. This operation produced new training examples, thus, new samples that belong to the class “Accepted” pronunciations were generated. For that purpose, we utilized the pitch-shifting method and the time-stretching technique that changes the speed/duration of sounds without affecting their pitch. To modify the speed of a signal, we resample the speech signal into two additional copies of the original training data with speed values of 90% and 110%. Additionally, pitch shifting changes the pitch of sounds without affecting their speed. We applied both techniques to half recordings of the Arabic isolated speech corpus that is augmented twice. Once the speech augmentation is performed, we corrupted the augmented data using the real-world UrbanSound8K dataset (Salamon et al., 2014). The urban sound data set contains 8732 clip sounds. Five environmental noises were considered: the air conditioner, the engine idling, car horn, children playing, and the jackhammer noises. The obtained sounds served to the training stage; finally, the training stage included 18045 “Accepted” noisy pronunciations. For the test stage of MDD, table 5.2 displays the distribution of testing data among three noise types (air conditioner, engine idling, and jackhammer).

Table 5.2 Distribution of the testing data among different noise types

	Noise type		
Accepted/Rejected	Air conditioner	Engine idling	Jackhammer
Accepted	1482	1236	1236
Rejected	959	959	959

The “Rejected” pronunciations are artificially recordings generated owing to audio augmentation techniques (Kharitonov et al., 2020). The generated pronunciations were designed to represent the deletion and insertion of phonemes, which are the common detected errors. After data preparation, the enhancement stage starts before the MDD, using the denoising DAE. The results of the SE achieved using the denoising DAE are fed to both models (DAE and FCN) to detect mispronunciation.

The second corpus described in (Bahi & Necibi, 2020), contains pronunciations from nine pupils, aging from 5 to 8 years; each of them uttered a set of 16 sequences (words or group of words). The chosen words included some difficulties for the young learners, such as the long vowels and the words written with more than one connected component (table 5.3).

Table 5.3 List of the pronounced sequences with their transcriptions (Bahi & Necibi, 2020)

#	Sequences in Arabic	Phonetic transcription	Translation	#	Sequences in Arabic	Phonetic transcription	Translation
1	صباح الخير	s`aba:ħu ʔalxajr	Good Morning	9	مسّن	Mussin	Aged
2	إلى اللقاء	ʔila ʔalliqa:ʔ	Good bye	10	متأخر	mutaʔaxir	Late
3	ليلة سعيدة	lajlatun saʕi:datun	Happy Night	11	فارغ	fa:riʁ	Empty
4	من فضلك	min fad`lik	Please	12	ثقيل	θaqi:l	Heavy
5	شكرا	ʃukran	Thanks	13	أسفل	ʔasfal	Down
6	جميل	ʤami:l	Beautiful	14	داخل	da:xil	Inside
7	قبيح	qabi: ħ	Ugly	15	بداخل	bida:xil	Inside of

8 قريب	qari:b	Near (close)	16 خارج	xa:ridʒ	Outside

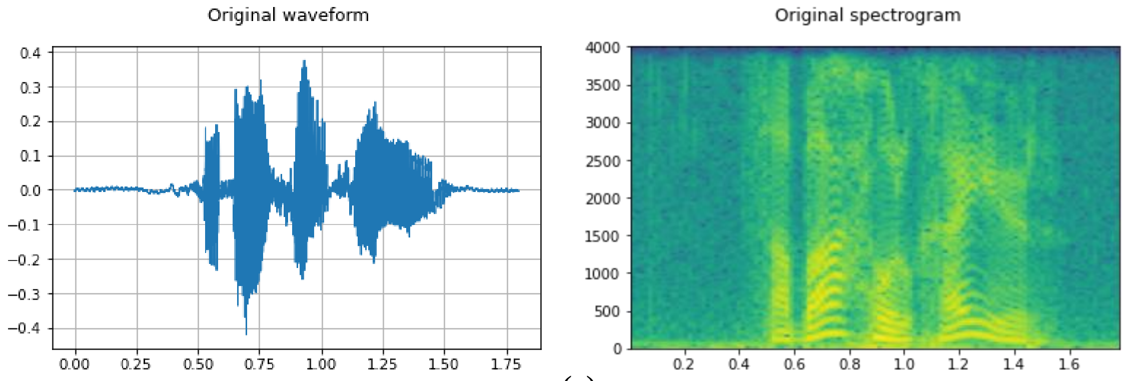
Among the pronunciations rated as “Accepted” by the expert, a small subset is included in the training dataset whereas, the remaining and “Rejected” ones are incorporated in the testing set. In particular, all utterances of speaker9 are excluded from the training set.

5.4.2 Data Augmentation to improve the performance of models

Data Augmentation refers to many techniques used for expanding the training data examples. Several methods have been conducted to explore the field in different directions for increasing interest in low resources languages, scalable large neural networks, and different tasks for machine learning (Feng et al., 2021). These techniques are commonly applied for multimodal tasks, including speech, image, and video. Some examples of these methods include Gaussian noise injection, time-stretching, and image rotations.

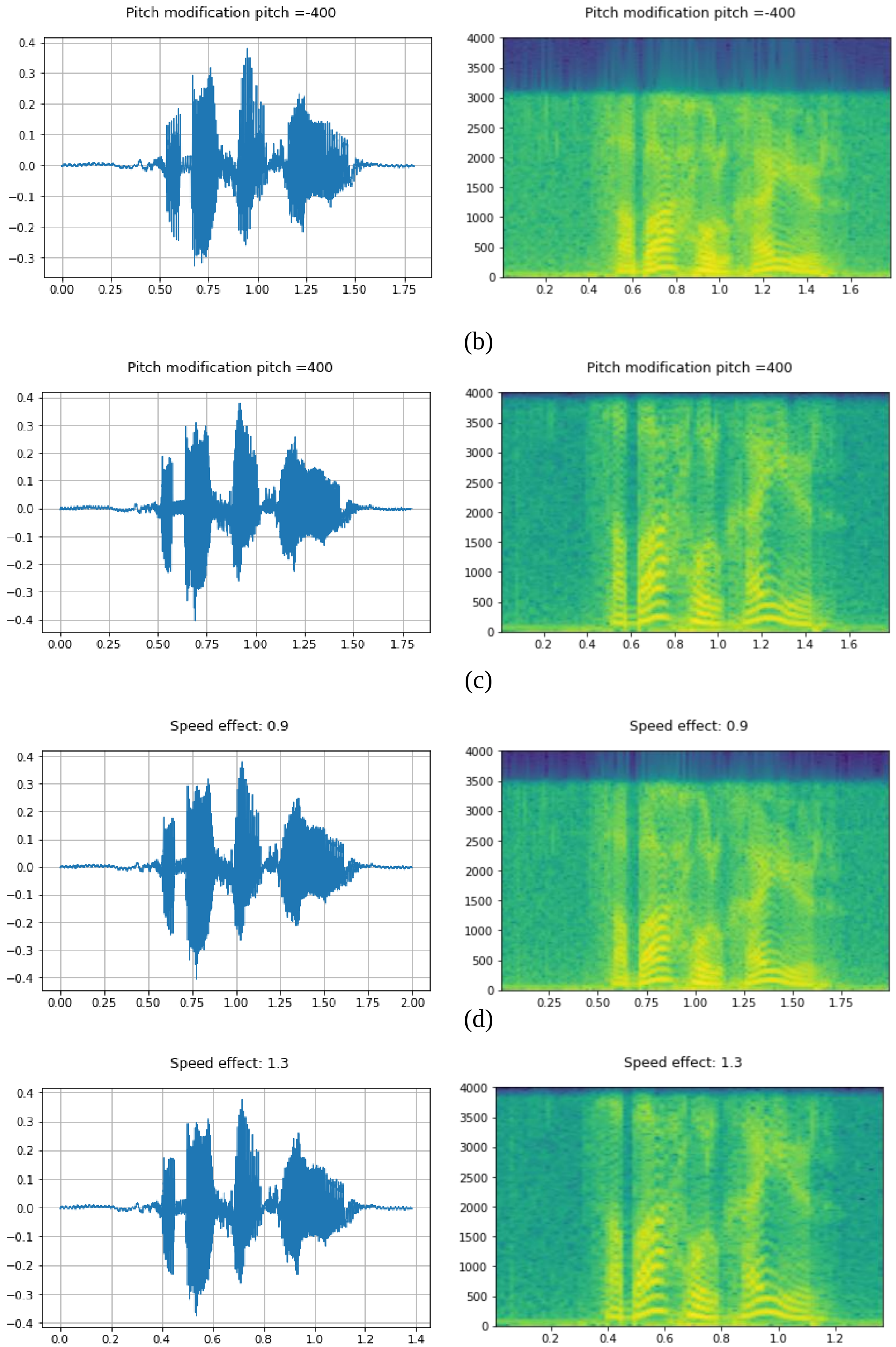
Many augmentations techniques have been proposed for speech tasks, such as, SpechAugment (D. S. Park et al., 2019) that deforms the input log-Mel spectrogram to improve the ASR performance. Furthermore, wave augmentation (Kharitonov et al., 2020) based time-domain library provides various techniques for speech data augmentation. We employed wave augmentation to overcome limitations in native and non-native speech data, in order to make models generalize better and to improve their performance on more unseen data.

Figure 5.6 depicts some examples of applying speech augmentation techniques to the sentence "صباح الخير". The waveforms and the spectrograms are generated using the [3]WavAugment library.



(a)

[3] https://github.com/facebookresearch/WavAugment

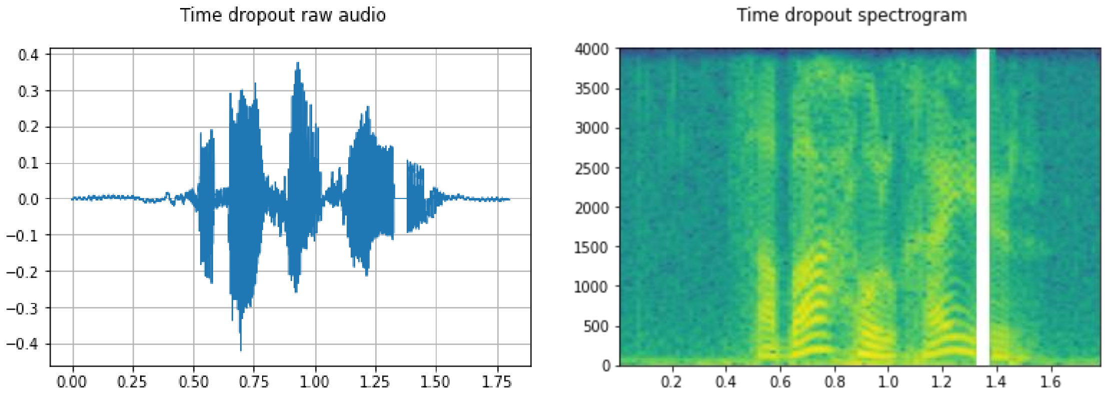

Pitch modification pitch =-400
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Pitch modification pitch =-400
4000
3500
3000
2500
2000
1500
1000
500
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Pitch modification pitch =400
−0.4
Pitch modification pitch =400
Speed effect: 0.9
2.00
Speed effect: 0.9
Speed effect: 1.3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Speed effect: 1.3

(b)

(c)

(d)

(e)



(f)

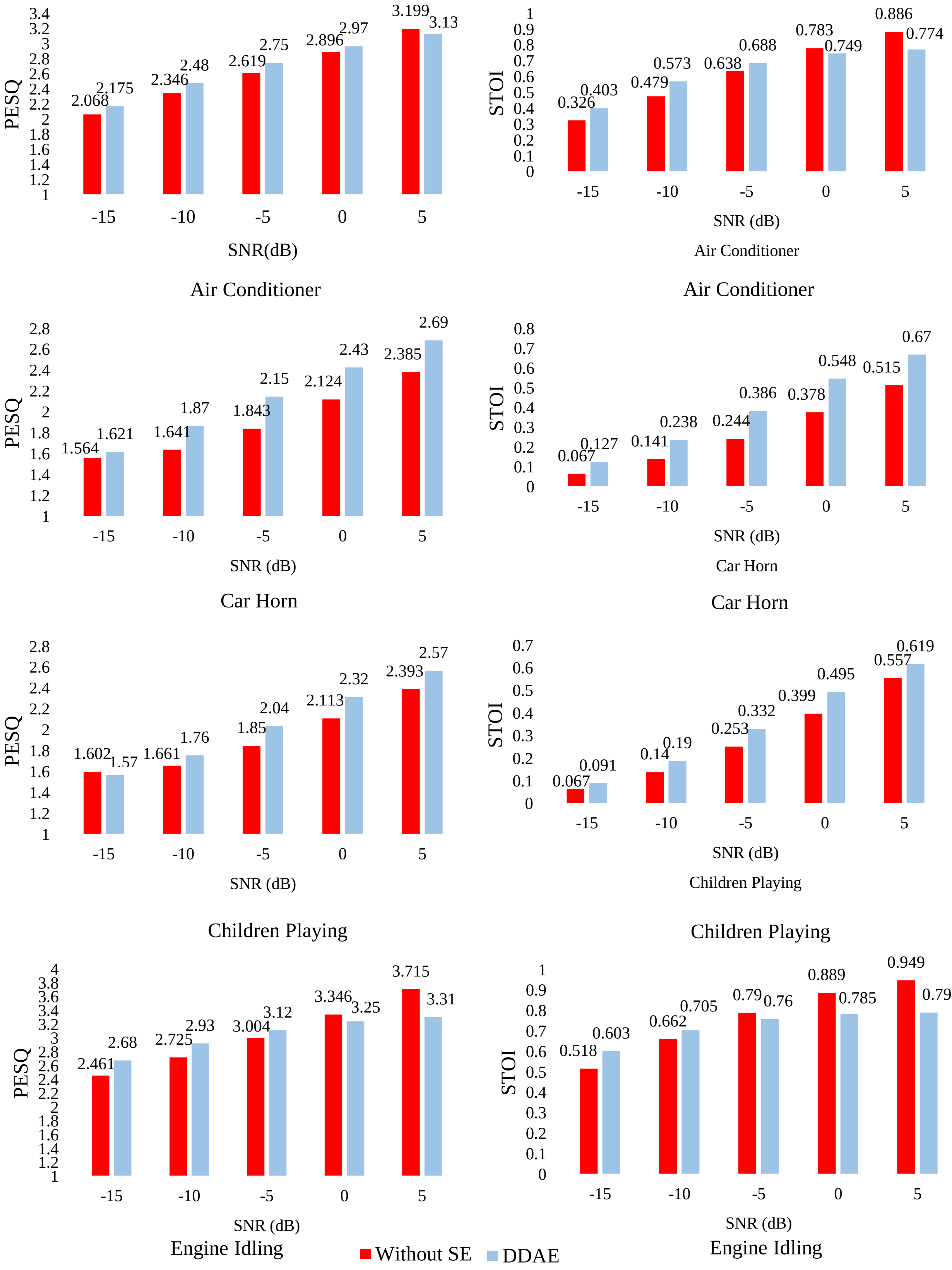
Figure 5.6 Augmented waveform and spectrogram for the sentence “صباح الخير”

5.5 Experiments and Results

The first experiments explored the capability of DDAE for SE under different noisy environments and at varying SNR levels (-15 dB, -10 dB, -5 dB, 0dB, and 5 dB). A second experiment investigated the capability of both models DAE, and the FCN to detect anomalies based on the enhanced raw waveforms. These experiments were firstly carried on the Arabic isolated speech corpus. Afterward, the augmented training data are considered to improve the performance of models. Furthermore, the effect of the model's depth on the performance results is analyzed to compare both architectures' depths in terms of accuracy (ACC) and FRR. Additional experiments were carried out on the second corpus to confirm the capability of the proposed approach to distinguish abnormal from accepted pronunciations.

5.5.1 Speech enhancement based on the DDAE

The DDAE SE architecture applied for SE consists of 257 neurons used for the input as well as the output layers. Two hidden layers represent its structure, each with 200 neurons DDAE (200, 200). The loss function is the MSE between the estimated log magnitude and the log magnitude of the clean original signal. The learning rate used for the fine-tuning phase is 0.0001. The effectiveness of the DDAE is validated using both SE metrics, speech quality (PESQ), and intelligibility (STOI). Figure 5.7 reports the performance results for the noisy and the DDAE model in terms of PESQ and STOI.


PESQ
3.4
3.2
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
2.068
2.175
2.346
2.48
2.619
2.75
2.896
2.97
3.199
3.13
-15
-10
-5
0
5
SNR(dB)
Air Conditioner
STOI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.326
0.403
0.479
0.573
0.638
0.688
0.783
0.749
0.886
0.774
-15
-10
-5
0
5
SNR (dB)
Air Conditioner
Air Conditioner
PESQ
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
1.564
1.621
1.641
1.87
1.843
2.15
2.124
2.43
2.385
2.69
-15
-10
-5
0
5
SNR (dB)
Car Horn
STOI
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.067
0.127
0.141
0.238
0.244
0.386
0.378
0.548
0.515
0.67
-15
-10
-5
0
5
SNR (dB)
Car Horn
Car Horn
PESQ
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
1.602
1.57
1.661
1.76
1.85
2.04
2.113
2.32
2.393
2.57
-15
-10
-5
0
5
SNR (dB)
Children Playing
STOI
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.067
0.091
0.14
0.19
0.253
0.332
0.399
0.495
0.557
0.619
-15
-10
-5
0
5
SNR (dB)
Children Playing
Children Playing
PESQ
4
3.8
3.6
3.4
3.2
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
2.461
2.68
2.725
2.93
3.004
3.12
3.346
3.25
3.715
3.31
-15
-10
-5
0
5
SNR (dB)
Engine Idling
STOI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.518
0.603
0.662
0.705
0.79
0.76
0.889
0.785
0.949
0.792
-15
-10
-5
0
5
SNR (dB)
Engine Idling
■ Without SE
■ DDAE

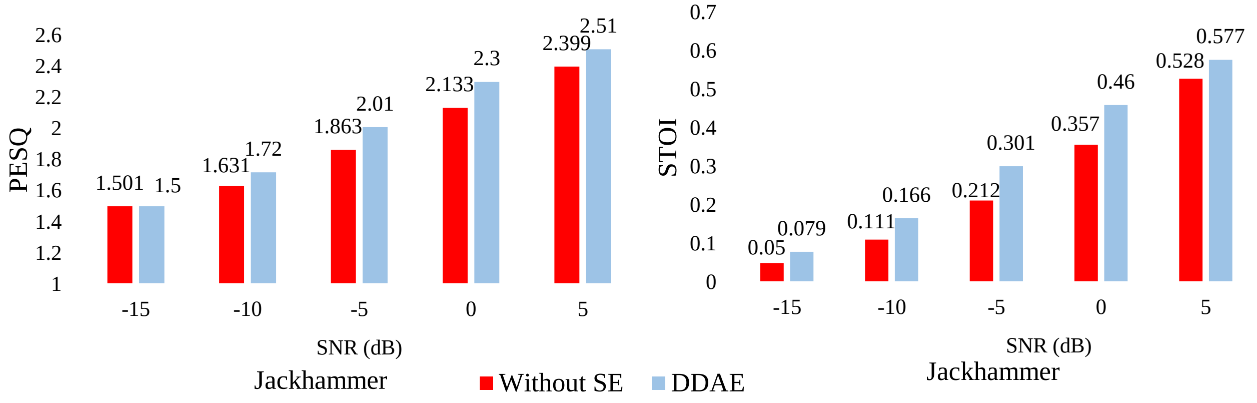


Figure 5.7 PESQ and STOI for DDAE-based SE method, with multiple noises types and at different SNR levels.

5.5.2 Mispronunciation detection based on the enhanced speech

Once the DDAE SE was conducted in different noisy environments at varied SNR levels, the MDD second stage was performed to assess enhanced speech using DAE and FCN architectures. The noisy environments considered in the second stage are air-conditioner, engine idling, and the jackhammer noises, at 0 dB SNR level. Table 5.4 reports the results obtained from the evaluation of the MDD system, in terms of accuracy (ACC) and FRR, without applying the SE previous stage.

Table 5.4 Performances of noisy environments for MDD, using DAE and FCN with waveform inputs

Metrics (%)	Model	Noise type			AVG (%)	Global AVG (%)
	Model	Air Conditioner	Engine Idling	Jackhammer		
ACC	DAE-4	68.04	66.01	59.08	64.37	64.74
	DAE-5	68.04	66.01	59.04	64.36	
	DAE-6	68	66.01	60	64.67	
	FCN-4	69.02	68.01	60	65.67	
	FCN-5	61.16	66.60	61	62.92	
	FCN-6	70.21	67.10	62.05	**66.45**	
FRR	DAE-4	22.26	27.5	37.86	29.20	29.66
	DAE-5	17.54	23.46	35.27	25.42	
	DAE-6	17.67	29.69	36.56	27.97	
	FCN-4	24.69	27.34	21.92	**24.65**	
	FCN-5	54.52	25.64	37.45	39.20	
	FCN-6	29.01	33	32.68	31.56	

Table 5.5 Performances of enhanced speech for the MDD based on DAE and FCN using waveform inputs

Metrics (%)	Model	Noise type			AVG (%)	Global AVG (%)
	Model	Air Conditioner	Engine Idling	Jackhammer		
ACC	DAE-4	72.05	70.11	71.04	**71.06**	**70.28**
	DAE-5	71.3	68.04	69.68	69.67	
	DAE-6	70.04	68.04	69.14	69.07	
	FCN-4	71.03	69.75	71.61	70.69	
	FCN-5	71.01	69.01	70.5	70.17	
	FCN-6	71.54	70.33	71.32	**71.06**	
FRR	DAE-4	7	5.33	8.18	6.83	**8.16**
	DAE-5	5.46	10.40	3.20	6.35	
	DAE-6	5.3	4.78	8.92	**6.33**	
	FCN-4	9.42	13.94	9.61	10.99	
	FCN-5	10.01	4.73	11.71	8.81	
	FCN-6	6.75	12.45	9.75	9.65	

5.5.3 Data augmentation effect

The aim here is to confirm the capability of the proposed one-class objective training using the speech data augmentation techniques. Audio speech augmentation techniques are highlighted, as they allowed to improve the performance of the model. In this experiment, we suppose to use the enhanced speech obtained after the SE stage.

We used the Arabic isolated speech corpus that includes solely accepted pronunciations. This dataset is split into two sets, the training set, and the testing set. From a part of the test set, we generated abnormal pronunciations standing for the "Rejected" pronunciation and the remainder testing utterances as an accepted pronunciation. Table 5.6 reports the results in terms of the accuracy and the FRR for all the models:

Table 5.6 Performances of the DAE and FCN with different depths using waveform inputs

	Without data speech augmentation											
	Training Dataset						Test Dataset					
Model	DAE-4	DAE-5	DAE-6	FCN-4	FCN-5	FCN-6	DAE-4	DAE-5	DAE-6	FCN-4	FCN-5	FCN-6
Acc.	91.86	94.61	82.83	96.37	96.62	97.79	83	83.10	82.83	77.61	83.13	87.03
FRR	8.13	5.38	5.36	3.62	3.37	2.20	6.4	5.79	5.36	20.12	15.3	5.48
	Using Data augmentation											
Acc.	98.37	97.97	98.38	98.31	97.95	98.28	90	81.09	82.73	86.79	85.01	87.54
FRR	1.62	2.02	1.61	1.68	2.04	1.71	8.78	10.91	7.80	15.36	16.34	6.7

5.5.4 Additional experiments

These experiments aim to confirm the capability of the proposed one-class objective training approach in pronunciation assessment, using DAE and FCN models. We experiment with a second CAPT-dedicated speech corpus that includes "Accepted" and "Rejected" pronunciations. Figure 5.8 reports the performance comparison of accuracy and FRR for both architectures DAE and FCN using waveforms.

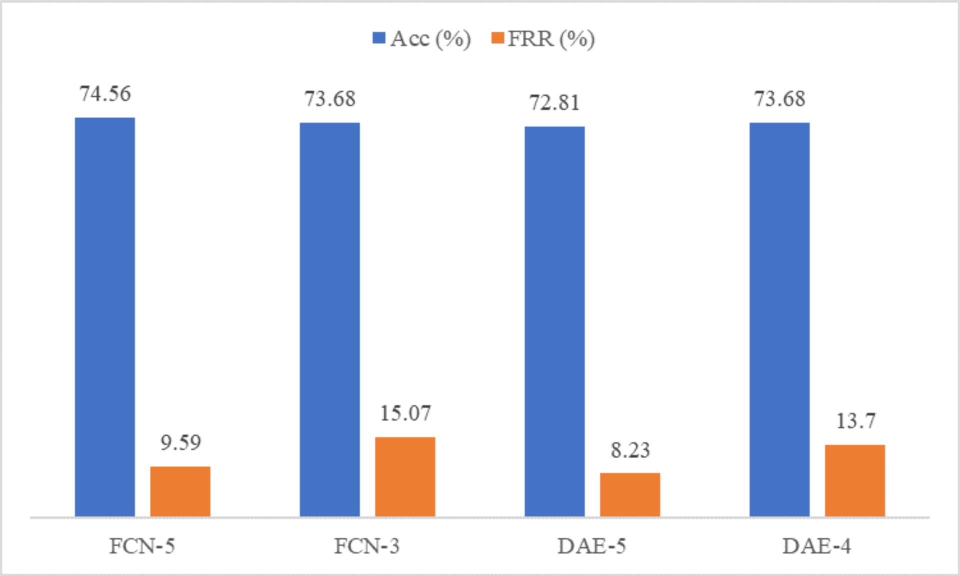


Figure 5.8 Performance comparison of the proposed approach for both models DAE and FCN with waveform inputs

As an illustration of the previous comparison in figure 5.8, we reported the evaluations of sequences pronounced by speaker9 in table 5.7 for both deep learning techniques.

Table 5.7 Illustration on Speaker9 of the various models' assessments

#Seq	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Expert	1	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0
FCN-5	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0
FCN-3	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0
DAE-5	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0
DAE-4	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0

5.6 Discussion

Experimental results on the DDAE-based SE show the effectiveness of the proposed approach. From figure 5.7, the DDAE outperformed the noisy speech (labeled by Without SE) at different environments in terms of the PESQ and STOI. The average improvements of the DDAE over different noisy environments at different SNR levels are 0.11 and 0.04 for PESQ and STOI, respectively. Once the SE was performed, the MDD was carried out. A comparison of the two tables' results (tables 5.4 and 5.5) reveals the effectiveness of SE for MDD over MDD in noisy environments. The average ACC and FRR for MDD after SE outperformed that without SE. After SE, the ACC achieved an average improvement of 5.54%, and the FRR average obtained is 8.16%.

On the other hand, MDD experimental results with the two corpora confirmed the potency of the proposed approach to distinguish between correct and incorrect pronunciations, even when the training dataset does not include rejected speech pronunciations. Further analysis of the data augmentation usage for the first corpus reveals the positive impact of speech augmentation techniques to expand the training dataset (Table 5.6). Moreover, Audio data augmentation allows large-scale tests by generating deviant pronunciations in the absence of dedicated CAPT corpora.

From tables 5.3 and 5.6, the small size of the training dataset can yield overfitting. Thus, applying audio augmentation techniques highlights the potency to expand training data and to improve the performance results (table 5.6). From table 5.6, it is worth noting that the average accuracy reaches 98.24 % and 98.18 % using DAE and FCN, respectively. The FRR values corresponding to DAE and FCN are 1.75 and 1.81, respectively. Using the second corpus, figure 5.8 depicts the performance comparison between DAE and FCN without performing the augmentation techniques. The Algerian pupils whose L1 language is Arabic confuse the pronunciation between similar phonemes, such as /θ/ (in seq. 12) which is commonly pronounced /t/ in Algeria. Even in this case, the assessment model reaches an average accuracy of 74.12% and 73.24% for FCN and DAE, respectively. In addition, the FRR is still encouraging for young pupils.

Finally, authors (Bahi & Necibi, 2020) have reported accuracy of about 62.5% for speaker9, and overall accuracy of 61.8%. In this work, the accuracy achieved for speaker9 is about 68.7% using the second corpus.

5.7 Conclusion

In this chapter, we tackled the problem of pronunciation assessment with the lack of nonnative datasets and in challenging noisy environments. First, we investigated the capability of DDAE to enhance speech in different noisy environments at various SNR levels. The results obtained in the first stage of SE provide better speech quality and intelligibility than noisy speech. Once the SE is conducted, the MDD using two algorithms (DAE and FCN) is performed, secondly. The results obtained from the MDD stage show the potency of the one class-objective training using DAE and FCN models that provide encouraging accuracy. In addition, the FRR rate encourages learners to pursue their learning. Moreover, we propose to use data augmentation techniques to expand the training dataset and enable large-scale experiments. Experimental results show the effectiveness of data augmentation techniques and their capability to generate deviant pronunciation.

On the other hand, deep learning techniques may not have been fully explored for CAPT due to the scarcity of non-native CAPT dedicated speech corpora. Hence, we propose to train solely deep learning algorithms on correct pronunciations to overcome the limited amount of non-native Arabic corpora.

Chapter 6: Conclusion and Future Work

6.1 Summary and Contributions

In this chapter, we summarize the contributions of this thesis in two points and discuss the future works.

This thesis deals with pronunciation learning in a ubiquitous environment. In particular, it addresses the Arabic pronunciation assessment in a ubiquitous CAPT system and refers to whether a fragment of speech was correctly pronounced or not. Ubiquitous technology can promote learner independence, or the capacity to control one's learning, and is rapidly gaining popularity as an effective way to improve foreign language skills, such as pronunciation skills.

One of the greatest issues in Arabic pronunciation learning is the lack of dedicated speech corpora. The thesis proposed two methods to overcome the lack of dedicated corpus for Arabic pronunciation learning in a ubiquitous environment. The first one lies in speech recognition in real-world environments, while the second one lies in mispronunciation detection.

6.1.1 Self-supervised speech enhancement

The approach used for this work, relying on unsupervised learning, is based on two deep autoencoders. The first one is an overcomplete autoencoder trained in an unsupervised way and aims to generate a clean version of the noisy transcoded speech. The second one is a denoising autoencoder that leverages the clean version, produced by the overcomplete autoencoder and is trained in a supervised manner.

We experimented with different configurations to implement the enhancement system. The experimental results highlighted the ability of the overcomplete deep autoencoders to discover relationships among the training data by mapping them in higher dimensions.

6.1.2 Mispronunciation detection under corpus scarcity conditions

As good pronunciations are more likely to be available than wrong ones, the thesis aims to tackle this issue alongside the scarcity of speech corpus dedicated to pronunciation assessment. For that purpose, the thesis presents two solutions. First, data augmentation techniques were helpful to expand the speech dataset. We used adequate techniques such as pitch modification and time-stretching to extend available datasets. Secondly, considering the lack of non-native Arabic speech corpus, we propose to detect mispronunciations according to the one-class objective training approach. Herein, the models are trained in an unsupervised way using solely correct pronunciations, and they are expected to detect wrong utterances during the test. For that purpose, a deep autoencoder and a fully convolution neural network were proposed.

6.2 Future Works

There are several possible directions for future research to extend current works. Moreover, some remaining open questions are worthy of further investigation.

6.2.1 Investigating OCAE

The results provided within speech enhancement promote the usage of the overcomplete DAE model that brings new perspectives in unsupervised learning. As future work, we intend to explore the capability of DAE-based architectures. For instance, explore overcomplete convolutional autoencoder (OCAE) for unsupervised speech enhancement and mispronunciation detection topics.

6.2.2 Toward a complete CAPT framework

It is appreciable that great opportunities are offered to Arabic, although research is still in its infancy and faces many problems. Altogether, many components required for such applications already exist. However, one of the biggest remaining challenges is to combine these many components into one that ideally is L1-independent, or at least easily configured for a different L1, without requiring a manually annotated non-native database.

6.2.3 Beyond the assessment, the feedback

As the tendency is personalized learning, future applications would pay more attention to integrating an intelligent virtual tutor that takes the role of a private tutor for the student; this is the case, in a real-life situation in learning Quran recitation, for example.

6.2.4 More focus on the front-end part: learner to the CAPT system interaction

The main aim of this thesis is to implement a ubiquitous framework for pronunciation learning, as described in figure 1.1. This framework contains two parts, the front-end, and back-end parts. We detailed two contributions: the unsupervised SE and mispronunciation detection from the back-end part. While the back-end is more discussed, the front-end needs more focus to investigate the learner to CAPT system interaction. For that purpose, some ideas and possible future work are:

- Enrich the interaction between the learner and the system, thus for example by attracting students using game-based reading and practice interactions.
- Personalized learning scenarios: personalization allows the students to track their learning performance.
- Mobile learning, ubiquitous learning, and pervasive learning.
- Allowing students to practice the language in the free form of speech. Spontaneous speech is a challenging topic.
- Multimodal feedbacks

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Steiner, B., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. www.tensorflow.org.

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *22*(10), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

Abdou, S. M., Hamid, S. E., Rashwan, M., Samir, A., Abd-Elhamid, O., Shahin, M., & Nazih, W. (2006). Computer aided pronunciation learning system using speech recognition techniques. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2*, 849–852.

Al-Marri, M., Raafat, H., Abdallah, M., Abdou, S., & Rashwan, M. (2018). Computer Aided Qur’an Pronunciation using DNN. *Journal of Intelligent and Fuzzy Systems*, *34*(5), 3257–3271. https://doi.org/10.3233/JIFS-169508

Al Hindi, A., Alsulaiman, M., Muhammad, G., & Al-Kahtani, S. (2014). Automatic pronunciation error detection of nonnative Arabic Speech. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2014*, 190–197. https://doi.org/10.1109/AICCSA.2014.7073198

Alalshekmubarak, A., & Smith, L. S. (2014). On improving the classification capability of reservoir computing for arabic speech recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8681 LNCS*, 225–232. https://doi.org/10.1007/978-3-319-11179-7_29

Alamdari, N., Azarang, A., & Kehtarnavaz, N. (2019). *Self-Supervised Deep Learning-Based Speech Denoising*. *1*. http://arxiv.org/abs/1904.12069

Almeman, K. (2018). The Building and Evaluation of a Mobile Parallel Multi-Dialect Speech Corpus for Arabic. *Procedia Computer Science*, *142*(2017), 166–173. https://doi.org/10.1016/j.procs.2018.10.472

Almeman, K., Lee, M., & Almiman, A. A. (2013). Multi dialect Arabic speech parallel corpora. *2013 1st International Conference on Communications, Signal Processing and Their Applications, ICCSPA 2013*. https://doi.org/10.1109/ICCSPA.2013.6487288

Alsabaan, M., & Ramsay, A. (2014). *Diagnostic CALL tool for Arabic learners*. *2014*, 6–11. https://doi.org/10.14705/rpnet.2014.000186

Alsunaidi, N., Based, L., & Altassan, M. (2018). ScienceDirect ScienceDirect ScienceDirect Abjad : Towards Interactive Learning Approach to Arabic Reading Abjad : Towards Interactive Learning Approach to Arabic Reading Based on Speech Recognition on Speech Recognition The 4th International Conference on. *Procedia Computer Science*, *142*, 198–205. https://doi.org/10.1016/j.procs.2018.10.476

Andersen, S., Duric, A., Astrom, H., Hagen, R., Kleijn, W., & Linden, J. (2004). Internet low bit rate codec (iLBC). IETF RFC3951. Retrieved September 7, 2018, from https://tools.ietf.org/html/rfc3951

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. The journal of the acoustical society of America, 50(2B), 637-655.Bahi, H., & Necibi, K. (2020). Fuzzy logic applied for pronunciation assessment. *International Journal of Computer-Assisted Language Learning and Teaching*, *10*(1), 60–72. https://doi.org/10.4018/IJCALLT.2020010105

Bahi, H., and Necibi, K. (2020). Fuzzy Logic Applied for Pronunciation Assessment. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 10(1), 60-72.

Bahi, H., & Sellami, M. (2005). An ASR based tool to detect dyslexia, Proceedings of the International Symposium of Programming Systems, Algeries, Algeria, pp. 117-122.

Bahi, H., & Sellami, M. (2001). Combination of vector quantization and hidden Markov models for Arabic speech recognition. *Proceedings ACS/IEEE International Conference on Computer Systems and Applications*, 96–100. https://doi.org/10.1109/AICCSA.2001.933957

Benesty, J., Makino, S., & Chen, J. (2005). *Speech Enhancement*. Springer. https://www.springer.com/gp/book/9783540240396

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic Evaluation and Training in English Pronunciation. *Proceedings of ICSLP 90*, 1185–1188.

Boll, S. F. (1979). Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209

Brandenburg, K. (1999). MP3 and AAC explained. In Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding. Audio Engineering Society.

Campillos Llanos, L. (2014). A Spanish oral learner corpus for computer-aided error analysis. Corpora, 9:2, 207–238.

Chen, L., Tao, J., Ghaffarzadegan, S., & Qian, Y. (2018). End-to-end neural network based automated speech scoring. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 6234–6238. https://doi.org/10.1109/ICASSP.2018.8462562

Chen, L. Y., & Jang, J. S. R. (2015). Automatic Pronunciation Scoring with Score Combination by Learning to Rank and Class-Normalized DP-Based Quantization. *IEEE Transactions on Audio, Speech and Language Processing*, *23*(11), 1737–1749. https://doi.org/10.1109/TASLP.2015.2449089

Chen, N. F., & Li, H. (2017). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*. https://doi.org/10.1109/APSIPA.2016.7820782

Chen, N. F., Tong, R., Wee, D., Lee, P., Ma, B., Li H.(2015). iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent, InterSpeech 2015, Dresden, Germany.

Chen, X., & Cheng, J. (2014). Deep neural network acoustic modeling for native and non-native Mandarin speech recognition. *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, ISCSLP 2014*, 6–9. https://doi.org/10.1109/ISCSLP.2014.6936617

Cheng, S., Liu, Z., Li, L., Tang, Z., Wang, D., & Zheng, T. F. (2020). ASR-Free Pronunciation Assessment, InterSpeech2020, Shangai, China.

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication*, *73*, 14–27. https://doi.org/10.1016/j.specom.2015.07.006

Cheng, J., Bernstein, J., Pado, U., & Suzuki, M. (2009). *Automatic assessment of spoken modern standard Arabic*. *June*, 1–9. https://doi.org/10.3115/1609843.1609844

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

Chu, W. C. (2003). ALGORITHMS SPEECH CODING Foundation and Evolution. In *John Wiley & Sons, Inc.* John Wiley & Sons.

Coalson J. (2001). *FLAC - Free Lossless Audio Codec*. Retrieved September 6, 2018, from https://xiph.org/flac/

Google Cloud Speech API, Speech-to-Text: Automatic Speech Recognition, Google Cloud. Retrieved July 21, 2021, from https://cloud.google.com/speech-to-text

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873. https://doi.org/10.1121/1.1471894

Dendani, B., Bahi, H., & Sari, T. (2021). Self-Supervised Speech Enhancement for Arabic Speech Recognition in Real-World Environments. *Traitement Du Signal*, *38*(2), 349–358. https://doi.org/10.18280/ts.380212

Dendani, B., Bahi, H., & Sari, T. (2020). Speech enhancement based on deep autoencoder for remote arabic speech recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12119 LNCS*, 221–229. https://doi.org/10.1007/978-3-030-51935-3_24

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.

Eskenazi, M. (2009). An overview of spoken language technology for education. Speech Communication, 51(10), 832–844. https://doi.org/10.1016/j.specom.2009.04.005

ETSI AMR (2001). Universal Mobile Telecommunications System (UMTS); AMR wideband speech codec; Feasibility study report. https:// https://www.etsi.org/deliver/etsi_tr/126900_126999/126901/04.00.01_60/tr_126901v040001p.pdf

ETSI GSM-FR (1998). Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON); Using GSM speech codecs within. https://www.etsi.org/deliver/etsi_ts/101300_101399/101318/01.01.01_60/ts_101318v01

0101p.pdf

ETSI GSM-HR (2000). Digital cellular telecommunications system (Phase 2+); Half rate speech. http://www.etsi.org

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. http://arxiv.org/abs/2105.03075

Fitt, S. (1995). The pronunciation of unfamiliar native and non-native town names, in Proc. of EuroSpeech, 2227-2230.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*(3), 401–418. https://doi.org/10.1177/0265532210364408

Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, *30*(2), 121–130. https://doi.org/10.1016/S0167-6393(99)00045-X

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (1997). Automatic pronunciation scoring for language instruction. *Ieee*, *2*, 1471–1474.

Franco, H., Neumeyer, L., Ramos, M., & Bratt, H. (1999). Automatic Detection Of Phone-Level Mispronunciation For Language Learning. *Learning, Proc. of Eurospeech 99*, 851–854. http://leoneu.github.io/pub/eurospeech99.pdf

Frihia, H., & Bahi, H. (2017). HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications. *International Journal of Speech Technology*, *20*(3), 563–573. https://doi.org/10.1007/s10772-017-9427-z

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, *116*(December 2019), 86–97. https://doi.org/10.1016/j.specom.2019.12.002

Fu, S. W., Hu, T. Y., Tsao, Y., & Lu, X. (2017). Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, *2017-Septe*, 1–6. https://doi.org/10.1109/MLSP.2017.8168119

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*, *08-12-Sept*, 3768–3772. https://doi.org/10.21437/Interspeech.2016-211

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-February*, 6–12. https://doi.org/10.1109/APSIPA.2017.8281993

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

Ge, F., Pan, F., Liu, C., Dong, B., Chan, S. duen, Zhu, X., & Yan, Y. (2009). An SVM-based Mandarin pronunciation quality assessment system. *Advances in Intelligent and Soft Computing*, *56*, 255–265. https://doi.org/10.1007/978-3-642-01216-7_27

Gibson, J., D. (2016). Speech Compression. *Information*, *7*(2), 32. https://doi.org/10.3390/info7020032

Goodfellow, I. (2016). *Deep Learning*.

Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., & Falavigna, D. (2019). Automatic Assessment of Spoken Language Proficiency of Non-native Children. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2019-May*(1), 7435–7439. https://doi.org/10.1109/ICASSP.2019.8683268

Gruhn, R., Cincarek, T., Nakamura, S. (2004). A multi-accent nonnative English database. In Proceedings of Acoustical Society of Japan, 195–196.

Hammami, N., Lawal, I. A., Bedda, M., & Farah, N. (2020). Recognition of Arabic speech sound error in children. *International Journal of Speech Technology*, *23*(3), 705–711. https://doi.org/10.1007/s10772-020-09746-3

Harrison, A. M., Lo, W. K., Qian, X. J., & Meng, H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *International Workshop on Speech and Language Technology in Education (SLaTE), 45-48.*

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

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hirsch, H. (2002). *THE INFLUENCE OF SPEECH CODING ON RECOGNITION PERFORMANCE IN TELECOMMUNICATION NETWORKS Test data*. *2002*(September).

Honig, F., Batliner, A., Weilhammer, K., Noth, E. (2009). Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners, in ISCA Workshop on Speech and Language Technology for Education (SLaTE).

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

Hu, W., Qian, Y., & Soong, F. K. (2013). A new DNN-based high quality pronunciation evaluation for computer-aided language learning (call). *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *August*, 1886–1890.

Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, *67*, 154–166. https://doi.org/10.1016/j.specom.2014.12.008

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing : a guide to theory, algorithm, and system development*. Prentice Hall PTR. https://dl.acm.org/citation.cfm?id=560905

ITU-T G.711 (1988). Pulse code modulation (PCM) of voice frequencies. Retrieved October 16, 2021, from https://www.itu.int/rec/T-REC-G.711/

ITU-T G.722 (2012). 7 kHz audio-coding within 64 kbit/s. Retrieved October 16, 2021, from https://www.itu.int/rec/T-REC-G.722

ITU-T G.726 (1990). *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*. Retrieved April 26, 2021, from https://www.itu.int/rec/T-REC-G.726-199012-I/en

ITU-T G.728 (1992). Coding of speech at 16 kbit/s using low-delay code excited linear prediction. Retrieved September 6, 2018, from https://www.itu.int/rec/T-REC-G.728/e

ITU-T G.729 (1996). Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). Retrieved September 6, 2018, from https://www.itu.int/rec/T-REC-G.729

ITU-T (2019). Software tools for speech and audio coding standardization, Telecommunication Standardization Sector (ITU-T), International Telecommunication Union. G.191 (01/2019).

Janet C. E. Watson. (2002). *The Phonology and Morphology of Arabic*. Oxford University Press.

Khan, A. F. A., Mourad, O., Mannan, A. M. K. B., Dahan, H. B. A. M., & Abushariah, M. A. (2013, February). Automatic Arabic pronunciation scoring for computer aided language learning. In 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA) (pp. 1-6). IEEE.

Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.-E., Douze, M., & Dupoux, E. (2020). *Data Augmenting Contrastive Learning of Speech Representations in the Time Domain*. 1–6. http://arxiv.org/abs/2007.00991

Khelifa, M. O., Elhadj, Y. M., Abdellah, Y., & Belkasmi, M. (2017). Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system. International Journal of Speech Technology, 20(4), 937-949.

Kim. Y.W., Kim, T., Ko, L., Choi, D.-L., Lee, Y. (2016). Non-native Speech Corpora for STiLL at SiTEC, O-COCOSDA, Bali, Indonesia.

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kleijn, W. B., Lim, F. S. C., Luebs, A., Skoglund, J., Stimberg, F., Wang, Q., & Walters, T. C. (2017). *Wavenet based low rate speech coding*. 3–7. http://arxiv.org/abs/1712.01120

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio Augmentation for Speech Recognition. *INTERSPEECH*. http://www.isip.piconepress.com/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems, 2012, Vol. 25, p. 1097-1105*. http://code.google.com/p/cuda-convnet/

Kumar, A., & Florencio, D. (2016). Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks. *Interspeech 2016*. https://doi.org/10.21437/Interspeech.2016-88

Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H., & Farinas, J. (2016). Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-Sept*(September), 2686–2690. https://doi.org/10.21437/Interspeech.2016-513

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113. https://doi.org/10.1109/72.554195

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

Lee, A. (2016). *Language-Independent Methods for Computer-Assisted Pronunciation Training*. Ph.D. thesis, Massachusetts Institute of Technology. https://doi.org/10.1109/UCC.2011.36

Lee, A., Chen, N. F., & Glass, J. (2016). Personalized Mispronunciation Detection and Diagnosis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 6145–6149.

Lee, A., & Glass, J. (2013). Pronunciation Assessment via a Comparison-based System. *SLaTE*, 122–126.

Lee, A., & Glass, J. (2015). Mispronunciation detection without nonnative training data. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 643–647.

Li, K., Qian, X., & Meng, H. (2017). Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *25*(1), 193–207. https://doi.org/10.1109/TASLP.2016.2621675

Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. H. (2016). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2016-May*, 6135–6139. https://doi.org/10.1109/ICASSP.2016.7472856

Lilly, B. T., & Paliwal, K. K. (1996). Effect of speech coders on speech recognition performance. *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference On*, *4*, 2344–2347 vol.4. https://doi.org/10.1109/ICSLP.1996.607278

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE, 67(12), 1586-1604.

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*, *42*(1995), 60–88. https://doi.org/10.1016/j.media.2017.07.005

Loizou, P. C. (2013a). Speech Enhancement: Theory and Practice, 2nd ed. Boca Raton, FL, USA: CRC Press.

Loizou, P. C. (2013b). Evaluating Performance of Speech Enhancement Algorithms: Listening Tests. In Speech Enhancement: Theory and Practice, 2nd ed (pp. 439-439). Boca Raton, FL, USA: CRC Press.

Lu, X., Matsuda, S., Hori, C., & Kashioka, H. (2012). Speech Restoration Based on Deep Learning Autoencoder with Layer-Wised Pretraining. *InterSpeech, Portland, OR, USA, 1504-1507*. http://www.isca-speech.org/archive

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*, 436–440.

Maas, A. L., Le, Q. V., O'Neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust ASR. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, *1*, 22–25. https://doi.org/10.21437/INTERSPEECH.2012-6

Maqsood, M., Adnan Habib, H., Nawaz, T., & Zeeshan Haider, K. (2016). A Complete Mispronunciation Detection System for Arabic Phonemes using SVM. *IJCSNS International Journal of Computer Science and Network Security*, *16*(3), 30.

McCree, A. V., & Barnwell, T. P. (1995). A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Transactions on Speech and audio Processing, 3(4), 242-250.

Menzel, W., Atwell, E., Bonaventura P. et al. (2000). The ISLE corpus of non-native spoken English. In: Gavrilidou, M, (ed.) In Proc. of LREC, Athens, Greece.

Metallinou, A., & Cheng, J. (2014). Using deep neural networks to improve proficiency assessment for children english language learners. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *September*, 1468–1472.

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., Makino, S. (2004). Development of English speech database read by Japanese to support CALL research. Internatioanl Congress on Acoustics (ICA), 577–560.

Moffitt, J. (2001). Ogg Vorbis—open, free audio—set your media free. Linux journal, 2001(81es), 9-es.

Moustroufas, N., & Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech and Language*, *21*(1), 219–230. https://doi.org/10.1016/j.csl.2006.04.001

Nazir, F., Majeed, M. N., Ghazanfar, M. A., & Maqsood, M. (2019). Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes. *IEEE Access*, *7*, 52589–52608. https://doi.org/10.1109/ACCESS.2019.2912648

Necibi, K., & Bahi, H. (2015). A statistical-based decision for arabic pronunciation assessment. *International Journal of Speech Technology*, *18*(1), 37–44. https://doi.org/10.1007/s10772-014-9248-2

Necibi, K., Frihia, H., & Bahi, H. (2015). On the use of decision trees for Arabic pronunciation assessment. *ACM International Conference Proceeding Series*, *23-25-November-2015*.

https://doi.org/10.1145/2816839.2816866

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, *30*(2), 83–93. https://doi.org/10.1016/S0167-6393(99)00046-1

Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1457–1460. https://doi.org/10.1109/ICSLP.1996.607890

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote, J. A., & Muller Levis, G. (2018). Directions for the future of technology in pronunciation research and teaching. Journal of Second Language Pronunciation, 4(2), 182–207. https://doi.org/10.1075/jslp.17001.obr

Oh, Y. R., Jeon, H., Song, H. J., Kang, B. O., Lee, Y., Park, J., & Lee, Y. (2017). *Deep-Learning Based Automatic Spontaneous Speech Assessment in a Data-Driven Approach for the 2017 SLaTE CALL Shared Challenge*.

Oh, Y. R., Park, K., Jeon, H. B., & Park, J. G. (2020). Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition. *ETRI Journal*, *0*(August 2019), 1–12. https://doi.org/10.4218/etrij.2019-0400

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019-Septe, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Park, S. R., and Lee, J. W. (2016). A fully convolutional neural network for speech enhancement. arXiv preprint arXiv:1609.07132.

Pascual, S., Bonafonte, A., & Serra, J. (2017). SEGAN: Speech enhancement generative adversarial network. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2017-Augus*(D), 3642–3646. https://doi.org/10.21437/Interspeech.2017-1428

Qian, X., Meng, H., & Soong, F. (2012). The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training. *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012. ISCA 2012*.

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

Raghavan, S., Meenakshi, N., Mittal, S. K., Yarra, C., Mandal, A., Kumar, K. R. P., & Ghosh, P. K. (2017). A comparative study on the effect of different codecs on speech recognition accuracy using various acoustic modeling techniques. 2017 23rd National Conference on Communications, NCC 2017. https://doi.org/10.1109/NCC.2017.8077042

Ramana, A., Parayitam, L., & Pala, M. (2012). Investigation of Automatic Speech Recognition Performance and Mean Opinion Scores for Different Standard Speech and Audio Codecs. *IETE Journal of Research*, *58*(2), 121. https://doi.org/10.4103/0377-2063.96179

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. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), 2, 749–752. https://doi.org/10.1109/ICASSP.2001.941023

Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, 1041–1044. https://doi.org/10.1145/2647868.2655045

Saleem, N., & Khattak, M. I. (2019). Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments. *International Journal of Interactive Multimedia and Artificial Intelligence*, *InPress*(InPress), 1. https://doi.org/10.9781/ijimai.2019.06.001

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

Schroeder, M., & Atal, B. S. (1985, April). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 10, pp. 937-940). IEEE.Semmlow, J. (2012). The Fourier Transform and Power Spectrum. In *Signals and Systems for Bioengineers* (pp. 131–165). Elsevier. https://doi.org/10.1016/b978-0-12-384982-3.00004-3

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Silovsky, J., Cerva, P., & Zdansky, J. (2011). Assessment of speaker recognition on lossy codecs used for transmission of speech. *ELMAR, 2011 Proceedings*, *September*, 14–16. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6044294

Sinder, D. J., Varga, I., Krishnan, V., Rajendran, V., & Villette, S. (2015). Recent Speech Coding Technologies and Standards. In *Speech and Audio Processing for Coding, Enhancement and Recognition* (pp. 75–109). Springer New York. https://doi.org/10.1007/978-1-4939-1456-2_4

Spanias, A. S. (1994). Speech Coding: A Tutorial Review. *Proceedings of the IEEE*, *82*(10), 1541–1582. https://doi.org/10.1109/5.326413

Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, *51*(10), 845–852. https://doi.org/10.1016/j.specom.2009.05.007

Sun, L., Mkwawa, I. H., Jammeh, E., & Ifeachor, E. (2013). Speech compression. In Guide to Voice and Video over IP (pp. 17-51). Springer, London.

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

Tabbaa, H. M. A., & Soudan, B. (2015). Computer-Aided Training for Quranic Recitation. *Procedia - Social and Behavioral Sciences*, *192*, 778–787.

https://doi.org/10.1016/j.sbspro.2015.06.092

Tan, K., & Wang, D. L. (2018). A convolutional recurrent neural network for real-time speech enhancement. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2018-Septe*(September), 3229–3233. https://doi.org/10.21437/Interspeech.2018-1405

Tan, Z.-H., & Varga, I. (2008). Automatic Speech Recognition on Mobile Devices and over Communication Networks. In *Network, Distributed and Embedded Speech Recognition: An Overview* (pp. 1–23). Springer, London. https://doi.org/10.1007/978-1-84800-143-5_1

Tao, J., Ghaffarzadegan, S., Chen, L., & Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2016-May*, 6140–6144. https://doi.org/10.1109/ICASSP.2016.7472857

Valin, J. M. (2016). Speex: A free codec for free speech. arXiv preprint arXiv:1602.08668.

Valin, J. M., Vos, K., & Terriberry, T. (2012). Definition of the Opus audio codec. IETF, September.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 1096–1103. https://doi.org/10.1145/1390156.1390294

Walker, W., Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., & Woelfel, J. (2004). *Sphinx-4: A flexible open source framework for speech recognition*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.4704

Wang, D., & Chen, J. (2018). Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1702-1726.

Wang, Y. B., & Lee, L. S. (2015). Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. *IEEE Transactions on Audio, Speech and Language Processing*, *23*(3), 564–579. https://doi.org/10.1109/TASLP.2014.2387413

Wang, Y., Narayanan, A., & Wang, D. L. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *22*(12), 1849–1858. https://doi.org/10.1109/TASLP.2014.2352935

Wani, M. A., Bhat, F. A., Afzal, S., & Khan, A. I. (2020). *Unsupervised Deep Learning Architectures* (pp. 77–94). Springer, Singapore. https://doi.org/10.1007/978-981-13-6794-6_5

Watson, J. C. (2002). The phonology and morphology of Arabic. Oxford University Press on Demand.

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. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9237*, 91–99. https://doi.org/10.1007/978-3-319-22482-4_11

Williamson, D. S., Wang, Y., & Wang, D. L. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *24*(3), 483–492. https://doi.org/10.1109/TASLP.2015.2512042

Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning*. Ph.D. thesis, University of Cambridge. https://www.repository.cam.ac.uk/handle/1810/251707

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2), 95–108. https://doi.org/10.1016/S0167-6393(99)00044-8

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

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

Young, S. (1997). Computer-assisted Pronunciation Teaching based on Automatic Speech Recognition. *Challenges, December*, 1–12.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Zhao, G., Sonsaat, S., Silpachai,A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., Gutierrez-Osuna, R., L2-ARCTIC: A Non-Native English Speech Corpus, in Proc. InterSpeech (2018) Hyderabad, India.

Zhao, H., Zarar, S., Tashev, I., & Lee, C. H. (2018). Convolutional-Recurrent Neural Networks for Speech Enhancement. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 2401–2405. https://doi.org/10.1109/ICASSP.2018.8462155

Zhao, Z., Liu, H., & Fingscheidt, T. (2018). 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