REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE YAHIA FARES DE MEDEA



Faculté de Technologie
Département du Génie Electrique

Mémoire de fin d'études de Master

Filière : Télécommunications

Spécialité : Systèmes des Télécommunications

Effect of speech bandwidth on speaker recognition performance using deep neural network (DNN)

Présentées par :

Mr. HADJERSI Youcef

Mr. HAMMOU Oussama

Proposé et Dirigé par:

Dr. KROBBA Ahmed

Promotion 2020

Acknowledgments

Throughout the writing of this dissertation we have received a great deal of support and assistance.

We would first like to thank our supervisor, dr.Krobba whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback helped us to sharpen our thinking and brought our work to a higher level.

In addition, the support, encouragements and comfort of our dear parents are forever appriciated beside their wise counsel and sympathetic ear. You are always there for us. And we would like to acknowledge our colleagues from our internship at the university for everything they have done to help us.

Finally, we could not have completed this dissertation without the support of our friends, who provided stimulating discussions as well as happy distractions to rest our mind outside of our research.

We dedicate this modest work to our dear parents, for their moral and financial support and for being the light of our life, who have always encouraged us to finish our studies, hoping them a long life and may God keep them.

• To our promoter Mr 'KROBBA .A'.

• To our brothers and our sisters.

• To all our dear friends, especially HICHEM.

• To all teachers in the Faculty of Science and Technology.

Oussama and Youcef

TABLE OF CONTENTS

LIST OF FIGURES i

LIST OF TABLES ii

LIST OF ABBREVIATIONS ii

CHAPTER I: 1

An overview of automatic speaker recognition technology 1

I.1 Introduction 4

I.2 Biometrics 4

I.3 Voice biometrics 4

I.4 Speech processing and speaker recognition 4

I.5 Different tasks of speaker recognition 5

I.5.1 Speaker identification system 5

I.5.2 Speaker verification System 6

I.6 General Architecture of speaker recognition system 7

I.6.1 Features extraction 8

I.6.1.1 Short-term feature extraction 8

I.6.1.2 Long-term feature extraction 10

I.6.2 Speaker modeling 10

I.6.3 Decision and performance evaluation 14

I.7 Applications of speaker recognition system 15

I.8 Conclusion: 16

CHAPTER II 17

Speaker recognition and mobile communication 17

II.1 Introduction 18

II.2 Voice recognition system over mobile communication 18
II.3 Speech coding 19
II.4 Factors affecting the performance of ASR in a mobile environment 20
II.4.1 Acoustic environment 21
II.4.2 Transmission channel noise 21
B)- Rayleigh channel 23
II.4.3 Effect of bandwidth 23
II.5. Different speaker recognition over mobile network 24
II.6 Conclusion 29
CHAPTER III 30
Experiments and results 30
III.1 Introduction 31
III.2 MSR Identity Toolbox with MATLAB 31
III.3 Experimental Setup 31
III.3.1 Datasets Description 31
III.3.2 Block diagram of ASV system without DNN 32
III.3.3 Block diagram of ASV system based on DNN 33
III.4 Experiments and results 34
III.4.1 ASV performance in Narrowband speech (NB) 34
III.4.2 ASV performance in wideband speech (WB) 35
III.4.3 ASV performance in BWE 36
III.5.Conclusion 38
Conclusion 39
Bibliography 40

LIST OF FIGURES

Figure I-1: Speech processing 5

Figure I-2: Speaker identification system. 6

Figure I-3:speaker verification 7

Figure I-4:General architecture of a speaker recognition system..... 8

Figure I-5: Various steps in the MFCC feature extraction procedure from a speech frame. ... 10

Figure I-6:GMM-UBM-based speaker modeling. 11

Figure I-7:Principle of support vector machines (SVM). 12

Figure I-8: Block diagram of a fully connected DNN architecture with 5dimensional input layer, 'H' 8-dimensional hidden layers and a 4-dimensional output layer. 14

Figure I-9:An illustration of target and non-target score distributions and the decision threshold. Areas under the curves with blue and red colors represent FAR and FRR errors, respectively..... 15

Figure II-1:General scheme of a speech-driven remote information system 19

Figure II-2:The effect of different levels of background music on speech-recognition accuracy..... 21

Figure II-3:Additive White Gaussian Noise..... 22

Figure II-4:The signal from Figure II.4 when it has been bandlimited to the frequencies present in a typical telephone signal(top),and its spectrogram (bottom).Information-bearing spectrographic patterns in the higher frequencies have been erased..... 24

Figure II-5:EMBEDDED MOBILE SPEECH RECOGNITION..... 25

Figure II-6:Network speech recognition 26

Figure II-7:distributed speech recognition 28

Figure III-1:Block diagram of ASV system without DNN 32

Figure III-2:block diagram of DNN-BWE..... 34

Figure III-3:Comparison of Speaker Verification System Performance using MFCCs and GFCCs features under different AMR Multi-Rate Narrowband speech codecs. 35

Figure III-4:Comparison of Speaker Verification System Performance using MFCCs and GFCCs features under different AMR Multi-Rate wideband speech codecs. 36

LIST OF TABLES

Tableau III-1: Performance comparison between BWE, WB and NB baseline based on i-vector-GPLDA system using different features extraction (MFCC and GFCC) 37

Tableau III-2: Performance comparison between BWE, WB and NB baseline based on DNN i-vector-GPLDA system using different features extraction (MFCC and GFCC)..... 37

LIST OF ABBREVIATIONS

ACELP : Algebraic Code Excited Linear Prediction.

AMR-NB : Adaptive Multi-Rate Narrowband

AMR-WB :Adaptive Multi-Rate Wideband.

ASI : Automatic Speaker Identification.

ASR : Automatic Speaker Recognition.

ASV : Automatic Speaker Verification.

AWGN : Additive White Gaussian Noise.

CMVN : Cepstral Mean and Variance Normalization

DNN : Deep Neural Networks.

EER : Equal Error Rate.

FA : False Acceptance.

FAR : False Acceptance Rate.

FR : False Rejection.

FRR : False Rejection Rate.

GFCC : Gammtone Frequency Cepstral Coefficients

GMM : Gaussian Mixture Model.

GMM_UBM : Gaussian Mixture Model_ Universal Background Model.

G-PLDA: Gaussian Probabilistic linear discriminant analisys

GSM : Global System Mobile.

ICSI: International Computer Science Institute.

JFA : Joint Factor Analysis.

LDC: Linguistic Data Consortium.

LPC : Linear Predictive Coding.

MFCC : Mel-Frequency Cepstral Coefficients.

NB: Narrowband

NIST: National Institute of Standards and Technology.

NSR : Network speech recognition

PLDA : Probabilistic linear discriminant analisys.

PSD : Power Spectral Density.

SRE: Speaker Recognition Evaluation.

SVM : Support Vector Machine.

VAD : Voice Activity Detector.

VoIP : Voice Internet Protocol

VSELP :Vector Sum Excited Linear Prediction.

1. Introduction

Speech is the primary means of communication between human beings. In fact, one can observe how difficult it is when communicating with difficulties or disabilities. Generally, people prefer to communicate by spoken words rather than the written ones. Technological advancement and especially computer science has created the need for new means of man-machine dialogue, speech is highly used in biometric systems due to its ease of transmission over long distances rather than the other means used in biometrics.

Speaker recognition is the process of automatically recognizing who is speaking by using speaker- specific information included in speech waves it deals with identifying a person from his/her voice. This type of biometric recognition is important in a number of applications, including: security systems, network access-control, forensics, automated customer services, etc. Individuals generally have distinct voices due to their unique vocal tract shape, larynx size, and other physiological properties of the human speech production system. Also, each speaker has their own way of speaking, including their accent, rhythm, intonation style, pronunciation patterns and choice of language and vocabulary [**1**].

The most advanced speaker recognition systems today utilize many of these speaker characteristics in different ways, with the short-term voiced excitation characteristics being exploited in most scenarios.

2. Motivation in this work

Recent years have witnessed a rapid deployment of high-speed speech transmission networks, primarily motivated by cost savings, ease of development, and integration of new services. The speech data are now sent over IP networks, such as VoIP, transporting telephone calls over the Internet. The NB transmitting the audio frequency range 300–3,400 Hz, which is only a small portion of the frequencies of the human speech. More recently, with the advent of IP networks, the dramatic expansion of available bandwidth favoured a drive towards wideband audio, extending the transmitted frequency range to 50–7,000 Hz [**2**]. Most recent work focused on exploiting the capability of deep neural net (DNN) given its success in several tasks in speech processing and analysis, and DNN based speech BWE has demonstrated improvement in automatic speech recognition (ASR) [**3**]. A feed-forward DNN trained feature of NB and WB data was incorporated in an ASR system.

In this work, we examines whether WB and WBE transmissions offer an improvement over NB transmissions on the speaker recognition performance

3. Organization of the thesis

This thesis is organized as follows:

In the first chapter, we introduce the basic concepts of speaker recognition. We describe the different tasks of speaker recognition (Identification and verification systems) .as well as the main components of speaker recognition such as features extraction, speaker modeling, decision, and the performance measurement of the system.

• The second chapter is totally dedicated to describe the applications of speech and speaker recognition over mobile environment.

• The third chapter, we presents the series of experiments and results.

Finally, the conclusions and future work

CHAPTER I:

An overview of automatic speaker recognition technology

I.1 Introduction

The present chapter gives a brief overview of the speaker recognition, this includes a description of the different tasks of speaker recognition (Identification and verification systems). We also present the general main components of speaker recognition such as features extraction, speaker modeling, and decision. The rest of this chapter presents the performance measurement of a speaker recognition system with different applications.

I.2 Biometrics

Biometrics is the science of establishing the identity of an individual based on the physical, chemical or behavioral attributes of the person. Biometrics is used more and more in applications of the everyday life. So with its beginnings at the end of the 19th century the biometric data were treated manually, today, with the data processing, the biometric systems are automated [4], it is based on 'Who you are' refers to his physiological characteristics such as fingerprints, iris, or face. And 'What you produce' refers to his behavioral patterns that characterize your identity such as the voice or the signature [5]. These physiological or behavioral characteristics are called biometric modalities. Biometrics such as we wants to use it today in the security systems aims to make an automatic recognition [4].

I.3 Voice biometrics

Voice biometric technology for authentication user is more convenient and accurate. This is because the biometrics characteristic if an individual are unique and belongs to the personal until the user dead. It is convenient for the user because nothing to be carried or remembered and would not scare the ID card being stolen or password being hacked. This is the only technique that currently allows you to recognize a person from a distance.

I.4 Speech processing and speaker recognition

Speech processing is a science located at the intersection of digital signal processing and language processing. Speech has the particularity, compared to other information processing signals, of being produced and perceived instantly by the brain, and for this reason speech processing tends to replace these functions with automatic systems.

Speech processing (Figure I.1) includes the speech coding, analysis, synthesis, and recognition. In recognition, we can find: speech recognition and speaker recognition. Speaker

recognition, unlike speech recognition, is particularly interested in extra-linguistic information conveyed by a speech signal.

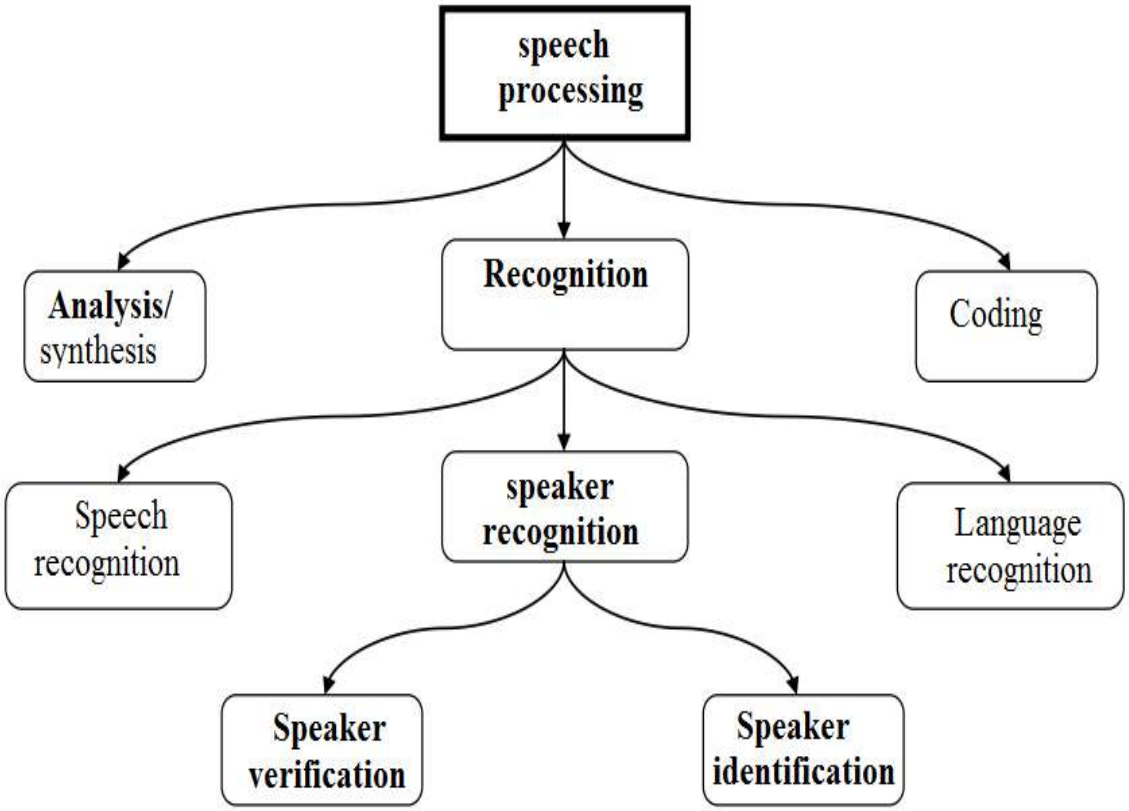


Figure I-1: Speech processing

I.5 Different tasks of speaker recognition

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual's information included in speech waves. Speaker recognition can be classified into identification and verification. Speaker recognition has been applied most often as means of biometric authentication.

I.5.1 Speaker identification system

The principle of automatic speaker identification, illustrated in (Figure I.2), consists of finding the identity of the associated speaker among a population of known speakers. From a

schematic point of view, a speech sequence is given at the input of the automatic speaker identification system.

For each known speaker in the system, the speech sequence is "compared" to a reference characteristic of the speaker. The identity of the speaker whose reference is closest to the speech sequence is given at the output of the ASI system [7]. Two modes are offered in ASI: identification in a closed set, for which it is assumed that the speech sequence is actually pronounced by a known speaker and identification in open set for which the speaker may not be known. In "open set" mode, the system must decide the reliability of its judgment by accepting or rejecting the identity it has found [7]. By virtue of its principle - determining an identity among potential identities - the performance of ASI systems deteriorates as the population of speakers increases [8,9].

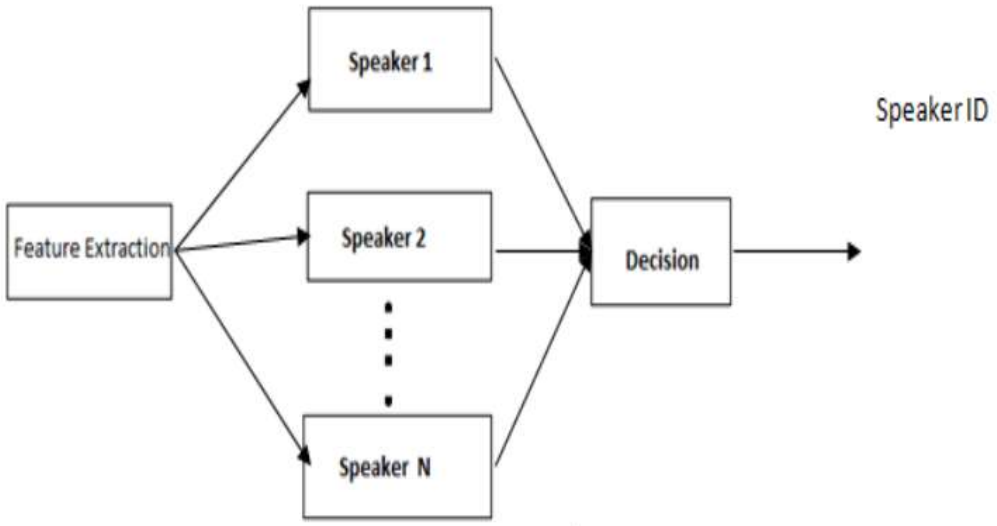


Figure I-2: Speaker identification system.

I.5.2 Speaker verification System

The Automatic Speaker Verification (ASV) is the decision-making process allowing to determine using a voice message, the veracity of the identity claimed by an individual (figure I.3). Identity and voice message are the two inputs to the ASV system. The identity, necessarily known to the system, automatically designates the reference characteristic of a speaker. A similarity measure based on the likelihood ratio is calculated between this

reference and the voice message and then compared to a decision threshold. If the similarity measure is above the threshold, the individual is accepted. Otherwise, the individual is considered an impostor and will be rejected.

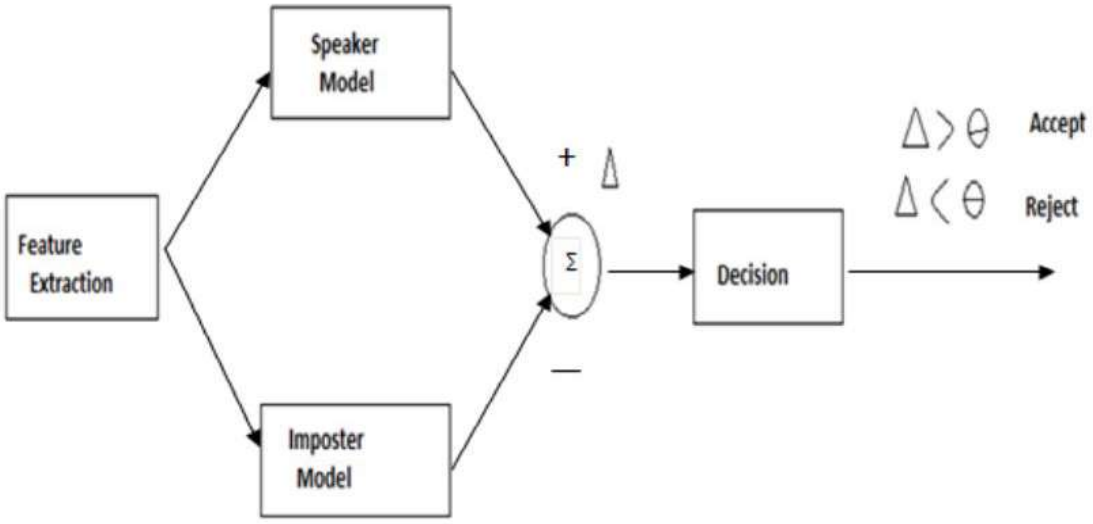


Figure I-3: Speaker verification.

I.6 General Architecture of speaker recognition system

A state-of-the art speaker recognition system consists of three main stages: the first is feature extraction where the speech signal is represented in a compact manner, the second is speaker modeling which can be defined as a process of describing feature properties for a given speaker, and the last stage is scoring or decision.The (Figure. I.4) shows the general structure of the speaker recognition system.

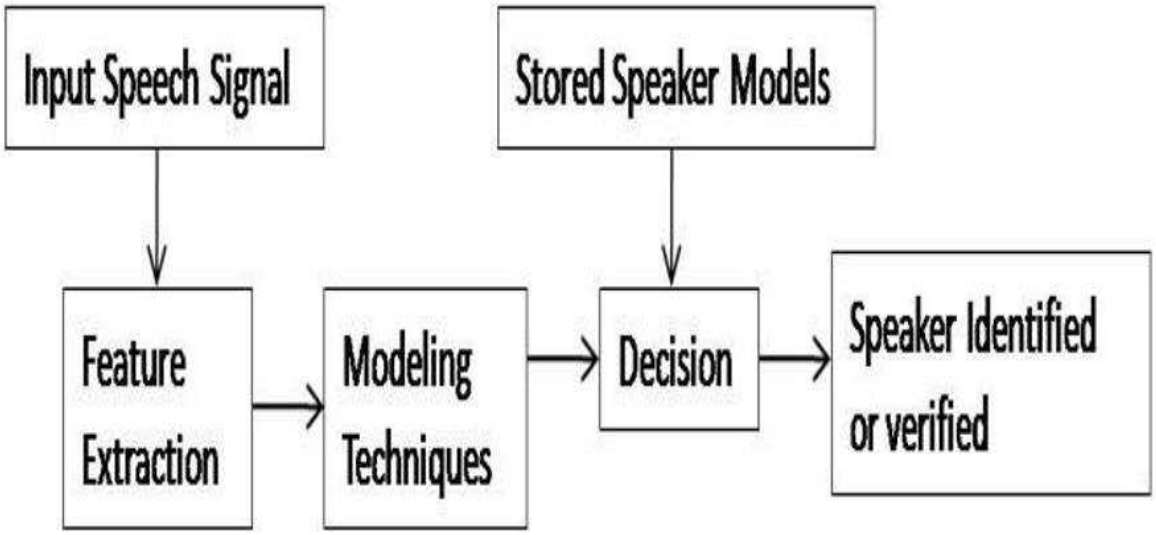


Figure I-4: General architecture of a speaker recognition system.

I.6.1 Features extraction

There are different methods of representing the speech signal. Some have been specifically developed by the acoustic study of speech production. Other parameters simulated the characteristics of the human ear. The features extraction can be classified into two main categories: short-term vocal tract characteristics of the speaker and prosodic features that model the speaker's long-term vocal tract characteristics.

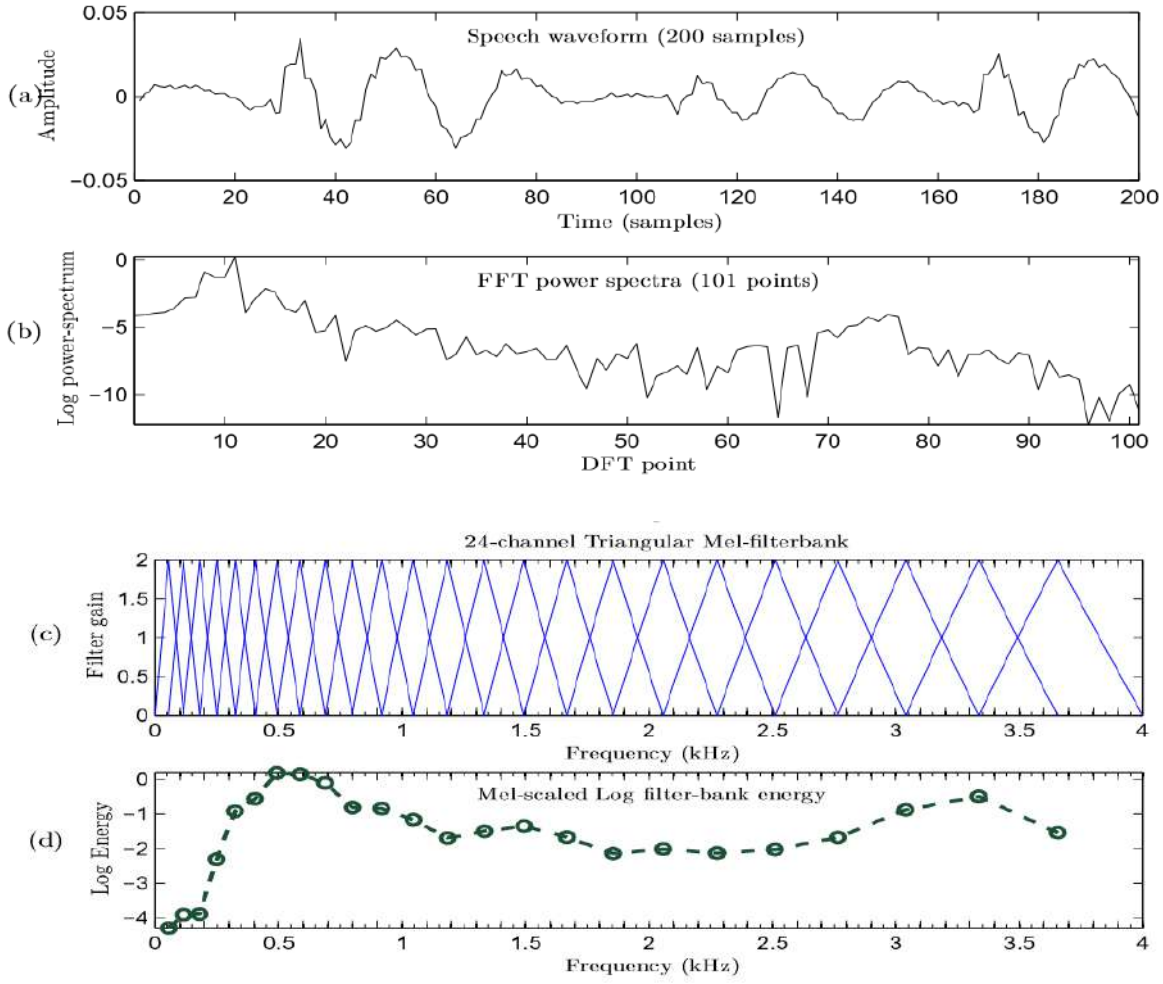
I.6.1.1 Short-term feature extraction

Short-time spectral representations have been among the most popular speech feature types in experiments on automatic speaker recognition during the past two decades. These are usually cepstral forms of the spectrum based on a linear prediction model, a filter bank, or more sophisticated perceptually motivated models of the speech signal.

a) Mel-frequency cepstral coefficients

The most popular short-time acoustic features are the Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) based features. To obtain the MFCC features from an audio recording, first the audio samples are divided into short overlapping segments. These segments are typically 20 − 25ms in duration. A typical 25 ms speech signal frame is shown in Figure I.5.

(a). The signal obtained in these segments/frames are then multiplied by a tapered window function (e.g., hamming, hanning, etc.) and the Fourier power-spectrum is obtained. Next, the logarithm of the spectrum is computed and the no-linearly spaced mel-space filter-bank analysis is performed. The latter process , the filter-bank analysis produces the spectrum energy in each channel (also known as the filter bank energy coefficients), representing different frequency bands with 24-channel filter bank. Finally, the MFCC coefficients are obtained by performing a Discrete Cosine Transform (DCT) on the filter-bank energy parameters and retaining a number of leading coefficients.



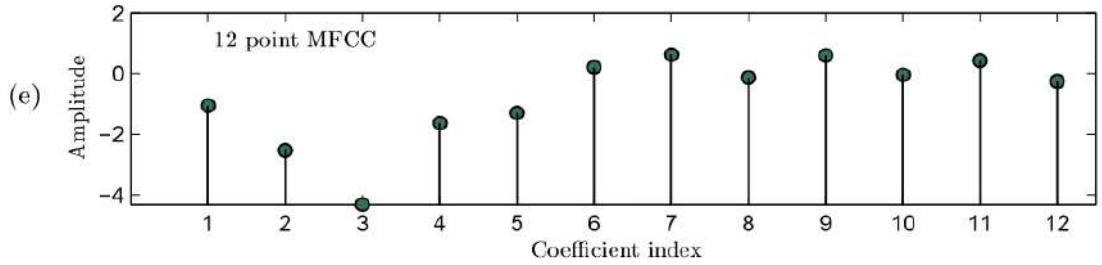


Figure I-5: Various steps in the MFCC feature extraction procedure from a speech frame.

I.6.1.2 Long-term feature extraction

Long-term feature extraction refers to non-segmental aspects of speech, including for instance syllable stress, intonation patterns, speaking rate and rhythm. One important aspect of prosody is that, unlike the traditional short-term spectral features (MFCC, GFCC, PLP), it spans over long segments like syllables, words, and utterances and reflects differences in speaking style, language background, sentence type.

I.6.2 Speaker modeling

Once audio segments are converted to feature parameters, the next task of the speaker recognition process is acoustic modeling. In general terms, we can define modeling as a process of characterizing the feature properties for a given speaker. The model must also provide a means of comparison with an unknown utterance. A modeling method is robust when the characterizing process of the features are not significantly affected by unwanted distortions, even though the features themselves incorporate the degradations. Any distortion in the acoustic signal directly affects the acoustic feature parameters, which in turn affects its match score with the speaker model, especially if the acoustic condition is unseen. Ideally, if features could be designed in such a way that no intra-speaker variation is present while inter-speaker discrimination is maximum, the simplest methods of modeling might suffice. In essence, the non-ideal properties of the feature extraction stage forces us to consider applying various techniques during the modeling phase so that the effect of the nuisance variations observed in the signal are minimized during the recognition process. Most speaker modeling techniques make various mathematical assumptions on the features (Gaussian distributed, for example). If these properties are not met by the data, we are essentially introducing imperfections during the modeling phase as well.

In this part, we will review the major modeling techniques that have been proposed in the past few decades and analyze how these methods help improve the robustness of speaker recognition systems.

a) GMM

A. Gaussian Mixture Model (GMM) is a generative model widely used in speaker recognition. It represents the state-of-the-art in this field. This model was introduced and apphed for the first time in speaker verification system [10].

b) GMM-UBM

The GMM-UBM (Universal Background Model) is a large GMM trained to represent a speaker-independent distribution of features. The corresponding training utterances are selected according to the types and quality of speech, as well as the composition of speech expected to be encountered during recognition(figure I-6).

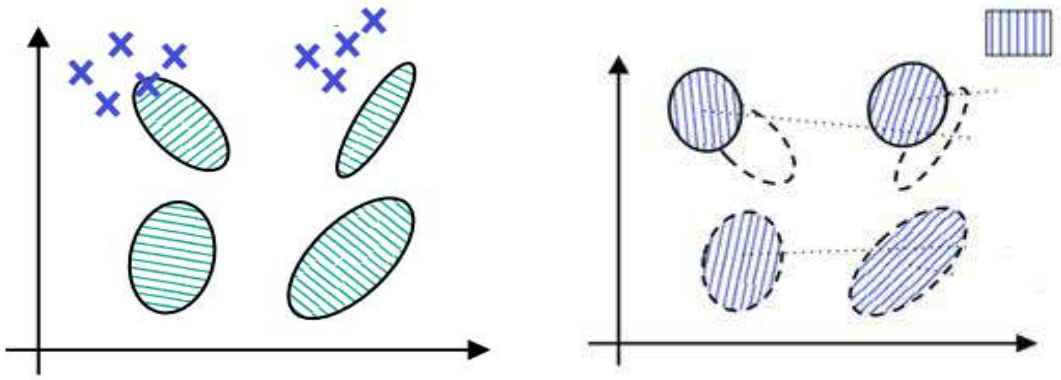


Figure I-7:GMM-UBM-based speaker modeling.

c) Support Vector Machines

Support vector machines (SVMs,called also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training

algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum (Figure I-8).

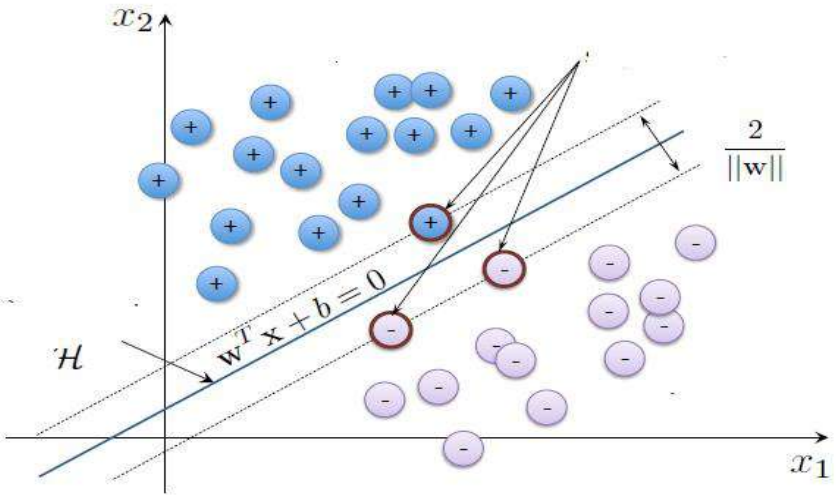


Figure I-9: Principle of support vector machines (SVM).

d) Joint factor analysis (JFA)

In a generative approach based on Gaussian mixture models, each speaker is represented by a GMM composed of C Gaussians. These Gaussians are learned in a continuous parameter space of dimension F. Each Gaussian is clo,gharacterized by a mean vector, a diagonal covariance matrix and a weight. A target speaker GMM is built by adapting the GMM components of the Universal Background Model (UBM) to the considered speaker's frames. The UBM is trained on a large set of speaker training data. Joint factor analysis [11] is a model that takes into account speaker and intersession variabilities in the context of the GMM framework. Traditionally used in conjunction with cepstral features, its application can be extended to other continuous features where GMM modehng is appropriate. The Joint factor analysis **(**JFA) model is based on a combination of classical MAP adaptation and eigenvoice for modeling speaker variability, and eigenchannel MAP for modeling

intersession variability. The intersession variability in the spectral speech features is generally caused by channel transmission effects. This is the reason for using the term channel variability rather than intersession variability in the context of spectral features. The key assumption in joint factor analysis is that the GMM supervector of speaker- and channel dependent M for a given utterance can be broken down into a sum of two supervectors:

$$M = s + c \tag{I.1}$$

where supervector s depends on the speaker and supervector c depends on the channel. The GMM supervector is a CF-dimensional vector obtained via concatenation of all Gaussian component means.

e) PLDA

PLDA, a generative probability model with which we can both extract the features and combine them for recognition. The latent variables of PLDA represent both the class of the object and the view of the object within a class. By making examples of the same class share the class variable. The usual LDA features are derived as a result of training PLDA, but in addition have a probability model attached to them, which automatically gives more weight to the more discriminative features. With PLDA, we can build a model of a previously unseen class from a single example, and can combine multiple examples for a better representation of the class.

f) DNN

Deep neural networks (DNNs) are powerful machine learning algorithms which are being extensively used in several areas including speech and image processing. DNNs are fully connected feed forward neural networks with more than one hidden layers. These hidden layers are used to transform the input feature vector into a probability distribution to estimate the output classes [10]. Fig.I-8 shows a basic DNN architecture with an input layer (x), 'H' hidden layers and an output layer.

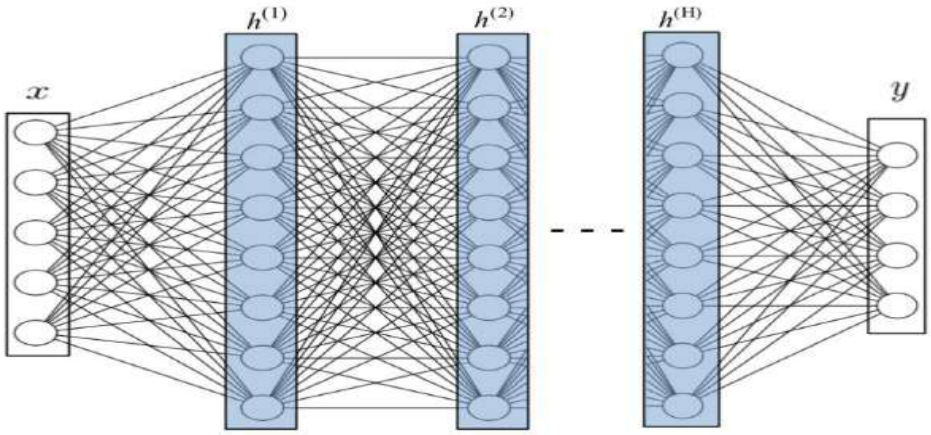


Figure I-10: Block diagram of a fully connected DNN architecture with 5dimensional input layer, 'H' 8-dimensional hidden layers and a 4-dimensional output layer.

I.6.3 Decision and performance evaluation

To measure the performance of a speaker verification system, we analyze two types of errors:

- False Acceptance (FA): this occurs when the system grants access to an impostor.
- False Rejection (FR): this occurs when the system denies access to an enrolled speaker.

From the speaker-detection point of view (a target speaker is sought), these are termed as: False alarm and Miss error, respectively. According by two error rates are defined as:

- False Acceptance Rate (FAR) $= \frac{\text{Total number of FA errors}}{\text{Total number of impostor speaker attempts}}$ (I.2)
- False Reject Rate (FRR) $= \frac{\text{Total number of FR errors}}{\text{Total number of enrolled speaker attempts}}$ (I.3)

A)- Equal error rate

The Equal Error Rate (EER) is defined as where the FAR and FRR values when become equal. That is, by changing the threshold, we find a point where the FAR and FRR become equal. This is shown in Figure I-9, where the shaded FAR and FRR areas are the same. EER is a very popular performance measure for speaker verification systems, since it allows a single number for system comparison. The soft scores from the automatic system are required to compute the EER. No actual hard decisions are made. It should be noted that, operating a speaker verification system on the threshold corresponding to the EER may not be desirable

for practical purposes. For high security applications, one should set the threshold higher, lowering the FA errors at the cost of miss errors. However, for high convenience, the threshold should be set lower. In the latter case, accepting an impostor speaker is not as critical as in high security applications.

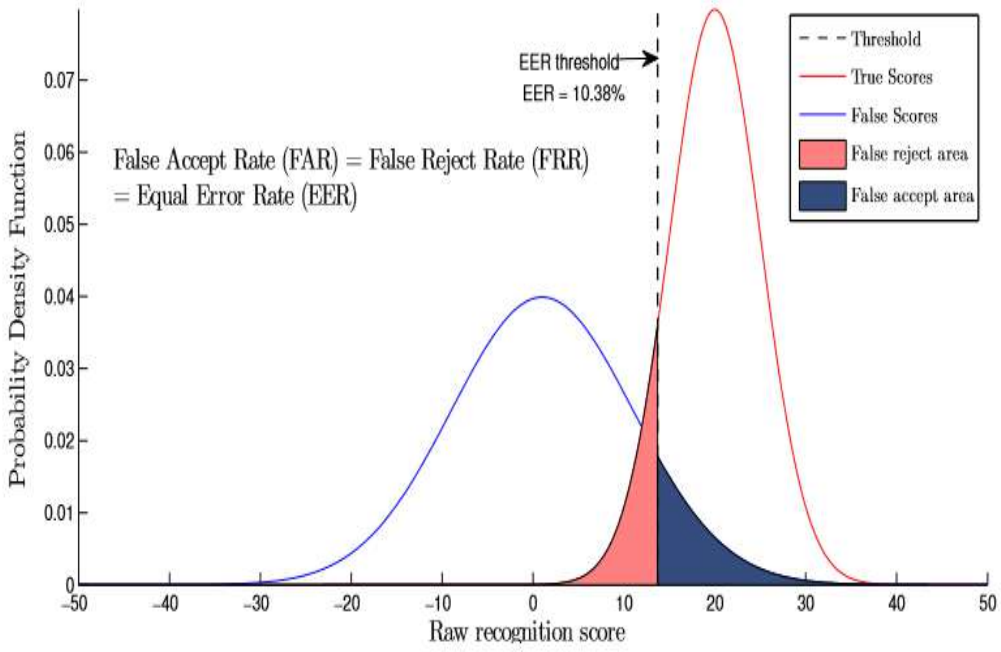


Figure I-11: An illustration of target and non-target score distributions and the decision threshold. Areas under the curves with blue and red colors represent FAR and FRR errors, respectively.

I.7 Applications of speaker recognition system

Speaker recognition has its vast applications in various fields. Following is the list of few technologies where speaker recognition has been or is currently used.

Access Control: Here speaker recognition can be used as a very effective biometric authentication. This includes controlling the access for computer network or websites required some subscription. This May be used for automated password reset services.

Forensic Science: Includes legal use of acoustic samples. Here by the use of forensic speaker comparison, one may conclude whether the two speech samples are from the same speaker or not. This introduces the concept of ear witness in many of the legal cases.

Transaction Authentication: For telephone banking, in addition to password account access control, higher levels of verification can be used for more sensitive transactions. More recent applications are in user verification for remote electronic and mobile purchases (e-and m-commerce).

Law Enforcement: Some applications are home-parole monitoring (call parolees at random times to verify they are at home) and prison call monitoring (validate inmate prior to outbound call).This Includes surveillance and speaker tracking.

Speech Data Management: In voice mail browsing or intelligent answering machines, use speaker recognition to label incoming voice mail with speaker name for browsing and/or action (personal reply).

Personalization: In voice-web or device customization, store and retrieve personal setting based on user verification for multi-user site or device. There is also interest in using recognition techniques for directed advertisement or services, where, for example, repeat users could be recognized or advertisements focused based on recognition of broad speaker characteristics.

I.8 Conclusion:

This chapter has been set out in the spirit of being a brief introduction of automatic speaker recognition technology, thus the tasks of ASR (identification and verification), we also described the main stages of speaker recognition system: features extraction,speaker modeling and decision.The next chapter will be entirely devoted to speaker recognition system on mobile environment.

CHAPTER II

Speaker recognition and mobile communication

II.1 Introduction

In this chapter, we describe the applications of speech and speaker recognition on mobile environment. The first part describes the general structure of the voice recognition system over mobile networks. The second part gives an overview of mobile network and speech coding standards commonly used in mobile communication. The third part presents the factors for degrading the performance of the recognition system. The fourth part deals with the different architectures allowing the integration of the speaker recognition system in mobile networks such as embedded speaker recognition, network speaker recognition, and distributed speaker recognition.

II.2 Voice recognition system over mobile communication

Speech is the most natural means of communication between humans. With the development of information technology and the massive use of the computer, the Human Dialogue-Interest in machine (DHM) using speech as a means of communication from both the scientific and industrial communities. On the other hand, mobile internet devices, ultra mobile PCs, internet tablets, smartphones and cellphones have widely proliferated the market. Accordingly there's a growing interest in developing applications that employ automatic speech and speaker recognition on mobile devices.

However, direct reproduction of algorithm that are suitable for desktop applications is either not possible or often leads to low performance on mobile devices [11]. Since these mobile devices are often used when the person is "on the move", variable acoustic environments and limited resources on the mobile device necessitates special arrangements [14].

Figure II-1 shows a general diagram of a remote information system using the speech and speaker recognition system. After recognition, the server provides and transmits the required information to the client. The information returned could be text, data or even a vocal response either generated by a text-to-speech system or previously recorded.

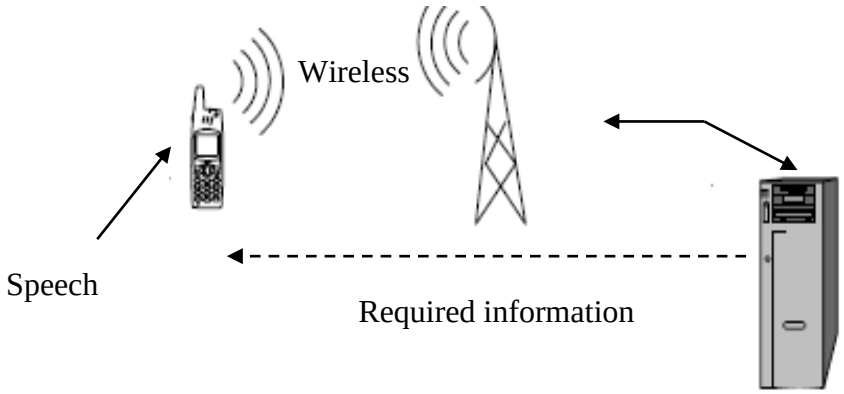


Figure II-1: General scheme of a speech-driven remote information system.

II.3 Speech coding

There exist different GSM speech coders; among them, we find the full rate, half rate and enhanced full rate coders. Their corresponding European telecommunications standards are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively. These coders work on a 13-bit uniform PCM speech input signal, sampled at 8 kHz. The input is processed on a frame-by-frame basis, with a frame size of 20 ms (160 samples). A brief description of these coders follows.

II.3.1 Full rate speech coder

The Full rate coder was standardized in 1987. This coder belongs to the class of Regular Pulse Excitation-Long Term Prediction—linear predictive (PCM) coders. In the encoder part, a frame of 160 speech samples is encoded as a block of 260 bits, leading to a bit rate of 13 kbps. The decoder maps the encoded blocks of 260 bits to output blocks of 160 reconstructed speech samples. The GSM full rate channel supports 22.8 kbps. Thus, the remaining 9.8 kbps are used for error protection. The FR coder is described in GSM 06.10 down to the bit level, enabling its verification by means of a set of digital test sequences which are also given in GSM. 06.10.

II.3.2 Half rate speech coder

The HR coder standard was established to cope with the increasing number of subscribers. This coder is a 5.6 kbps VSELP (Vector Sum Excited Linear Prediction) coder from Motorola (Gerson and Jasiuk 1993). In order to double the capacity of the GSM cellular system, the half rate channel supports 11.4 kbps. Therefore, 5.8 kbps are used for error protection. The

measured output speech quality for the HR coder is comparable to the quality of the FR coder in all tested conditions, except for tandem and background noise conditions. The normative GSM 06.06 gives the bit exact ANSI-C code for this algorithm, while GSM 06.07 gives a set of digital test sequences for compliance verification.

II.3.3 Enhanced full rate speech coder

The EFR coder was standardized later. This coder is intended for utilization in the full rate channel, and it provides a substantial improvement in quality compared to the FR coder (Järvinen 1997). The EFR coder uses 12.2 kbps for speech coding and 10.6 kbps for error protection. The speech coding scheme is based on Algebraic Code Excited Linear Prediction (ACELP).

II.3.4 Adaptive multi-rate wideband coder

The G.722.2 AMR-WB algorithm is used as an internet wideband audio codec for VoIP applications with an audio band of 50 – 7000 Hz instead of the 200 – 3400 Hz used with traditional telephony. The increased bandwidth improves the intelligibility and naturalness of speech significantly. G.722.2 describes the detailed mapping from input blocks of 320 speech samples in 16 bit uniform PCM format to encoded blocks of 132, 177, 253, 285, 317, 365, 397, 461 and 477 bits and from encoded blocks of 132, 177, 253, 285, 317, 365, 397, 461 and 477 bits to output blocks of 320 reconstructed speech samples. The coding scheme for the multi-rate coding modes is the so called Algebraic Code Excited Linear Prediction Coder, hereinafter referred to as ACELP. The multi-rate wideband ACELP coder is referred to as AMR-WB. G.722.2 also utilizes an integrated Voice Activity Detector (VAD). The sampling rate is 16 000 samples/s leading to a bit rate for the encoded bit stream of 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 or 23.85 kbit/s.

II.4 Factors affecting the performance of ASR in a mobile environment

When compared to ASR systems on desktops or servers, implementation of accurate speech recognizers on mobile devices is challenged by several factors, including: limited available storage space (language and acoustic models have to be smaller, which leads to low performance), cheap and variable microphones, often far away from speaker's mouth, no hardware support for floating point arithmetic, low processor clock-frequency (algorithms implement a tradeoff between speed and accuracy), small cache of 8-32 KB, and highly

variable and challenging acoustic environments ranging from heavy background traffic noises to a small room with reverberation of multiple speakers speaking simultaneously. Moreover, cellphones consume a lot of energy during algorithm execution [11], which is a significant factor for consideration for speech application in the developing world, where prior deployments have demonstrated that electricity issues could hamper use of mobile applications in everyday settings [15].

II.4.1 Acoustic environment

In practical applications of automatic speech recognition, speech is often distorted by a background noise. Because of this distortion, speech features are distorted, and therefore there is a mismatch between the training (clean) and testing (noisy) conditions. This mismatch severely degrades the performance of speech recognizers [16].

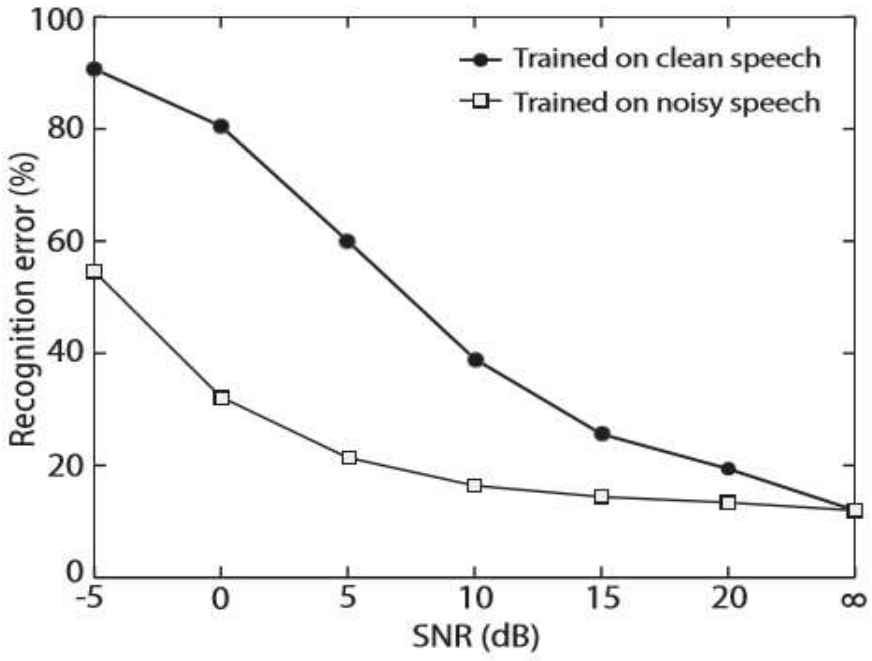


Figure II-2: The effect of different levels of background music on speech-recognition accuracy.

II.4.2 Transmission channel noise

The transmitted signal is commonly degraded by an additive noise inherent in this transmission medium. The noise can be quite destructive when the receiver moves away from the transmitter, since the signal strength diminishes. This is known as *path loss*. However, the most destructive degradation of the radio channel is due to the *multipath* phenomenon.

A) - Additive white Gaussian noise channel

The performance of a digital communication system is quantified by the probability of bit detection errors in the presence of thermal noise. In the context of wireless communications, the main source of thermal noise is addition of random signals arising from the vibration of atoms in the receiver electronics. The term additive white Gaussian noise (AWGN) originates due to the following reasons:

The noise is additive, i.e., the received signal is equal to the transmitted signal plus noise. This gives the most widely used equality in communication systems.

$$r(t)=s(t)+w(t) \qquad \text{(II-1)}$$

which is shown in Figure II.3. Moreover, this noise is statistically independent of the signal.

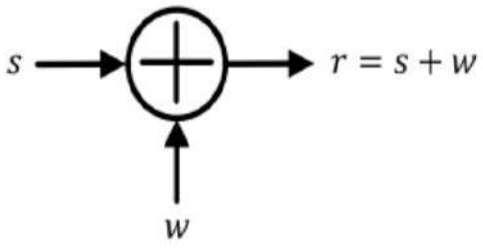


Figure II-3: Additive White Gaussian Noise.

Just like the white color which is composed of all frequencies in the visible spectrum, white noise refers to the idea that it has uniform power across the whole frequency band. As a consequence, the Power Spectral Density (PSD) of white noise is constant for all frequencies ranging from-∞ to +∞. AWGN channel, namely additive due to adding the amount of noise to send signals cannot multiply, Gaussian because noise in this channel will be amount randomly in normal. Therefore, cannot determine the amount of noise. Otherwise, when the rejecting amount of definite noise will get the receiver without any noise, white because of power for every frequency equal [19]. Therefore, at each frequency, the noise level and frequency domain are fixed. In the channel, the supposition of Gaussian noise is permitted in current causes. The noise is defined through front-end. This noise has energy spectral intensity. The AWGN channel is the perfect pattern for communication satellite. It is not the ideal model for

obtained links because of overlap, multitude, covering relief, etc. An AWGN channel can be used for imitation noise and its capacity can express in the equation (II.2):

$$C=\frac{1}{2}\log\left(1+\frac{p}{n}\right) \quad \text{(II.2)}$$

where *P* represents the maximum channel power.

B)- Rayleigh channel

Fading assumed as the attenuation produced by spread media or considered as the perversion in wireless communications. It either is already with reason shadowing from difficulty simulating the wave spread or with reason multipath spread [18]. Also, the Fading already modify with radio frequencies, time, and geographical position.

Rayleigh fading channel:

In civilized passage vehicles, building and other objects, the signal come at the recipient RX, from the sender TX on different paths. When there are different signal paths among the receiver and transmitter, A Rayleigh constrains the total signal at the recipient and can change the probability intensity Function of Rayleigh fading channel and can be expressed in equation (II.3) :

$$p_R(r)=\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) \quad \text{for } r\geq 0 \quad \text{(II.3)}$$

where:

r: the envelope amplitude of RX signal.

σ: the difference of random variable.

II.4.3 Effect of bandwidth

The bandwidth permitted to the signal is often limited. For instance, older analog telephone lines only transmit the 300–3400-Hz frequency band of the signal,the remaining frequencies are filtered out prior to transmission. Modern cellphone and VoIP channels can transmit higher bandwidths (e.g., 8 kHz), but still the most commonly used cut-off frequency is 4 kHz. This can result in significant loss of useful spectral information as shown by the example in Figure II-4.

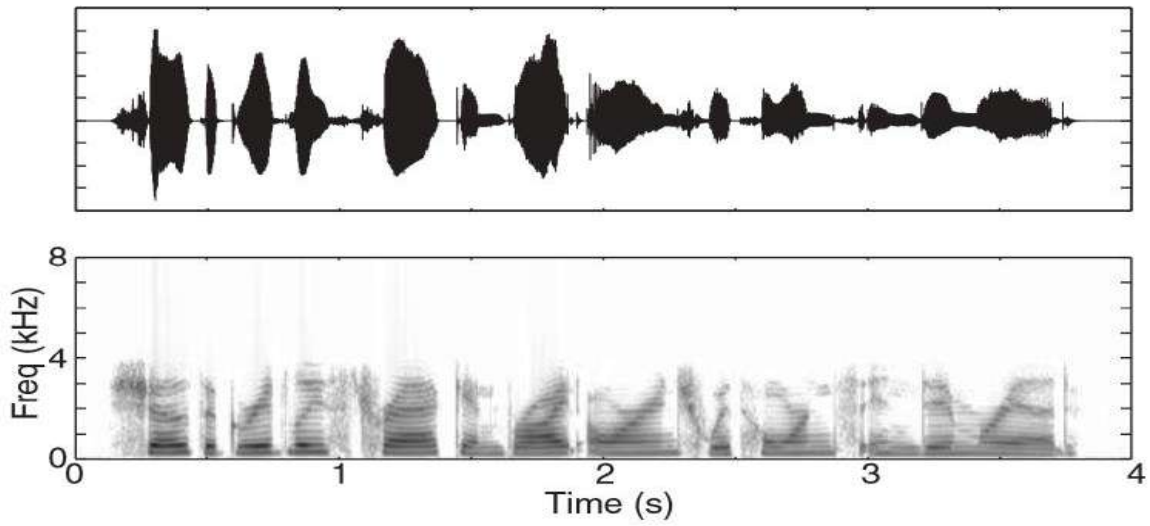


Figure II-4: The speech signal has been limited by bandwidth.

Communication channels introduce impairments to the transmitted signal like bandwidth limitation and speech coding introduce different distortions in the original speech, which augments the nondesirable within-speaker variability. Besides, mismatch between enrolment and test utterances can be generated if these are transmitted through different communication channels, causing a decrease in speaker recognition performance. Although the effects of coded-decoded NB and WB speech have not been compared before for automatic speaker recognition—this is one of the purposes of this our work—the advantages of the frequencies beyond NB have already been assessed for clean, unprocessed speech [2].

II.5. Different speaker recognition over mobile network

a) Embedded mobile speaker recognition

In the case of embedded mobile speaker recognition, the entire process of speaker recognition, including feature extraction and search happens locally on the mobile device, as shown in Figure II.5. However, it is important to note that even though feature extraction is a computationally intensive operation, it consumes only 2% of the processing time in case of medium-sized vocabularies and even less for large vocabulary recognition tasks [11].

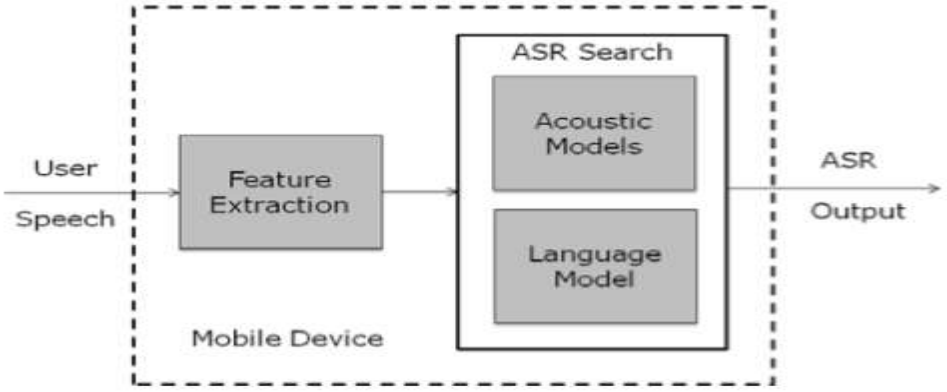


Figure II-5: Embedded mobile speaker recognition.

Advantages

The main advantage of this mode is that it does not rely on any communication with a central server, and hence can work in conditions where there are no networks. Thus, the ASR system is always ready for use, and since the audio data is recorded and decoded on the same device, the recognition is not affected by the quality of the data transmitted. Moreover, there are no additional costs involved with transmitting and receiving information back from the server. Most importantly, such systems are not affected by the latency that's involved with transmitting and receiving information from a server. Latency becomes a major issue in developing regions because cellular data connections in such regions are costly, slow and unreliable with frequent call drops.

Disadvantages

The disadvantage of this approach is that mobile devices are still not comparable to full-fledged servers that can perform complex computations. They lack in terms of speed and memory, which in turn restricts the type of applications which can be supported [12]. To achieve reliable performance, modifications need to be made to every sub-system of the ASR to take both factors into account.

b) Network speaker recognition

Practically all complications caused by the resource limitations on mobile devices can be avoided by shifting both the ASR front-end and the back-end from the terminal to the remote server. Such an ASR architecture is referred in the literature to as network speech recognition Figure II.6. Unlike embedded ASR, the NSR architecture can augment not only PDAs but

also “thin” terminals, e.g. cellular phones or regular phones, with a very large vocabulary ASR. Another advantage of NSR relies in the fact that it can provide access to recognizers based on different grammars or even different languages. Furthermore, the content of the ASR vocabulary often may be confidential, thus prohibiting its local installation. Finally, NSR allows a seamless upgrade and modification of the recognition engine.

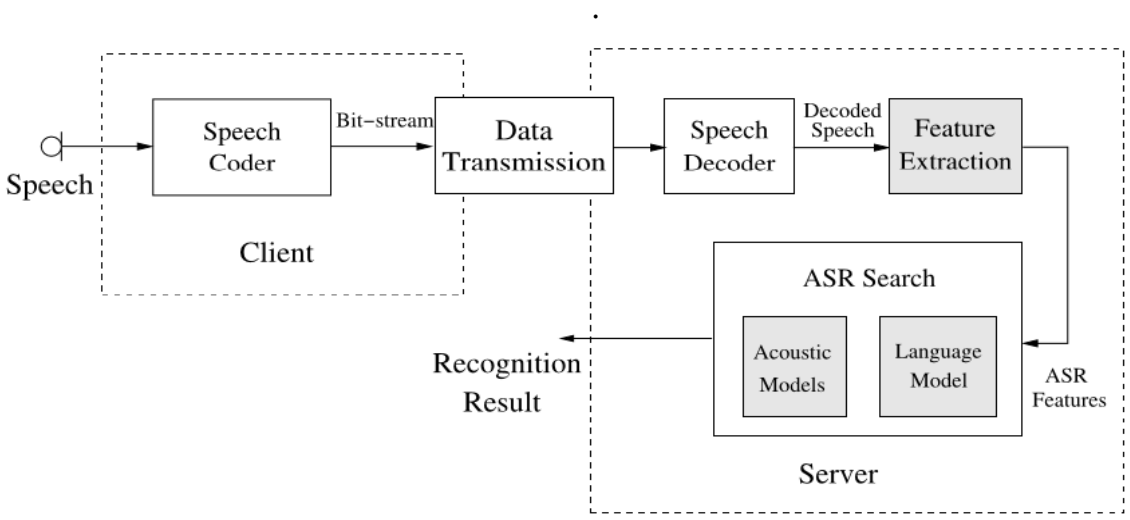


Figure II-6: Network speaker recognition.

Advantages

Network speaker recognition is a commonly used mode, and has several advantages to it: the burden of post-processing audio and extracting meaning from it is given to a powerful dedicated machine. This machine is generally a high configuration server capable of doing complex computation fast, which is essential for real-time speech recognition systems. This offers significant advantage in terms of speed and accuracy. In addition, the implementation of ASR may be confidential to the developers, and this approach avoids any local distribution or installation, which provides an easy way to upgrade or modify the central speech recognition system. Another significant advantage of this mode is that it can be used for speech recognition with low-end mobile devices such as cheap cellphones. This can be particularly useful for users in the developing region who own low-end cellphones that are only capable of making and receiving voice calls, and may not support complex speech recognition tasks

.

Disadvantages

In spite of the advantage of increased resources, this mode of recognition has some disadvantages to it that makes this mode less attractive. One of the most important cons of this approach is the performance degradation caused by low bitrate codecs, which becomes severe in the presence of data transmission errors or background noise [11]. Another disadvantage of this mode is that acoustic models on the central server need to account for large variations in the different channels (quality of microphone and audio card) on each terminal mobile device, which spans a large number of conditions. Furthermore, an important issue related to ASR design in this mode is the server-side architecture which should be capable of supporting requesting from hundreds of clients simultaneously and processing them without any additional queuing delays. Another reason that makes this option less attractive is that each data transfer over the telephone network can cost money for the end user. This is a significant problem for users in the rural parts of developing regions who would want to use applications that provide a continual service with no or just a one-time fee. Server-based speech dialog systems, as commonly used in customer self care, are a subclass of this mode.

C) Distributed speaker recognition

Distributed speaker recognition represents a client-server architecture, where the recognition process is shared between the terminal device and the server. The computationally less costly parts, i.e. feature extraction and vector quantization are performed on the terminal device, whilst the computation of temporal derivatives and the costly ASR search are performed on the remote server (see Figure II.7).

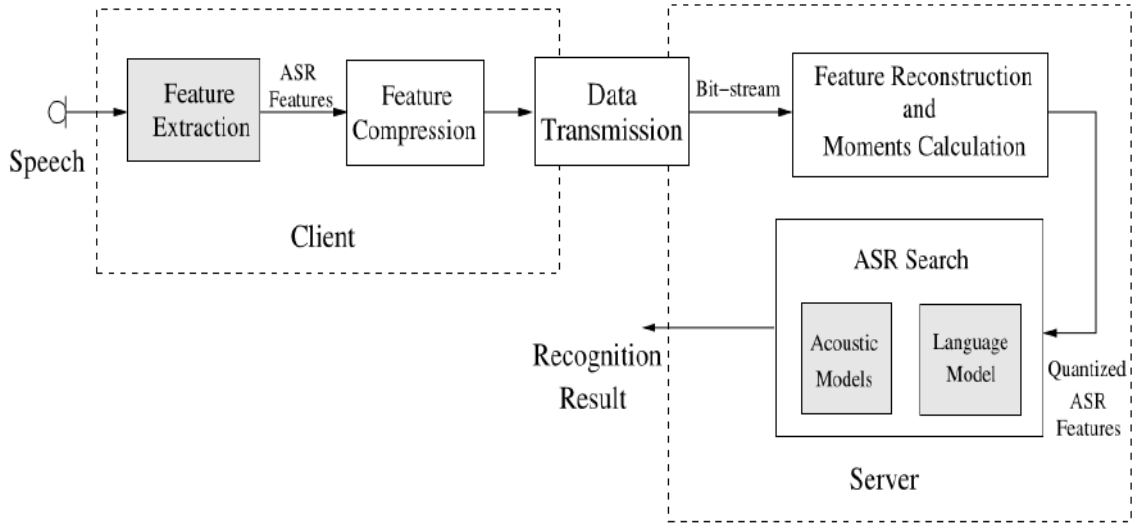


Figure II-7: distributed speaker recognition.

Advantages

The advantages of this mode are similar to those of the previous mode where the entire speech recognition is done at the server end. In addition, since ASR does not really need high quality speech, but rather a set of acoustic parameters, such as feature vectors, this mode can benefit from no loss in data due to speech coding, transmission and decoding at low bit-rates, as in the previous mode. Subsequently, better methods to improve word error rates (or similar metrics depending on the requirements of the specific task) can be developed, since normalization and adaptation can be specific to device and speech codec.

Disadvantages

The major disadvantage of this mode still remains cost and the need of continuous and reliable cellular connection, as in mode 2, although the requirements are somewhat relaxed. Also, since the feature extraction is happening on the mobile device and the ASR search takes place on the server end, there's a need for standardized feature extraction processes that account for variabilities arising due to differences in channel (microphone and audio data card), multi-linguality, variable accents, Lombard effect [13], and gender differences, etc.

II.6 Conclusion

This chapter detailed the ASR on mobile environment and the Factors affecting its performance as background noise, channel noise and bandwidth effect. We have also summarized the three existing ASR systems that are currently used for speech applications on mobile devices, and their advantages and disadvantages.

CHAPTER III

Experiments and results

III.1 Introduction

The transmission channel may introduce severe distortions that degrade the speaker recognition system performance. Particularly, there may mismatch between enrolment and test utterances. The aim of this chapter is to study the effects of speech bandwidth and speech coding on speaker verification. We will present the experiences and results obtained after evaluating the speaker recognition systems in several scenarios. Our goal is to develop speaker recognition systems with DNN and without DNN train for Bandwidth Extension (BWE) wideband (WB) and narrowband (NB) baselines. The protocols for the development and evaluation of this technique are described in this chapter based on two type of features extraction such as MFCCs and GFCCs.

III.2 MSR Identity Toolbox with MATLAB

A MATLAB toolbox for speaker-recognition research, contains a collection of MATLAB tools and routines that can be used for research and development in speaker recognition. It provides researchers with a test bed for developing new front-end and back-end techniques, allowing replicable evaluation of new advancements. It will also help newcomers in the field by lowering the "barrier to entry," enabling them to quickly build baseline systems for their experiments. Although the focus of this toolbox is on speaker recognition, it can also be used for other speech related applications such as language processing, dialect, and accent identification [20]

III.3 Experimental Setup

III.3.1 Datasets Description

a) NIST 2008

NIST 2008 speaker recognition evaluation test Set,was developed by the Linguistic Data Consortium (LDC) and National Institute of Standards and Technology (NIST). It contains 942 hours of multilingual telephone speech and English interview speech along with transcripts and other materials used as test data in the 2008 NIST Speaker Recognition Evaluation (SRE). The 2008 evaluation was distinguished from prior evaluations, in particular those in 2005 and 2006, by including not only conversational telephone speech data but also conversational speech data of comparable duration recorded over a microphone channel involving an interview scenario. The speech data in this release was collected in 2007 by LDC at its human subjects data collection

Laboratories in Philadelphia and by the International Computer Science Institute (ICSI) at the university of california, berkley. This collection was part of the Mixer 5 project, which was designed to support the development of robust speaker recognition technology by providing carefully collected and audited speech from a large pool of speakers recorded simultaneously across numerous microphones and in different communicative situations and/or in multiple languages [21].

III.3.2 Block diagram of ASV system without DNN

The i-vector approach to speaker recognition can be divided into three phases: i-vector feature extraction, inter-session variability compensation and G-PLDA modeling and scoring using the batch likelihood ratio. The block diagram of the proposed ASV system is shown in Fig III-1

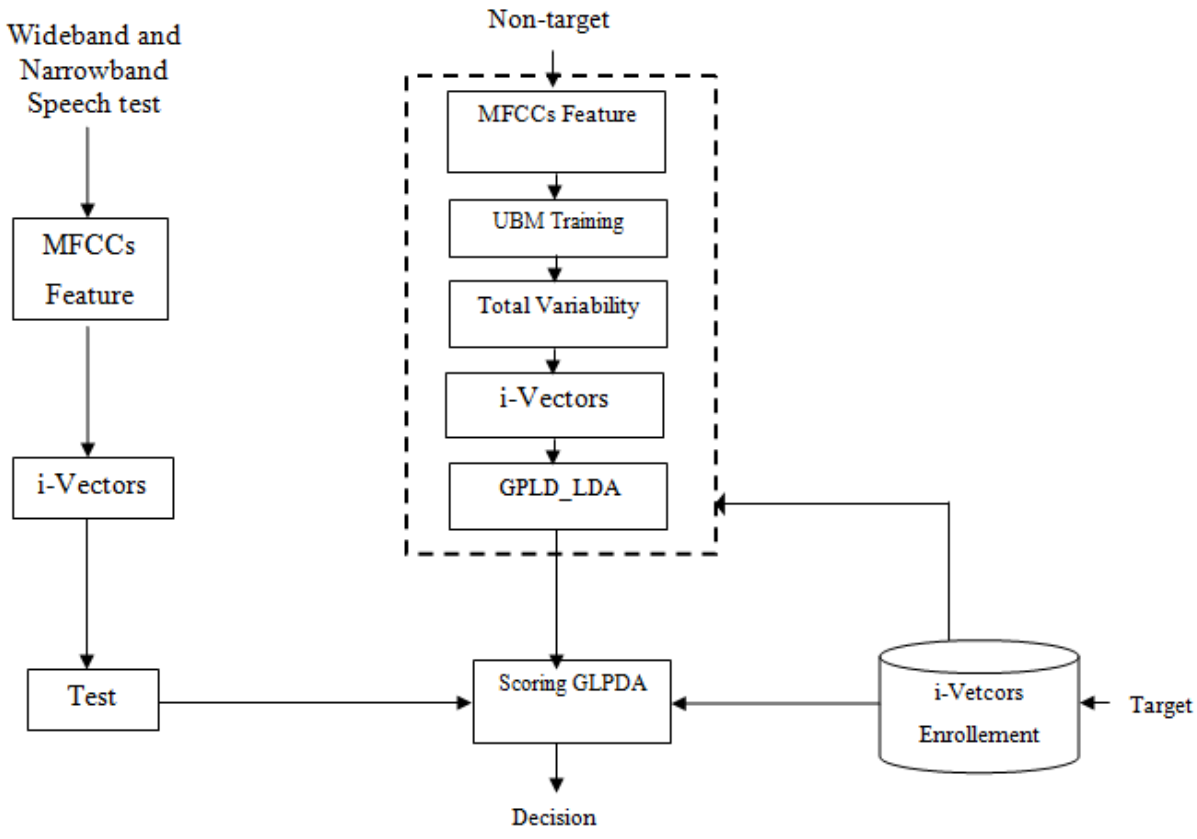


Figure III-1: Block diagram of ASV system without DNN.

a) Feature extraction

In this work, MFCC and GFCC features were computed in frames of 25 ms at intervals of 10 ms. The MFCCs and GFCCs set comprised 19 coefficients, including the zero (energy) coefficient.The 60 dimensional features are obtained by including log-energy, delta and acceleration coefficients. Finally, the features are processed through Cepstral Mean and Variance Normalization (CMVN) utilizing a 3-sec sliding window.

b) Speaker modelling

The speaker models were obtained from clean training speech data. The clean waveforms are transcoded by passing them through a coding and decoding AMR-WB and AMR-NB codecs. After feature extraction, each speaker model is adapted from a 512-component in which the UBMs are trained using the entire database. For the total variability matrix training, the UBM training dataset is used. The EM training is performed throughout five iterations. We use 400 total factors (i.e., the i-vector size is 400) then LDA is applied to reduce the dimension of the i-vector to 200, then length normalization is applied. In the process of variability compensation and scoring, a G-PLDA model with adding noise is used. In practice, the MSR Identity Toolbox is used to implement the i-vector G-PLDA processing.

c) Performance measure

The performance of the verification system is evaluated in terms of the Equal Error Rate (EER). EER corresponds to the detection threshold for equal False Alarm Rate (FAR) and False Rejection Rates(FRR).

III.3.3 Block diagram of ASV system based on DNN

In the Figure III-2, we train a DNN-i-vector system on NB and WB feature to obtain the BWE feature.

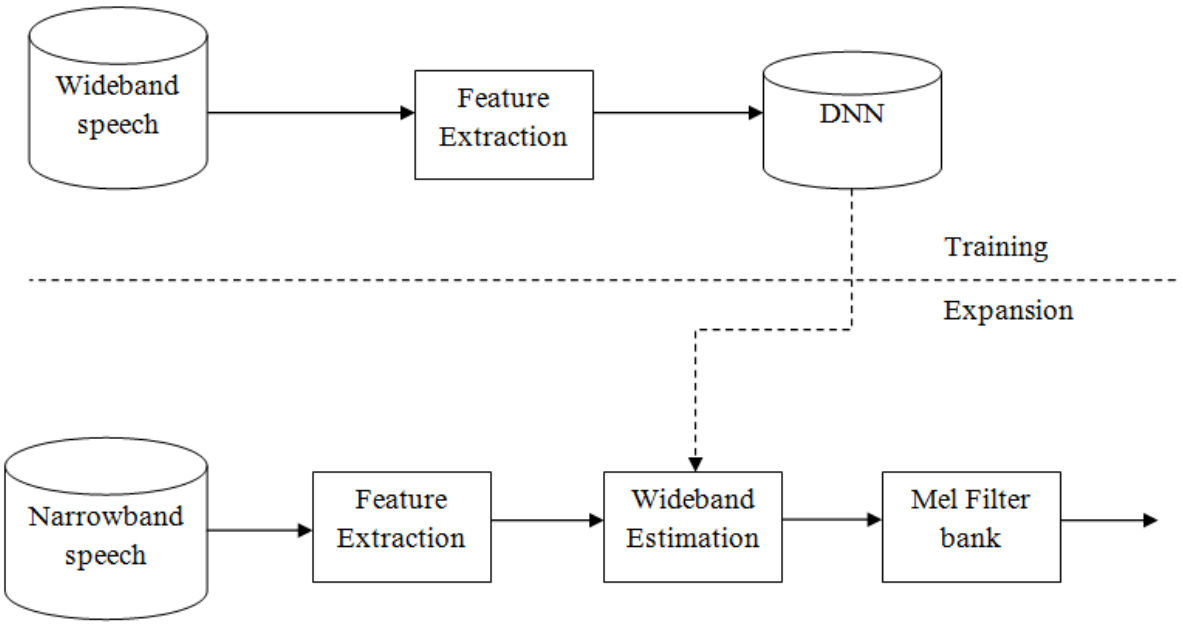


Figure III-2:block diagram of DNN-BWE.

III.4 Experiments and results

For this section, we conducted a series of experiments in order to measure the influence of bandwidth for the performance of ASV system based on three different speech codec (NB, WB and BWE).

III.4.1 ASV performance in Narrowband speech (NB)

In the first experiment, we assess the accuracy of our ASV system under different AMR Multi-Rates Narrowband speech codecs based on mismatched condition two feature extraction techniques (MFCCs and GFCCs). The results are shown in Figure III.2 for the ASV system experiments in terms of EER. From the results of Figure III.2, it can be seen that the GFCCs better performance in comparison with the MFCCs over all codecs and over all the training conditions.

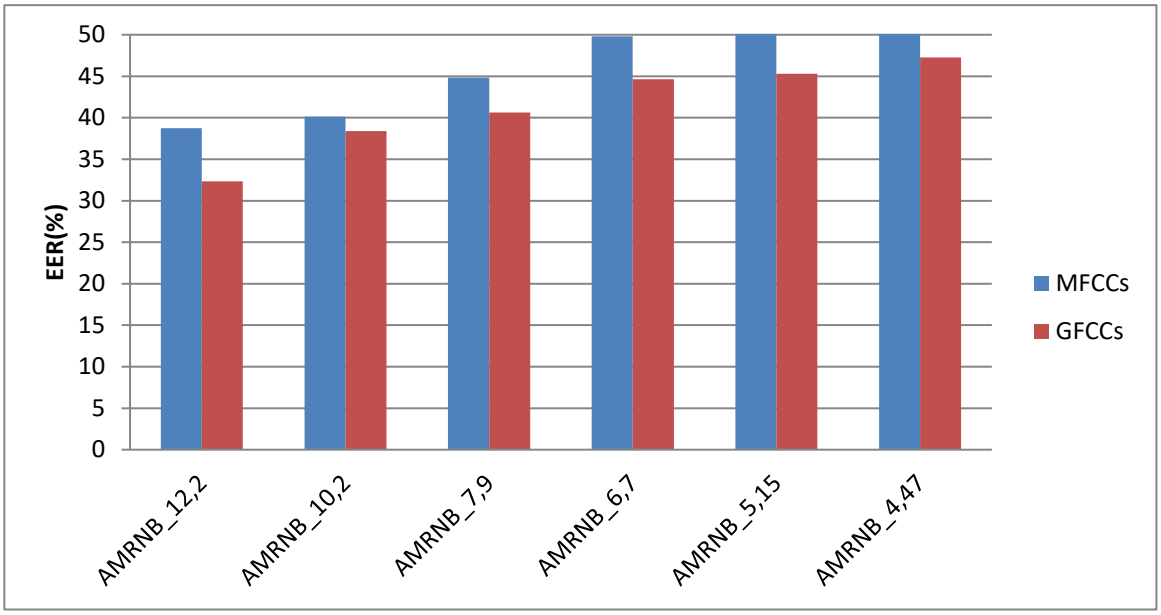


Figure III-3: Comparison of speaker verification system performance using MFCCs and GFCCs features under different AMR Narrowband speech codecs.

III.4.2 ASV performance in wideband speech (WB)

In other experiment, when we used the different AMR wideband speech codecs based on mismatched condition two feature extraction techniques (MFCCs and GFCCs). Figure III-4 present results for the ASV system experiments in terms of EER in wideband speech codecs. It can be seen that the GFCCs better performance in comparison with the MFCCs over all codecs and over all the training conditions.

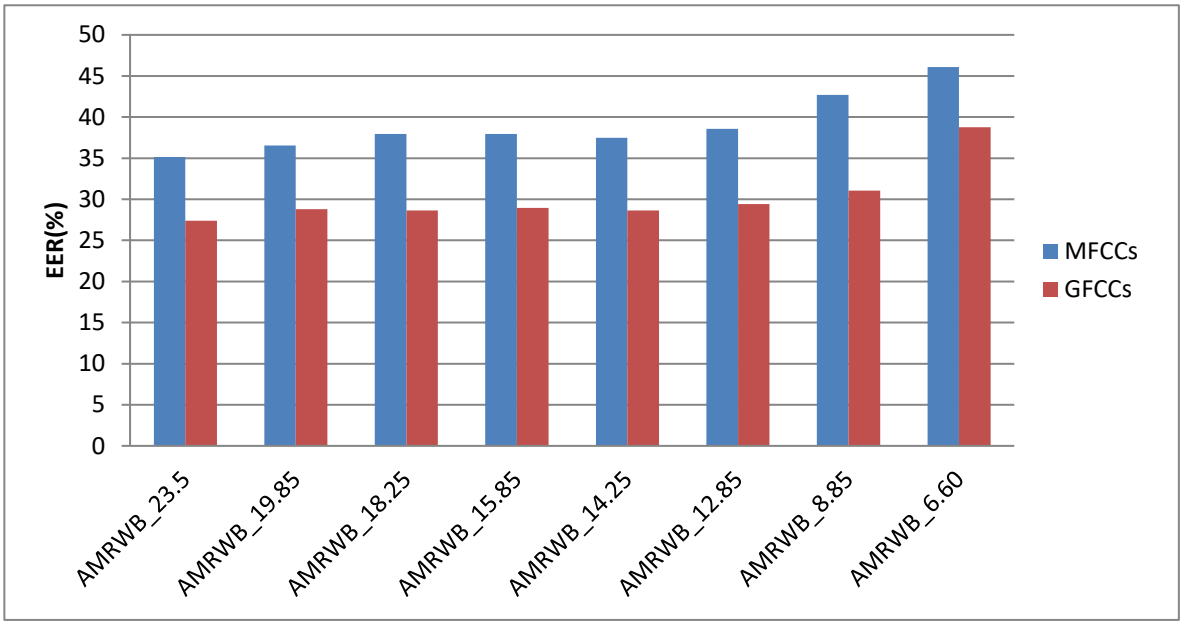


Figure III-4: Comparison of speaker verification system performance using MFCCs and GFCCs features under different AMR wideband speech codecs.

III.4.3 ASV performance in BWE

In this section, we compare the performance between two different categories of systems based on feature extractions (MFCCs and GFCCs). The first category is based on i-victor GPLDA modelling. The second category is based on the combination of DNN and i-vector-GPLDA. The results are shown in Table III-1 and Table III-2 for the ASV systems experiments in terms of EER. Frome the Table III-1, we can see that WB provides slightly improved the performance of ASV system compared to NB baseline under two type of features extraction MFCC and GFCC. The proposed BWE achieves much better performance over the baseline system when using WB and NB under two type of features extraction MFCC and GFCC.

Tableau III-1: Performance comparison between BWE, WB and NB baseline based on i-vector-GPLDA system using different features extraction (MFCC and GFCC).

i-vector GPLDA system	EER(%)	
	MFCC	GFCC
NB baseline	38.73	32.25
WB	35.25	25.5
BWE	29.60	21.85

Frome the Table III.2, we can see also that WB improved the performance of ASV system compared to NB baseline under two type of features extraction MFCC and GFCC. Here, again we observe the superiority of the proposed BWE system over the baseline system WB and NB.

Tableau III-2: Performance comparison between BWE, WB and NB baseline based on DNN i-vector-GPLDA system using different features extraction (MFCC and GFCC).

DNN-i-vector GPLDA system	EER(%)	
	MFCC	GFCC
NB Baseline	34.60	28.65
WB	28.50	20.35
BWE	24.75	19.90

III.5.Conclusion

In this chapter, we presented three independent case studies: (1) training the i-vector-GPLDA speaker recognition system on under different AMR Multi-Rate narrowband and AMR Multi-Rate wideband speech codecs, (2) training the i-vector-GPLDA speaker recognition system on NB, WB and BW extended speech and (3) training the DNN i-vector-GPLDA system on NB, WB and BW extended speech by using the different feature extraction MFCC and GFCC. We observed that, for NIST 2008 dataset, (NB, WB and BWE data) to train the i-vector extractor and GPLDA backend did not give improvements over the DNN i-vector extractor and GPLDA backend.

Conclusion

As part of this work, we tackled the problem of the robustness of speaker recognition system in the real mobile environment. We examined and compared the effect of speech bandwidth by NB, WB, and BWE and speech coding on the speaker recognition performance.

We started with the study the influence of speech coding by using the AMR narrowband and AMR wideband speech codecs to different type of feature extraction. The AMR-WB at 23.05 kbit/s offer statistically significantly bettre speaker recognition performance than the AMR-NB at 4.75 kbit/s. The low bitrate coding of the AMR-WB significantly reduces the speaker recognition with respect to the other NB codecs based on the GFCC feature.

A comparative study is included about the the efficiency of these two systems such as the i-vector-GPLDA system and DNN-i vector GPLDA system. The various experimental results that we obtained show a reduction in the EER error rate. Through a series of experiments using NIST 2008 corpora of narrowband, wideband and wideband extension, we demonstrated that the proposed system with DNN-i vector GPLDA is able to significantly outperform the system with i-vector GPLDA to reduce the effect of speech wideband.

As perspectives for this work, it will be interesting to:

a) Make the proposed parameterization techniques cooperate with other types of feature such as long term feature.

b) Study the impact of background noises for speaker recognition system over mobile environment.

Bibliography

[1]- Hansen JHL, Hasan T. Speaker recognition by machines and humans: A tutorial review. IEEE Signal Process Mag 32:74–99, 2015.

[2]- Gallardo LF. Human and automatic speaker recognition over telecommunication channels. Springer Science + Business Media, Singapore, 2016.

[3]- K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in Sixteenth Annual Conference of the International
Speech Communication Association, 2015.

[4] L. Allano, La Biométrie multimodale : stratégies de fusion de scores et mesures de dépendance appliquées aux bases de personnes virtuelles, PhD thesis, Institut National des Télécommunications, 2009.

[5] A.K. Jain, P. Flynn, A. Ross, Handbook of Biometrics, Springer, New York, USA, 2008.

[6] J. Fierrez-Aguilar, Adapted Fusion Schemes for Multimodal Biometric Authentication, PhD Thesis, University of Madrid, 2006.

[7] SAYOUD, Halim. Reconnaissance automatique du locuteur approche connexionniste. 2003. Thèse de doctorat.

[8] BLOUET, Raphaël. Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées. 2002. Thèse de doctorat. Rennes 1.

[9] BOCCARDI, Federico et DRIOLI, Carlo. Sound morphing with Gaussian mixture models. In : Proc. DAFx. 2001. p. 44-48.

[10] Maas, A.L., Qi, P., Xie, Z., Hannun, A.Y., Lengerich, C.T., Jurafsky, D. and Ng, A.Y., "Building DNN Acoustic Models for Large Vocabulary Speech Recognition," arXiv preprint arXiv,, 1406.7806, Jan. 2015.

[11] Schmitt, A., Zaykovskiy, D., Minker, W. Speech recognition for mobile devices. International Journal of Speech Technology, vol. 11, number 2, pp. 63-72, 2008

[12] Novak, M. Towards large vocabulary ASR on embedded platforms. In Proc. of InterSpeech, 2004.

[13] Junqua J.C. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93 (1): 510–24, 1993.

[14]Rose, R.C., Partharathy, S. A tutorial on ASR for wireless mobile devices. In International Conference on Spoken Language Processing, 2002.

[15]Kumar, A., Tewari, A., Shroff, G., Chittamuru, D., Kam, M., and Canny, J. An exploratory study of unsupervised mobile learning in rural India. In Proc. of ACM Conference on Human Factors in Computing Systems (CHI '10), Atlanta, Georgia, April 10-15, 2010.

[16]C. Mokbel, G. Chollet, Speech recognition in adverse environments: speech enhancement and spectral transformations, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (Toronto, Ontario, Canada, 1991)

[17] X. LEI, P. FAN, "On the Error Performance of M-ary Modulation Schemeson Rician Nakagami Fading Channels", Wireless Personal Communications, vol. 53, no 4, p. 591-602, 2010.

[18] N. V. Panicker, A. K. Sukesh, "BER Performance Evaluation of Different Digital Modulation Schemes for Biomedical Signal Transceivers under AWGN and Fading Channel Conditions", International Journal of Engineering and Advanced Technology (IJEAT), Vol. 3 Issue 5, June 2014.

[19] X. LEI, P. FAN, "On the Error Performance of M-ary Modulation Schemeson Rician Nakagami Fading Channels", Wireless Personal Communications, vol. 53, no 4, p. 591-602, 2010.

[39]-]- Seyed Omid Sadjadi, Malcolm Slaney, Larry Heck; "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research, 2013, https://www.microsoft.com/enus/research/wp-content/uploads/2013/09/MSR-Identity-Toolbox-v1_1.pdf

[21] - NIST Year (2008) Speaker recognition evaluation plan, Technical report, NIST. https://catalog.ldc.upenn.edu/LDC2011S08, Accessed 19 September 2020

Résumé :

Ce travail étudie le problème général posé par le système de reconnaissance du locuteur dans les environnements mobiles avec un effet d'intérêt sur la bande passante et le codage de la parole. Il s'agit d'une étude comparative sur les performances du système de reconnaissance du locuteur sous différents codecs vocaux AMR Multi-Rate et les bandes passantes (NB bande étroite, WB large bande et BEW large bande extension) basée sur l'approche DNN.

Mots clés: Reconnaissance du locuteur, Réseau neuronal profond DNN, NB, WB, BWE

Abstract:

This work studies the general problem posed by the speaker recognition system in mobile environments with an effect of interest the bandwidth and speech coding. This is a comparative study on the performance of the speaker recognition system under different AMR Multi-Rate voice codecs and the bandwidths (NB narrowband, WB wideband and BEW wideband extension) based on the DNN approach.

Keywords: Speaker recognition, Deep Neural Network, NB, WB, BWE.

الملخص :

يدرس هذا العمل المشكلة العامة التي يطرحها نظام التعرف على السماعات في البيئات المتنقلة مع تأثير الاهتمام على النطاق الترددي وتشفير الكلام. هذه دراسة مقارنة حول أداء نظام التعرف على السماعات في ظل مختلف برامج ترميز الصوت متعددة المعدلات وعرض النطاق الترددي على أساس نهجDNN

الكلمات المفتاحية:

التعرف على السماعات ، الشبكة العصبية العميقة ، NB,WB,BWE