Towards Expressive Arabic Text to Speech

By

Doaa Gamal Madany Taya

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND ELECTRICAL COMMUNICATIONS ENGINEERING

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2014

Towards Expressive Arabic Text to Speech

By
Doaa Gamal Madany Taya

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND ELECTRICAL COMMUNICATIONS ENGINEERING

Under the Supervision of

Prof. Dr. Mohsen Rashwan
.....
Professor,
Electronics and Electrical Communications Department
Faculty of Engineering, Cairo University

Dr. Hossam A. H. Fahmy
.....
Associate professor,
Electronics and Electrical Communications Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2014

Towards Expressive Arabic Text to Speech

By

Doaa Gamal Madany Taya

A Thesis Submitted to the

Faculty of Engineering at Cairo University

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

In

ELECTRONICS AND ELECTRICAL COMMUNICATIONS ENGINEERING

Approved by the

Examining Committee

_____

Dr. Khairy el-barbary, External Examiner (Suez Canal University)

_____

Prof. Sherif abdou, Internal Examiner

_____

Prof. Mohsen Rashwan, Thesis Main Advisor

_____

Dr. Hossam A. H. Fahmy, Member

FACULTY OF ENGINEERING, CAIRO UNIVERSITY

GIZA, EGYPT

2014

Engineer's Name: Doaa Gamal Madany Taya
Date of Birth: 16/11/1987
Nationality: Egyptian
E-mail: Doaa_Gamal@eng.suez.edu.eg
Phone:
Address: 6b, el-zohour district, el-sheikh zayed, Ismailia, Egypt.
Registration Date: 1/10/2010
Awarding Date
Degree: Master of Science
Department: Electronics and Electrical Communications Engineering
Supervisors:

Prof. Dr. Mohsen Rashwan

Dr. Hossam A. H. Fahmy

Examiners:

Dr. Khairy el-barbary (External Examiner)
Prof. Sherif abdou (Internal Examiner)
Prof. Mohsen Rashwan (Thesis Main Advisor)
Dr. Hossam A. H. Fahmy (Member)

Title of Thesis:

Towards Expressive Arabic Text to Speech

Key Words:
Expressive Speech Synthesis; Emotion Conversion; Prosody Conversion; Unit Selection.

Summary:
In this thesis, an emotion conversion system has been proposed for expressive Arabic speech; this system combines the transformation of both spectral and prosodic parameters. Four parameters have been transformed to obtain the desired expression: pitch, duration, energy, and spectral envelope. The effect of converting each speech parameter in our system is studied and the overall emotion conversion system performance is evaluated. In pitch conversion, the effect of using two different intonation units (words-syllables) and using different pitch detectors on the converted speech is also studied.

Acknowledgments

First and foremost, I am thankful to “Allah” for guiding me to finish this work.

Special thanks to my supervisors: Prof. Mohsen Rashwan, and Prof. Hossam Fahmy. They have been my mentors and advisors throughout this work, guiding me all the way through this important stage in my life, and developing in me the spirit of productive and moral research.

Special thanks to Prof Sherif Abdou, at faculty of computer science, Cairo University, for his great support and sincere advices.

I am grateful to my family for their continuous support, help and advice.

I would also like to express my gratitude to my professors: Dr. Ahmed Magdy, Prof Khairy el-Barbary, and Prof Atef Ghuniem at the Electrical Engineering Department, Faculty of Engineering, Suez Canal University, for their support and sincere advices.

Special thanks to my colleagues and students at Faculty of Engineering, Suez Canal University, who participated in the system evaluation and who always support and encourage me.

Many thanks are posed to The Research & Development International Company (RDI®) for providing the database used in this research and for its support of this work and participation in the system evaluation.

Special thanks to Mr. Gamal Khalaf-allah, Eng. Ebtesam, and Mr. Ahmed Farouk at Egyptian Radio and Television Union (ERTU) for their great help in recording the speech corpus.

Finally, I’m grateful to Egypt Scholars for their scientific research lectures. I have benefited a lot from these lectures.

Dedication

To my mother, my father, my brother and my sisters.

Table of Contents

List of Tables

List of Figures

List of Abbreviation

TTS	Text to Speech
MRI	Magnetic Resonance Imaging
IVR	Interactive Voice Response
MLSA	Mel-Log Spectrum Approximation
MSD	Multi-Space probability Distributions
HMM	Hidden Marcov Model
HTS	HMM-based Speech Synthesis System
LPC	Linear Predictive Coding
LSF	Line Spectral Frequencies
TD-PSOLA	Time Domain Pitch Synchronous Overlap Add
DTW	Dynamic Time Warping
RAPT	Robust Algorithm for Pitch Tracking
HSR	Harmonic-to- Subharmonic- Ratio.
MBSC	Multi-Band Summary Correlogram
GMM	Gaussian Mixture Model
EM	Expectation Maximization
ML	Maximum Likelihood
CART	Classification and Regression Tree
MOS	Mean Opinion Score
IEEE	Institute of Electrical and Electronics Engineers
ITU	International Telecommunication Union

Abstract

Emotion conversion using a small speech corpus is very important for expressive text to speech systems. Applying the unit selection paradigm for intonation conversion has been widely used for different languages. Different intonation units were used depending on the linguistic characteristics of different languages.

In this thesis, an emotion conversion system is proposed for expressive Arabic speech. This system combines the transformation of both spectral and prosodic parameters of speech based on the linguistic context. Four speech parameters are transformed to obtain the desired expression: pitch, duration, energy, and spectral envelope. The linguistic features of Arabic speech which are responsible for different intonation are studied. Unit selection is used for pitch conversion and the effect of using different intonation units and different pitch detectors is studied. We also study the effect of converting each of the four speech parameters, using our proposed system, on different expressions. Finally we evaluate the overall emotion conversion system for different expressions.

Subjective tests were carried out to evaluate the system on three target expressions: sadness, happiness and questioning. Results show the effectiveness of both syllable and word units as the basic intonation unit for pitch conversion, however using syllable unit gives higher expressiveness for sadness and happiness. Results also show that converting pitch contours using our system is dominant for happiness and questioning and highly affects the sadness, while duration conversion affects only sadness, and spectral conversion affects only happiness.

After analyzing the training corpus; we have proposed decreasing the energy level for sadness and results show the effectiveness of energy level decrease in sadness.

Finally, the evaluation of the overall emotion conversion system for expressive speech shows that the proposed system managed to add an acceptable expressiveness in Arabic speech with a good quality ($MOS \approx 3$) for sadness and happiness. The same results can be obtained for questioning if only the pitch contour is converted, since spectral conversion degrades the output quality of questioning without increasing the expressiveness and duration conversion has no effect on questioning.

Chapter1: Introduction

Text to speech system (TTS) is a machinery system which generates human-like speech from any input text. Speech synthesizers used in TTS have been developed over the years as the memory resources become available and cheaper; as a result, large enhancements in the quality and the intelligibility of the synthesized speech have been achieved.

The task of learning a TTS machine to convert text into speech can be represented as learning a child reading a text aloud. Imagine if this child reads all expressive sentences in a neutral manner without illustrating questioning, surprise, sadness, happiness, fear, and other expressions; the utterances may be unintelligible. The same thing for the TTS machine, which makes adding expressiveness to TTS systems required to increase the naturalness and intelligibility of the system.

1.1 Expressive TTS literature review

Many studies in the literature were done in the field of expressive TTS, starting with studying the acoustic correlates of emotions to find speech parameters which are responsible for emotion [1-5], passing by designing good corpora for emotional speech [6, 7], and finally studying how to generate expressive speech.

1.1.1 *Acoustic correlates of emotions*

Several studies were made to find the effective parameters of the speech signal that are responsible for different emotions. Results in [1] proved that there are systematic changes in the spectral envelopes of neutral and emotional speech signals. A copy-synthesis approach is used in [2-5] and their results proved that modifying only the prosodic parameters(pitch-duration- intensity), or only the spectral parameters isn't sufficient to achieve a successful emotion conversion, so both prosodic and spectral parameters should be modified to achieve a successful emotion conversion, however combining both spectral and prosodic transformation increases the identification rate of each emotion at the expense of decreasing the output speech quality [5].

1.1.2 *Generating expressive speech*

Three approaches for adding expressiveness to TTS systems were proposed in the literature for different speech synthesizers. These approaches are: expressive unit selection (or playback), expressive HMM-based synthesis, and emotion conversion; as summarized in Figure 1-1.

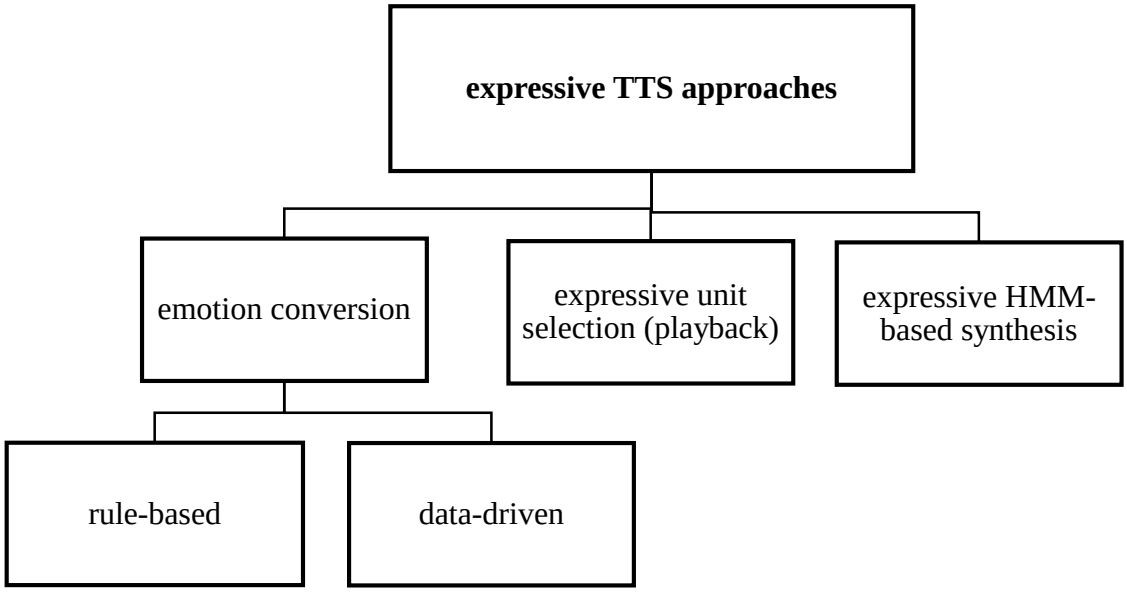


Figure 1-1 literature work of generating expressive speech

1.1.2.1 *Expressive unit selection*

Expressive unit selection, also called playback, is an approach used to obtain expressive speech in concatenative synthesizers; where a database containing different speaking styles is recorded, and at the synthesis time; appropriate units are selected from this database and concatenated to obtain expressive speech.

This approach was used in early diphone concatenative synthesizers [8]; where a diphone inventory was recorded in different speaking styles by the same speaker, and at the synthesis time target diphones were selected from the desired emotion inventory and concatenated to obtain expressive speech.

Recent unit selection concatenative synthesizers, which store large database of different speech units (phonemes, diphones, words, and phrases) to reduce the concatenation distortion, make use of the playback approach to achieve expressiveness [9-12].

1.1.2.2 *Expressive HMM-based synthesis*

Another approach for expressive TTS is to model the speech parameters (spectrum, excitation, and phoneme duration) and linguistic features of different emotions using HMMs. At the synthesis time, these parameters are used to synthesize the speech in the desired style [13, 14].

1.1.2.3 *Emotion conversion*

Emotion conversion is used for adding expressiveness to the text to speech system using a small corpus recorded in different styles. Emotion conversion for expressive TTS is independent of the synthesizer type, unlike the playback and HMM-based approaches; where

emotion conversion approach can be applied to the output speech of any synthesizer or to the parameters used for speech generation in a parametric HMM synthesizer.

In emotion conversion approach, a set of speech parameters are transformed to the desired emotion. Emotion conversion uses either rule-based techniques or statistical techniques to convert the neutral speech to the desired emotion. A major issue in emotion conversion is that it is difficult to compare the results of two different works, since the emotion conversion system is affected by the linguistic features of the language, the quality and expressiveness of the training data, and the size of the database (for statistical approach).

1.1.2.3.1 Rule-based emotion conversion

Rule-based approach was used to convert an input utterance to the desired emotion using a set of pre-determined rules [15-21]. Rules are designed for different parameters of speech signal such as F0 level and range, speech tempo, possibly loudness, voice quality, and the number and duration of pauses.

1.1.2.3.2 Data-driven emotion conversion (statistical approach)

In data-driven emotion conversion, an efficient-size of parallel training corpus is used to train statistical models or to build a unit selection framework, which are used later to convert the neutral speech into the desired emotion. Two main tasks are exist in the statistical emotion conversion: mapping neutral speech parameters to their emotional values, and then modifying them in the speech waveform without affecting the quality of the converted waveform.

Different machine learning techniques have been proposed to map the speech parameters to different emotions. Several techniques were explored for pitch conversion such as Linear Modification Model (LMM), Gaussian Mixture Model (GMM), and Classification and Regression Tree (CART). A comparison between those methods was presented in [22] and results show that the best of them is CART if trained with a large corpus. In [23] Hidden Marcov Models (HMMs) and unit selection were proposed for pitch modification, and results show that unit selection outperforms HMM-based pitch conversion.

For spectral envelope mapping, the most popular learning technique used for emotion conversion is GMM [24-26]. A comparison between three methods for transforming spectral envelopes from neutral to emotional speech is presented in [27] and results show that weighted frame mapping and GMM based transformations are slightly better than the weighted codebook mapping.

Several signal processing techniques for prosody modification were introduced in the literature. The most popular one of them is PSOLA (TD-PSOLA, FD-PSOLA, and LP-PSOLA) [28]. A prosody modification using the instants of significant excitation is presented in [29]. Some approaches for prosody modification represent the speech signal in a parametric form like STRAIGHT [30, 31] or HSM [32], and then modify the prosody parameters.

1.2 Motivation and contribution of the thesis

In this work we employ data-driven emotion conversion to obtain expressive Arabic speech. As mentioned above that emotion conversion is affected by the linguistic features of the language. The Arabic language is classified as a stress timed language and word stress is predictable and regular. In contrast to English, which is also stress-timed language, the unstressed syllables are pronounced more clearly with neutral vowels.

In our work, the linguistic characteristics of Arabic language which affects the expression are studied. An emotion conversion system is proposed for expressive speech, this system combines the transformation of both spectral and prosodic parameters. Pitch, duration, energy, and spectral envelope are transformed to obtain the desired expression. In pitch conversion, we evaluate the effect of using different intonation units (words-syllables) and using different pitch detectors on the quality and the expressiveness of the output speech.

The effect of changing different speech parameters using our system is evaluated and the overall system performance is also evaluated.

1.3 Thesis Organization

First, in chapter 2 and chapter 3, we give a background in text to speech systems and a more detailed review of expressiveness in TTS. Chapter 4 includes the overall description of the emotion conversion system used for expressive Arabic speech, while the system modules are illustrated in detail in chapters 5. Chapter 6 shows the experiments carried out to evaluate the system and their results, and chapter 7 includes the final conclusions and proposals for future work.

Chapter2: Overview of Text To Speech systems

Text to Speech system (TTS) is a machinery system which generates speech from an input text. The task of learning a machine how to convert text into speech can be represented as learning a child reading a text aloud. Three tasks are required for reading aloud; the first task is to decode the input text to realize its phonetic transcription (or speakable form), the second task is to convey the intonation of the sentences properly to sound natural, and the final task is pronouncing the phonetic transcription of the text with the appropriate intonation. The same for TTS, it can be divided into three main blocks: Text and Phonetic Analysis, Prosody Generation, and Acoustic Synthesis [33]. The basic blocks of TTS are shown in Figure 2-1.

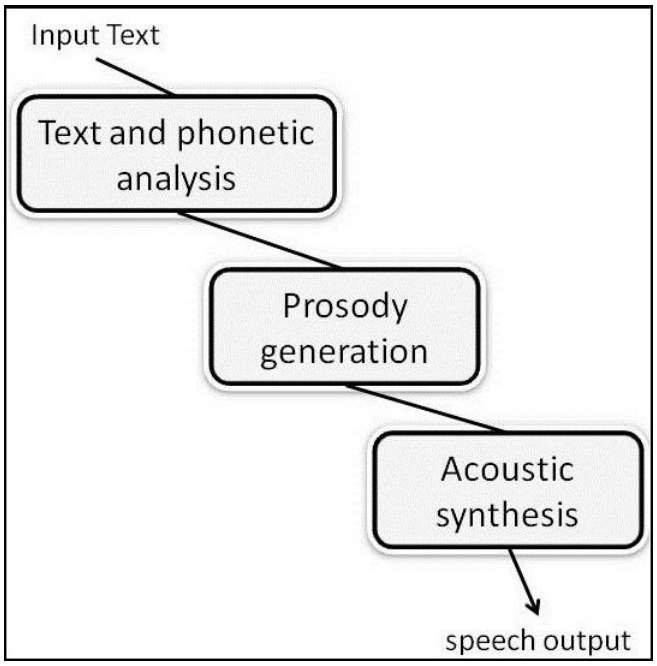


Figure 2-1 Text To Speech system.

2.1 Text and phonetic Analysis

Text and phonetic analysis module in TTS is responsible for analyzing the input text to generate its phonetic transcription and the main information needed for the prosody analysis and acoustic synthesis modules. Figure 2-2 shows the basic blocks of the text and phonetic analysis module, and they are described in the following subsections.

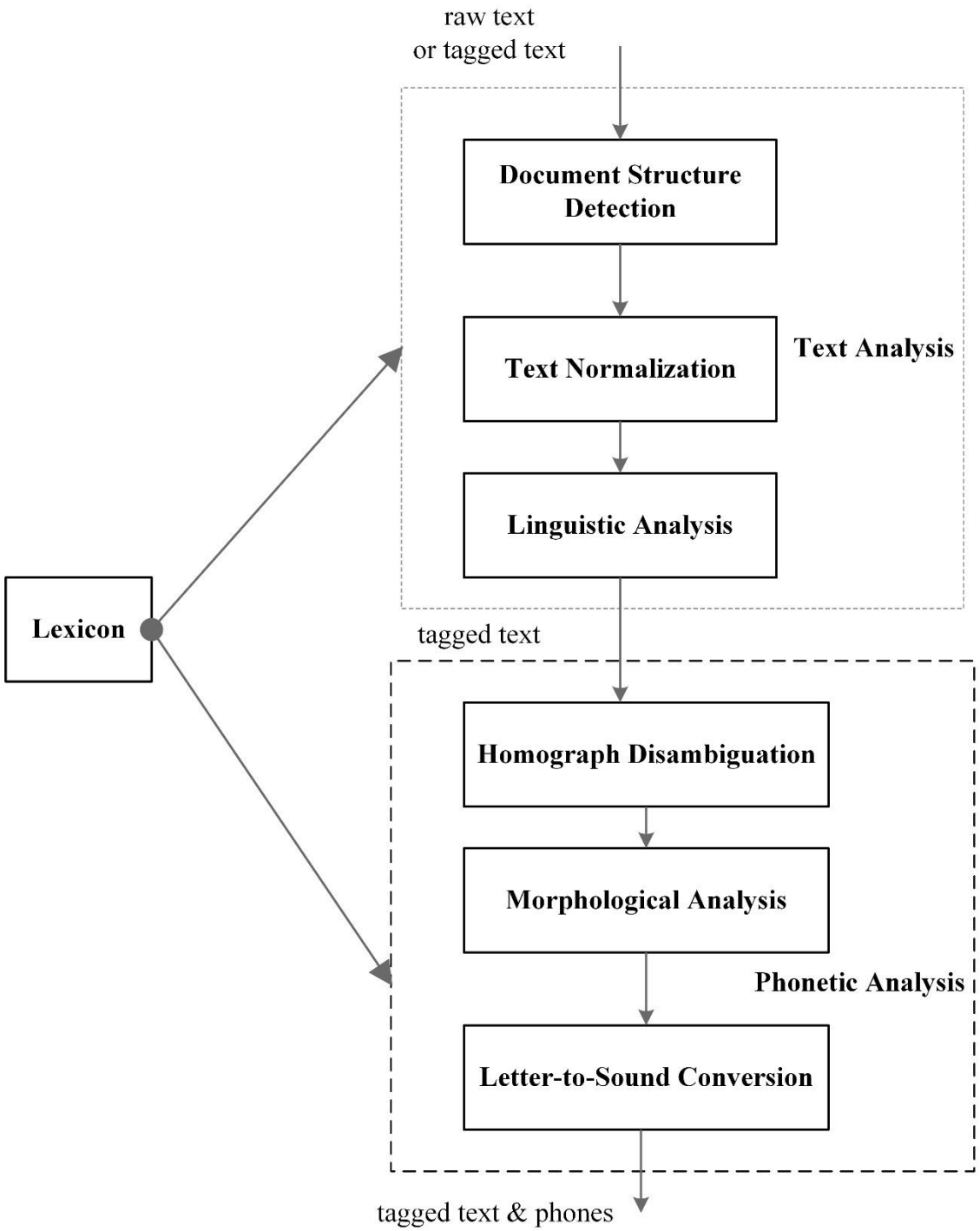


Figure 2-2 Modularized function blocks for text and phonetic analysis [33].

2.1.1 *LEXICON*

TTS lexicon or dictionary is considered the most important resource for text and phonetic analysis module. TTS lexicon should provide the text and phonetic analysis module with the following information [33]:

- *Inflected forms of lexicon entries.*
- *Phonetic pronunciations for each lexicon entry.*
- *Morphological analysis capability.*
- *Abbreviation and acronym expansion and pronunciation.*
- *Attributes indicating word status, including proper-name tagging, and other special properties.*
- *List of speakable names of all common single characters.*
- *Word part-of-speech (POS) and other syntactic/semantic attributes.*
- *Other special features, e.g., how likely a word is to be accented, etc.*

2.1.2 *DOCUMENT STRUCTURE DETECTION*

TTS detects the structure of the input document to control and regulate prosody for different document structures. For example, the pitch range at the start of a new paragraph should be higher than at the mid-paragraph sentences, and it narrows at the final few clauses. Different document structures should be detected by TTS such as:

- *Chapter and Section Headers*
- *Lists or bulleted items*
- *Paragraphs*
- *Sentences*
- *E-mail*
- *Web Pages*
- *Dialog Turns and Speech Acts*

2.1.3 *TEXT NORMALIZATION*

Text normalization (TN) is the process of transforming abbreviations or non-alphabetic symbols into normalized orthography. This process is performed using lexicon. Table 2-1 shows different types of symbols in Arabic and their normalized texts.

Table 2-1 Example of text normalization in Arabic

Text type	Example	normalized orthography
Abbreviations and Acronyms	أ.د.	الأستاذ الدكتور
Numbers	375	ثلاثمائه وخمسة وسبعون
Phone Numbers	02233331414	صفر–إثنين–إثنين–ثلاثه– ثلاثه– ثلاثه– ثلاثه–واحد–أربعه–واحد–أربعه
Dates	1992/5/27	السابع والعشرون من مايو عام ألف وتسعمائه واثنين وتسعين
Times	8:30 ص	الثامنه والنصف صباحًا

2.1.4 *LINGUISTIC ANALYSIS*

Linguistic analysis is used for syntactic and semantic parsing. TTS can use natural language processing (NLP) systems, which produce structural and semantic information about the sentences, for linguistic analysis.

Linguistic analysis supports the phonetic analysis (homograph disambiguation-morphological analysis) and prosody generation modules with the following information [33]:

- Word part of speech (POS) or word type (e.g., noun, verb, or preposition).
- Word sense, e.g., *river bank* vs. *money bank*.
- Phrasal cohesion of words, such as idioms, syntactic phrases, clauses, sentences.
- Modification relations among words.
- Anaphora (co-reference) and synonymy among words and phrases.
- Syntactic type identification, such as questions, quotes, commands, etc.
- Semantic focus identification (emphasis).
- Semantic type and speech act identification, such as requesting, informing, narrating, etc.
- Genre and style analysis.

2.1.5 *HOMOGRAPH DISAMBIGUATION*

In written languages, some words have the same spelling but vary in pronunciation, since they have different syntactic/semantic meanings. POS tags and semantic analysis are used in English to resolve homograph disambiguation. In Arabic TTS, Automatic Arabic Phonetic Transcriptor (Diacritizer/Vowelizer) is used for all words in the text to resolve Homograph variation. An example of this in Arabic is the following sentences:

(تعود المدرس على ضرب التلاميذ – ضرب المدرس التلاميذ)

After using the Automatic Arabic Diacritizer, these sentences will be:

(تَعَوَّدَ الْمُدَرِّسُ عَلَى ضَرْبِ التَّلَامِيذِ – ضَرَبَ الْمُدَرِّسُ التَّلَامِيذْ)

It is obvious that the word (ضرب) in these sentences has two different pronunciation; although they have the same spelling.

2.1.6 *MORPHOLOGICAL ANALYSIS*

In TTS, when an input word doesn't exist in the lexicon, and can be analyzed into shorter units (morphemes), morphological analyzer is used to analyze this word in terms of stems and affixes. The pronunciations of these morphemes can be then combined with some adjustment to obtain the pronunciation of the input word.

2.1.7 *LETTER-TO-SOUND CONVERSION*

Letter-to-sound conversion, also known as grapheme-to-phoneme conversion, is the process of converting an input text to its phonetic pronunciation form via dictionary lookup and using different information generated from other blocks of text and phonetic analysis module. For example, the phonetic pronunciation of Arabic word (ظلام) is (/D`ala:m/), and of the English word (dogs) is (/ˈdɒgz/).

When the dictionary lookup fails to find an input word, a set of rules for pronunciation is used to obtain the phonetic pronunciation of this word.

2.2 Prosody Generation

Prosody generation provides the TTS system of how to utter different texts. Sheridan illustrated the importance of prosody generation as follows [34]:

"Children are taught to read sentences, which they do not understand; and as it is impossible to lay the emphasis right, without perfectly comprehending the meaning of what one reads, they get a habit either of reading in a monotone, or if they attempt to distinguish one word from the rest, as the emphasis falls at random, the sense is usually perverted, or changed into nonsense."

Prosody or intonation of an utterance can be generated using different parameters of speech signal such as pitch, phoneme duration, and loudness. The value of these parameters

can be estimated from an input text using either a set of pre-designed rules or statistical methods.

2.3 Acoustic Synthesis

The acoustic synthesis module is responsible for generating the output speech waveform. The input to this module is the phonetic transcription of the text and its associated prosody. For high-quality speech production, the original text with tags are included in the input to the acoustic synthesizer.

Acoustic synthesizers have been developed over the years, they can be divided into three generations depending on the quality of their output speech [33, 35]. In the following subsections we describe an overview of these generations.

2.3.1 *First generation acoustic synthesizers*

First generation synthesizers were the main synthesis techniques until 1980s and they are used less today. They were based on vocal tract models. The most famous synthesizers from the first generation are the formant synthesizer and articulatory synthesizer.

2.3.1.1 *Formant synthesizer*

Formant synthesizers make use of the source-filter model shown in Figure 2-3, where a source signal is fed into the vocal tract model to generate the sound. The vocal tract is considered as a filter with variable resonance frequencies (formants), formants vary with time according to the pronounced phoneme, and the source signal is either a sequence of periodic impulse train of one pitch period for voiced sounds, a white noise for obstruent sounds, or both for voiced obstruents.

It has been shown that only the first three formants can be used by the listeners to discriminate sounds, and so the higher formants may be used to add naturalness to the output speech. The locus theory indicates that formant frequencies within a phoneme tend to reach a stationary value named the *target.* The formant values in the transition regions between different phonemes are interpolated. Rule-based methods were used to find the value of these targets for different phonemes. Data-driven methods to generate the formant values have also been proposed [36].

A single formant can be implemented using a second-order IIR filter with the following transfer function:

$$H_i(z) = \frac{1}{1 - 2e^{-\pi b_i}\cos(2\pi f_i)\, z^{-1} + e^{-2\pi b_i} z^{-2}} \tag{2.1}$$

$$f_i = F_i / F_s \tag{2.2}$$

$$b_i = B_i / F_s \tag{2.3}$$

Where F_i, B_i, and F_s are the formant center frequency, formant bandwidth, and sampling frequency, respectively.

A filter with several resonances can be obtained by connecting several second-order sections in series (*cascade synthesizer*) or by adding them (*parallel synthesizer*). Usually the first three formants were used for speech production

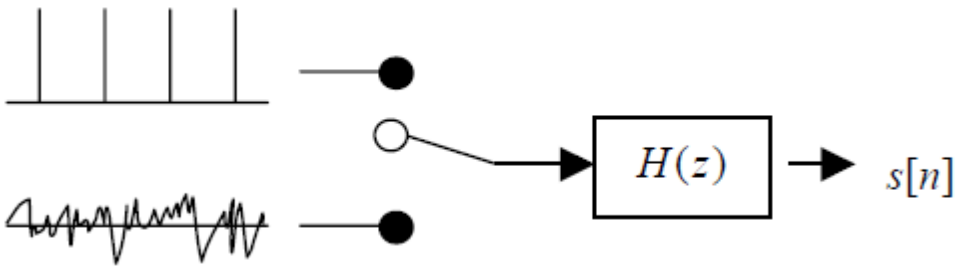


Figure 2-3 Source-filter model for voiced and unvoiced speech [33].

The output speech of formant synthesizer is intelligible but far from natural because of the simplicity of the model and the excitation signal.

2.3.1.2 *Articulatory synthesizer*

Articulatory synthesizer is another speech synthesizer which belongs to the first generation. The output speech is created by simulating the air flow through the vocal tract and the mechanical motions of human articulators which are responsible for speech production as shown in Figure 2-4.

Two difficulties in articulatory synthesis makes the quality of their output speech not comparable to that of the formant synthesis:

1- Finding the right model which simulates the human articulators accurately and at the same time is practical and easy to design and control.

2- Finding the correct model parameters, since they can't be estimated directly from the speech signal, rather they are estimated using X-rays and magnetic resonance imaging (MRI) by placing sensors in the vocal tract which change the way by which speech is generated.

2.3.2 *Second generation acoustic synthesizers (diphone-concatenative synthesizers)*

Over time, memory has become cheaper such that it becomes possible to store speech waveforms instead of recording parameters to produce speech. In the diphone-concatenative synthesizer, speech units (diphones) are recorded, and these units are concatenated to produce any new word sequence. Different signal processing algorithms (like PSOLA) are used to modify the concatenated speech to eliminate the concatenation distortion due to

discontinuities between units, and to modify the prosodic parameters (pitch- duration- intensity) of the output speech.

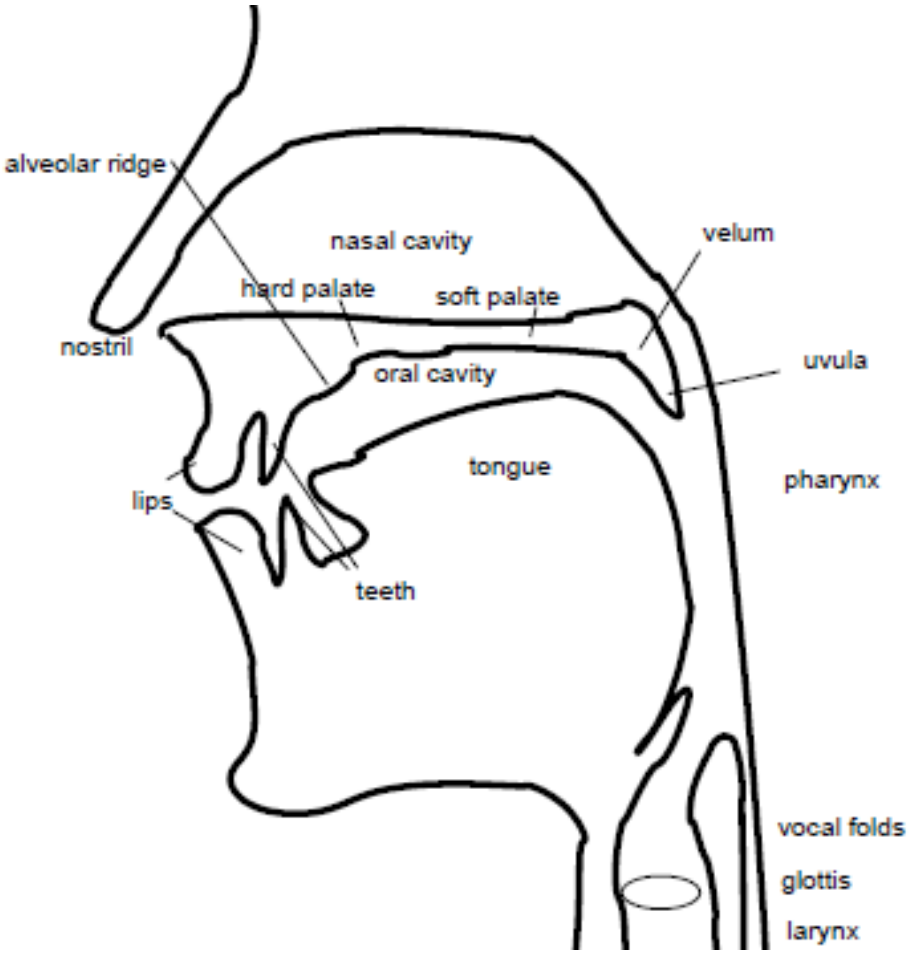


Figure 2-4 vocal organs [35].

2.3.3 *Third generation of acoustic synthesizers*

These are the most used synthesizers today because of their high accuracy. Third generation of acoustic synthesizers is data-driven which means that a very large database is used for these synthesizers. Two types of synthesizers exist in the third generation: HIDDEN MARKOV MODEL (HMM) SYNTHESIZERS, and UNIT SELECTION SYNTHESIZERS.

The main advantage of HMM synthesizers over the unit selection synthesizers is the high efficient use of memory, since they use orders of magnitude less memory to store the HMM parameters than storing the original speech waveforms as in the case of unit selection synthesizers.

2.3.3.1 *HIDDEN MARKOV MODEL SYNTHESIZER*

In the HMM-based speech synthesis [37], also called statistical parametric synthesis, speech waveforms are generated using trained context-dependent HMMs which models different speech parameters such as spectrum, excitation, and phoneme duration.

A well-known example of HMM-based speech synthesizer is HMM-based Speech Synthesis System (HTS) [38]. The system as shown in Figure 2-5 consists of two stages: training stage and synthesis stages.

Heiga Zen, and et al, illustrated the stages of HTS in [38] as follows:

"*The training part is similar to that used in speech recognition systems. The main difference is that both spectrum (mel-cepstral coefficients, and their dynamic features) and excitation (logarithmic fundamental frequencies ($log\, F_0$) and its dynamic features) parameters are extracted from a speech database and modeled by context-dependent HMMs (phonetic, linguistic, and prosodic contexts are taken into account). To model variable dimensional parameter sequence such as ($log\, F_0$) with unvoiced regions properly, multi-space probability distributions (MSD) are used. Each HMM has state duration probability density functions (PDFs) to capture the temporal structure of speech. As a result, the system models spectrum, excitation, and durations in a unified HMM framework. The synthesis part does the inverse operation of speech recognition. First, an arbitrarily given text to be synthesized is converted to a context-dependent label sequence, and then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Second, state durations of the utterance HMM are determined based on the state duration PDFs. Third, the speech parameter generation algorithm generates the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the corresponding speech synthesis filter (mel-log spectrum approximation (MLSA) filter for mel-cepstral coefficients).*"

If any modification is desired in this system, like emotion conversion or voice conversion, HMM parameters are modified using various techniques such as adaptation, interpolation, eigen voice, or multiple regression to achieve the desired modification.

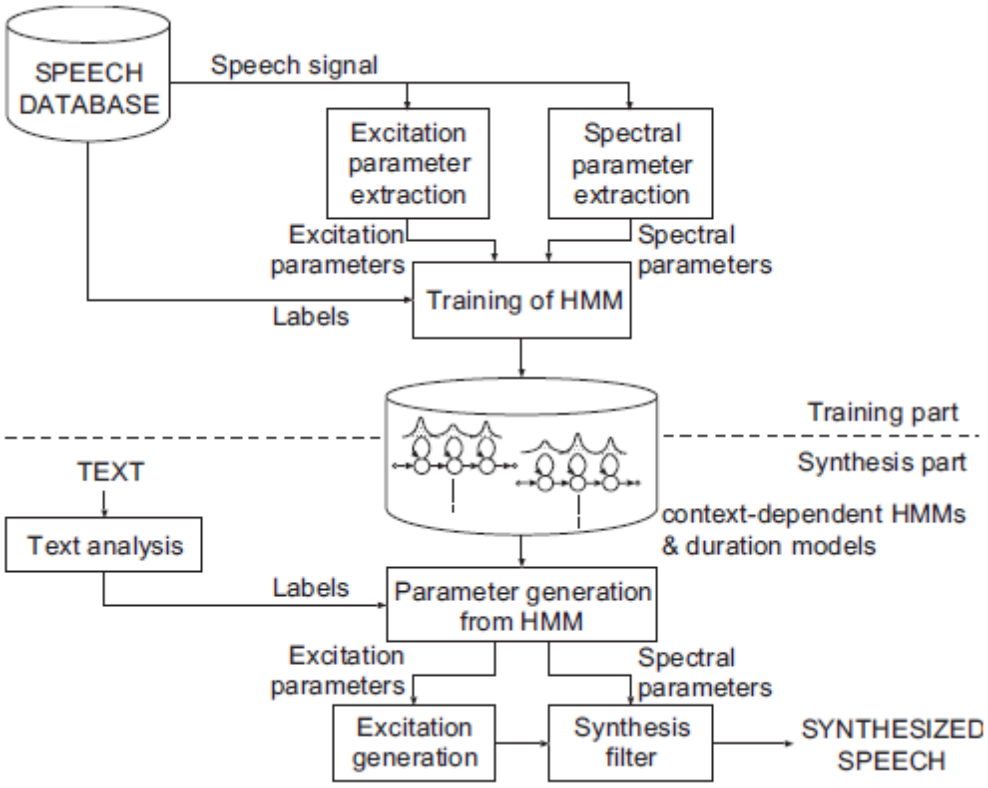


Figure 2-5 overview of HTS system [38].

2.3.3.2 *UNIT SELECTION SYNTHESIS*

Unit selection is the highest quality synthesis technique till now [33, 35, 39]. The main disadvantage of unit selection synthesis is the inefficient use of memory, since they require large memory to store a very large number of units which are used later to generate the synthesized speech.

Unit selection synthesis is an extension of diphone-concatenative synthesis, where in unit selection synthesis, an algorithm selects the best sequence of units to be concatenated to generate the desired utterance.

Different types of units with different lengths are used, including phonemes, diphones, syllables, words, and phrases. These units have different linguistic and acoustic (prosodic) features. Unit selection concatenation process can be seen as Huang, and et al said [33]:

"Unfortunately, this is equivalent to assembling an automobile with parts of different colors: each part is very good yet there is a color discontinuity from part to part that makes the whole automobile unacceptable. Speech segments are greatly affected by coarticulation, so if we concatenate two speech segments that were not adjacent to each other, there can be spectral or prosodic discontinuities. Spectral discontinuities occur when the formants at the concatenation point do not match. Prosodic discontinuities occur when the pitch at the concatenation point does not match."

A simple way to minimize the concatenation discontinuities is to use long units such as words, phrases or sentences if available, to obtain less concatenations; but when these long units aren't available, we will have to concatenate small units to obtain the desired utterance; which causes concatenation distortion and requires modification by some signal processing techniques.

To find the best sequence of units to be concatenated, a veterbi search algorithm is used to find the set of units with the lowest total cost. The total cost is the sum of the two following cost functions:

- *Target cost*: measures the distance between the desired specifications and the candidate units. These specifications are based on a set of linguistic features resulting from the text and prosodic analysis of the TTS, such as: target word, phone sequence, phone duration, pitch, and intonation.
- *Concatenation cost*: measures the discontinuity of concatenating different units based on the acoustic features.

The target cost between the desired target unit t_i and the candidate unit u_i, is a weighted sum of the distances (sub-costs) between the target and candidate feature vectors. If we have p target features, then the target cost will be:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} \omega_j^t C_j^t(t_i, u_i) \tag{2.4}$$

The concatenation cost between two adjacent units u_i, u_{i-1}, is a weighted sum of q concatenation sub-costs between these units:

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} \omega_k^c C_k^c(u_{i-1}, u_i) \tag{2.5}$$

The total cost between the target unit t_i, and the candidate unit u_i is the sum of the target and concatenation costs, target and concatenation costs are illustrated in Figure 2-6:

$$C(u_i, t_i) = C^t(u_i, t_i) + C^c(u_{i-1}, u_i) \tag{2.6}$$

The best sequence of units for a given utterance is the sequence which minimizes the total cost of the whole utterance:

$$\breve{u} = \underset{u}{\operatorname{argmin}}\{C(u_{1:n}, t_{1:n})\} \tag{2.7}$$

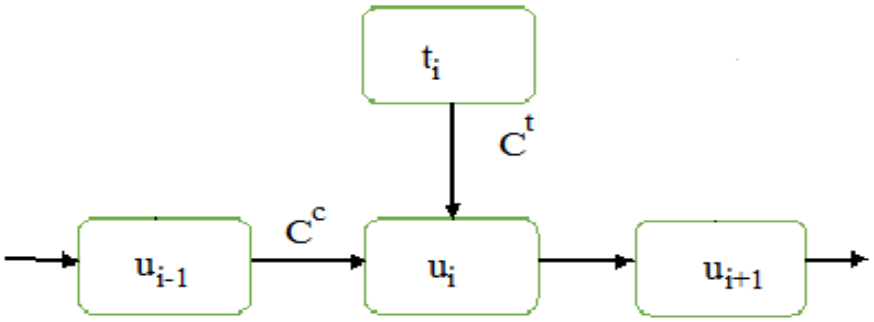


Figure 2-6 target and concatenation costs.

2.4 Applications of Text To Speech systems

Text To Speech systems have many applications in our life, such as: Aid-to-the handicapped, Education, interactive voice response (IVR) systems, entertainment applications, and speech-to-speech translation [33, 40].

- *Aid-to-the handicapped*: The most important application of TTS is the aid to blind people in various areas such as: reading books, easy use of different devices like computers and talking calculators.
- *Education:* Text to speech systems can be used in learning different languages.
- *Entertainment applications:* like reading e-mail while driving, or other applications which belong to car navigation systems.
- *Interactive voice response (IVR) systems*: IVR is a system that allows interaction between a computer and humans. The input to this system is entered by a keypad or by a speech recognition module and the output speech is produced either from pre-recorded messages or from a text to speech system. An important use of IVR systems is in the field of customer service in companies, hospitals, and etc.; which reduces the cost and increase the service efficiency.
- *Speech-to-speech translation (S2S)*: speech-to-speech translation between different languages enables cross-lingual communication. S2S requires speech recognition , machine translation and text-to-speech synthesis [41].

Chapter3: Expressiveness in Text To Speech systems

Generating an expressive speech from a TTS system is required to increase the naturalness and intelligibility of the output of these systems. Usually, a number of emotions or expressions are considered as fundamental emotions needed for TTS and the other emotions can be derived from them. Fundamental emotions vary in different studies, but several studies deal with anger, fear, surprise, happiness, sadness, and disgust as the fundamental emotions. In this chapter, we illustrate the acoustic correlates of emotions and various approaches to obtain expressive speech from TTS.

3.1 Acoustic correlates of emotions

Emotions may be defined as: mental states with identifiable effects on physiology and behavior. Different studies in the literature were done to find the acoustic correlates of emotion. Janet E. Cahn described the physiological and acoustical effects of emotion in [21] as follows:

"When emotion affects physiology the corresponding effects on speech show up primarily in the fundamental frequency (F0) and timing. Thus, with the arousal of the sympathetic nervous system – as with fear, anger or joy – heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is correspondingly loud, fast and enunciated, with strong high frequency energy. With the arousal of the parasympathetic nervous system – as with boredom or sadness – heart rate and blood pressure decrease and salivation increases, producing speech that is slow, low-pitched and with little high frequency energy".

Studies on the speech parameters, which are responsible for emotion, show that both prosodic parameters (including pitch, duration, and intensity), and spectral parameters of the speech signal are responsible for emotion in speech [2-5].

3.2 Approaches for expressive TTS

Different approaches for expressive TTS were proposed in literature including expressive unit selection for concatenative synthesizers, expressive HMM-based for HMM synthesizers, and both rule-based and data-driven emotion conversion for any type of speech synthesizer; since they operate on the output speech waveform itself regardless of the components of the text to speech system.

3.2.1 *Expressive unit selection*

One approach to obtain expressiveness in unit selection synthesis is to record either the complete set or just a small set of the original neutral database in different speaking styles.

In case of limited domain speech synthesis, in which the neutral database size is small to cover a specific area of speech, the complete database is recorded in different emotions and at the synthesis time the units from the desired emotion are concatenated to obtain the output utterance [10, 42]. In [42] an example of expressive unit selection in a limited domain was proposed for the application of computer games, where about 200 sentences from the poker domain in addition to 400 sentences which cover the most important diphones in German were recorded in different emotions. Different expressive styles can be obtained by switching from one emotional database to another.

For large neutral databases, usually just a small set of the neutral database is recorded in different emotions and combined with the original database and at the synthesis time the selection of the best candidate unit would be according to a cost function. As in the basic unit selection synthesis; the cost function consists of a target cost and a concatenation cost, the new thing in expressive unit selection is that the target cost includes a style cost with an appropriate weight. The style cost is either symbolic or acoustics-based.

Symbolic style cost is determined according to the distance between the desired emotion and the emotion of the candidate unit in the database which is represented by a label attached to it. The style label can be attached manually to the emotional corpus [12], or automatically to the overall corpus using an automatic labelling scheme trained by the small emotional corpus [43, 44]. The reason for labelling the complete database using this model is that some of the neutral units may sound emotional and this is used to increase the size of the emotional corpus.

Acoustics-based style cost is determined using the acoustic distance between the target and the candidate units. This distance is calculated according to pitch and voice quality parameters [12].

3.2.2 *Expressive HMM-based synthesis*

In the basic HMM-based synthesizer, speech parameters (spectrum, excitation, and phoneme duration) are modelled using context-dependent HMMs which are used later to produce the synthesized speech. The same for expressive HMM-based synthesizer, either each emotion is modeled individually using separate HMMs, or all emotions are modelled by a single model and the emotion of each utterance is treated as a contextual feature with other linguistic features [13]. A comparison between these two methods in [13] shows that both methods have the same performance. However, using a single model for all styles enables interpolation between different emotions to obtain new emotions which weren't found in the training database as proposed in [14].

3.2.3 *Rule-based emotion conversion*

Rule-based emotion conversion is used to obtain expressive speech, where a set of rules is designed to convert the effective acoustic parameters of speech to their emotional values. Different speech parameters affect the emotion in the speech signal; they can be divided into 4 categories: pitch, timing, voice quality and articulation parameters.

The pitch parameters describe the features of fundamental frequency, while the timing parameters are used to control different durations in the speech utterance, and the voice

quality represent the parameters of the speech spectrum. Precision of articulation is also used to convert the emotion.

Table 3-1 shows some speech parameters, which affect the emotion in the speech signal. Different sets of these parameters are used in [17-21] for emotion conversion.

Rules for transforming different speech parameters are set based on: analysis of parallel neutral-emotional corpus to find the variation of the parameters between different emotions[15-17], systematic parameter change to obtain the optimal variation of each parameter for each emotion[18], or using rules of prior studies in the literature [19-21].

A rule-based system for emotion conversion is presented in [21], each speech parameter has a known minimum and maximum values and varying this parameter between its min and max affects the output emotion. Modification of different parameters for different emotions is shown in Table 3-2. Parameter modification is performed on a scale centered at zero and varies from negative ten to positive ten. Zero represents the parameter value of neutral affect, while negative ten and positive ten represent the minimum and maximum values of this parameter, respectively. Output samples of this system can be found on the following link: http://alumni.media.mit.edu/~cahn/emot-speech.html.

The main problem of rule-based emotion conversion system is the tedious manual analysis required to obtain good rules, so a limited set of parameter variations can be captured [23].

Table 3-1 different acoustic parameters which are responsible for emotion in speech signal

Parameter category	parameter	Parameter meaning
Pitch parameters	Accent shape	The rate of F0 change for any pitch accent in the utterance.
	average pitch	The average F0 for the utterance.
	contour slope	The overall trend of the pitch range for the utterance − whether it expands, remains level or contracts.
	final lowering	The rate and direction of F0 change at the end of an utterance, whether it rises or falls.
	pitch range	The bandwidth of the range bounded by the lowest and highest F0 for the utterance.
	reference line	The F0 to which the pitch contour appears to return following a high or low pitch excursion.
Timing parameters	Exaggeration	The degree to which pitch accented words receive exaggerated duration as a means of emphasis.
	fluent pauses	The frequency of pausing between syntactic or semantic units.
	hesitation pauses	The frequency of pausing within a syntactic or semantic unit. These pauses often occur after the first function word in a clause.
	speech rate	The rate of speech. It affects the number of syllables or words spoken per minute and the duration of pauses.

	Stress frequency	The ratio of stressed to stressable words in an utterance. Stressable words may legitimately receive a pitch accent in accord with sentence semantics. Stressed words are those stressable words which actually do receive pitch accents.
Voice quality	phonation type	The phonation type can either be a modal, falsetto, breathy, creaky, or tense voice
	Breathiness	The amount of frication noise that may be co-present with non-fricative phonemes.
	Brilliance	The ratio of low to high frequency energy.
	Loudness	The amplitude of the speech signal.
	Spectral tilt	The degree to which intensity drops off as frequency increases. The overall slope of the spectrum is an indication of spectral tilt.
	Laryngealization	Describes the creaky voice phenomena in which there is minimal subglottal pressure, a small open quotient, a narrow glottal pulse and an irregular fundamental period. The speech of older speakers is often laryngealized.
	pause discontinuity	Describes the smoothness or abruptness of a pause onset.
	pitch discontinuity	Describes the smoothness or abruptness of F0 transitions throughout the utterance.
	Tremor or vocal jitter	Refers to irregularities between successive glottal pulses. It was observed in recordings of fearful utterances.
	lip-spreading	This feature affects mainly the happiness emotion and is implemented by raising the frequencies of the first two formants by a given rate.
Articulation parameter	Precision of articulation	Describes the degree of slurring or enunciation for all phoneme classes.

Table 3-2 example of speech parameter modification for different emotions [21]

Parameter	Angry	Disgusted	Glad	Sad	Scared	Surprised
Accent shape	10	0	10	6	10	5
average pitch	-5	0	-3	0	10	0
contour slope	0	0	5	0	10	10
final lowering	10	0	-4	-5	-10	0
pitch range	10	3	10	-5	10	8
reference line	-3	0	-8	-1	10	-8
fluent pauses	-5	0	-5	5	-10	-5
hesitation pauses	-7	-10	-8	10	10	-10
speech rate	8	-3	2	-10	10	4
Stress frequency	0	0	5	1	10	0
Breathiness	-5	0	-5	10	0	0
Brilliance	10	5	-2	-9	10	-3
Laryngealization	0	0	0	0	-10	0
Loudness	10	0	0	-5	10	5
pause discontinuity	10	0	-10	-10	10	-10
pitch discontinuity	3	10	-10	10	10	5
Precision of articulation	5	7	-3	-5	0	0

3.2.4 *Data-driven emotion conversion*

The basic idea of data-driven emotion conversion, also referred as statistical emotion conversion, is to train a statistical model or build a unit selection framework using a sufficient size of neutral-emotional corpus to be used later for converting speech parameters to the desired emotion. As in rule-based emotion conversion, speech parameters which are responsible for emotion are prosodic parameters (pitch- duration- energy) and spectral parameters. Data-driven emotion conversion requires the use of suitable machine learning techniques to map the speech parameters to their emotional values and appropriate signal processing algorithms to modify the speech parameters in the waveform without affecting the quality.

LMM, GMM, CART, HMM, and codebook mapping are well known machine learning techniques which are used for emotion conversion [22-26, 32], while we can find that PSOLA [28], LPC [23] STRAIGHT [30, 31] or HSM (Harmonic plus Stochastic Model) [32] are familiar signal processing algorithms which are used for analysis-synthesis and modification of the speech signal for emotion conversion. In the remaining sections of this chapter, we explain different signal processing and machine learning techniques which are used in our system.

3.3 Signal processing and machine learning algorithms for data-driven emotion conversion system

The implementation of data-driven emotion conversion involves three topics: finding the effective speech parameters which are responsible for each emotion in the speech signal, mapping each of these parameters to their emotional values using an appropriate machine learning technique, and modifying these parameters in the speech waveform using a suitable signal processing algorithm. In this section, we explain different signal processing and machine learning techniques which are used in our system, while finding the speech parameters which are responsible for emotion is covered in the next chapter.

3.3.1 *Linear Predictive Coding*

Linear Predictive Coding (LPC) is one of the most powerful analysis-synthesis techniques of the speech signal [45]. In LPC, the current time-domain sample $s[n]$ is predicted as a linear combination of the past time-domain samples $s[n-1], s[n-2], \ldots, s[n-M]$.

$$s[n] \approx \hat{s}[n] = -\sum_{i=1}^{M} a_i s[n-i] \tag{3.1}$$

where $\hat{s}[n]$ is the predicted sample, and a_i, $i = 1,2,\ldots,M$ are the LPC coefficients. The prediction error can be written as:

$$e[n] = s[n] - \hat{s}[n] = s[n] + \sum_{i=1}^{M} a_i s[n-i] = \sum_{i=1}^{M} a_i s[n-i] \tag{3.2}$$

where $a_0 = 1$. By taking the z-transform of Equation (3.2), we will obtain:

$$E(z) = S(z) + \sum_{i=1}^{M} a_i S(z)\, z^{-i} = S(z)\left[1 + \sum_{i=1}^{M} a_i\, z^{-i}\right] = S(z)A(z) \tag{3.3}$$

where

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i} = \sum_{i=0}^{M} a_i z^{-i} \tag{3.4}$$

Equation (3.3) can be written as follows:

$$S(z) = E(z)\frac{1}{A(z)} \tag{3.5}$$

From the above equation, the speech signal can be considered as the output of an all-pole filter whose transfer function is $H(z) = \frac{1}{\mathrm{A(z)}}$ and excited by the error signal E[z]. Thus the excitation signal can be obtained by inverse filtering the speech signal:

$$\mathrm{E(z)} = \frac{\mathrm{G} * \mathrm{S(z)}}{\mathrm{H(z)}} \tag{3.6}$$

where, G is the filter gain.

3.3.1.1 *Spectral estimation via LPC*

LPC is mainly used for spectral estimation of speech, so we illustrate the relation between LPC spectrum and the speech spectrum. The total prediction error using Parseval's theorem is:

$$E_p = \sum_{\mathrm{n}=-\infty}^{\infty} \mathrm{e}^2(\mathrm{n}) = \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \left|\mathrm{E}\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|^2 \mathrm{d}\omega \tag{3.7}$$

Substituting Equation (3.6) into Equation (3.7) we obtain,

$$E_p = \frac{\mathrm{G}^2}{2\pi} \int_{\omega=-\pi}^{\pi} \frac{\left|\mathrm{S}\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|^2}{|\mathrm{H}(\mathrm{e}^{\mathrm{j}\omega})|^2} \mathrm{d}\omega \tag{3.8}$$

The main idea of LPC parameter calculation is to find the coefficients that minimize the prediction error $\mathrm{E_p}$. From Equation (3.8), this minimization can be seen as minimizing the average ratio of the speech spectrum to its LPC spectrum. Thus the LPC spectrum is used to model the speech spectrum; the higher the LPC order is, the more spectral details that can be modeled by LPC.

3.3.1.2 *Line Spectral Frequencies*

An equivalent representation of LPC is the line spectral frequencies (LSFs) [46]. LSFs are more useful than LPC in many applications because of their good interpolation and quantization properties. LSFs are considered as the roots of the following two polynomials:

$$\begin{aligned} P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \\ = 1 - \left(a_1 - a_p\right)z^{-1} - \left(a_2 - a_{p-1}\right)z^{-2} - \cdots \\ - \left(a_p - a_1\right)z^{-p} + z^{-p+1} \end{aligned} \tag{3.9}$$

$$\begin{aligned} Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \\ = 1 - \left(a_1 + a_p\right)z^{-1} - \left(a_2 + a_{p-1}\right)z^{-2} - \cdots \\ - \left(a_p + a_1\right)z^{-p} + z^{-p+1} \end{aligned} \tag{3.10}$$

where $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_p z^{-p}$ is the inverse linear predictor filter. Thus it is clear that

$$A(z) = [P(z) + Q(z)]/2 \tag{3.11}$$

The equations (3.9), (3.10) can be written as

$$P(z) = \prod_{i=1,3,5,\dots} (1 - 2cos\omega_i z^{-1} + z^{-2}) \tag{3.12}$$

$$Q(z) = \prod_{i=2,4,6,\dots} (1 - 2cos\omega_i z^{-1} + z^{-2}) \tag{3.13}$$

where ω_i are the LSF parameters. Using equations (3.9-3.13) the LPC parameters $[a_1, a_2, \dots, a_p]$ can be converted to their LSF representation $[\omega_1, \omega_2, \dots, \omega_p]$.

3.3.2 *TD-PSOLA*

Time-Domain Pitch-Synchronous Overlap and Add method (TD-PSOLA) is the most popular algorithm for time-scale and pitch-scale modifications [35, 47, 48].

3.3.2.1 *TD-PSOLA analysis-synthesis without modification*

The analysis process of TD-PSOLA is to divide the speech signal into pitch-synchronous overlapped frames (short-term signals, $x_m(n)$). This is achieved by multiplying the speech waveform by overlapping windows (generally Hanning window, $h_m(n)$) centered at a predetermined pitch-marks, t_m, and usually extends to the next and the previous pitch-marks

$$x_m(n) = h_m(t_m - n)x(n) \tag{3.14}$$

Pitch-marks are set at a pitch-synchronous rate for voiced segments (one pitch-mark for every pitch period) and at a constant rate for unvoiced segments. Glottal Closure Instants are often used as pitch-marks in PSOLA.

To synthesize these frames; their centers are placed back on the original pitch-marks and the overlapping regions are added together, this process is known as overlap and add synthesis. The output speech waveform of PSOLA analysis-synthesis is perceptually indistinguishable from the original waveform. The analysis-synthesis process using TD-PSOLA is illustrated in Figure 3-1.

3.3.2.2 *Time-scale & pitch-scale modification using TD-PSOLA*

The main idea of time-scale modification using TD-PSOLA is to eliminate or duplicate some frames to achieve the desired expansion or compression of the speech waveform without affecting the pitch contour, this operation is illustrated in Figure 3-2.

Pitch-scale modification is carried out by recombining the speech frames on modified pitch-marks which are set synchronous with the new pitch values as shown in Figure 3-3.

In both cases of modifications; new pitch-marks which achieves the desired modification are calculated, then the frames which correspond to those pitch-marks are determined and centered on those new pitch-marks. Finally, the overlapping regions are added together to obtain the modified speech waveform.

3.3.2.3 *Computation of synthesis pitch-marks for pitch modification*

Let t_a^i be the i^{th} analysis pitch-mark, t_s^i be the i^{th} synthesis pitch-mark, and $a(t)$ be the pitch-scale modification factor. The analysis pitch-marks are set at a pitch synchronous rate for voiced frames (one pitch-mark for each pitch period), then

$$t_a^{i+1} = t_a^i + P\left(t_a^i\right) \tag{3.15}$$

where $P\left(t_a^i\right)$ is the local pitch period in the interval $t_a^i \le t < t_a^{i+1}$. The same for the synthesis pitch-marks:

$$t_s^{i+1} = t_s^i + P'\left(t_s^i\right) \tag{3.16}$$

where $P'\left(t_s^i\right)$ is the modified pitch around time-instant t_s^i, and equals the mean value of the scaled original pitch period over the frame $t_s^i \to t_s^{i+1}$:

$$P'\left(t_s^i\right) = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)}\, dt \tag{3.17}$$

$$\alpha(t) = a\left(t_a^i\right) \qquad for\ t_a^i \le t < t_a^{i+1} \tag{3.18}$$

Substituting Equation (3.17) in Equation (3.16), we will obtain

$$t_s^{i+1} = t_s^i + \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)}\, dt \tag{3.19}$$

If we have the synthesis pitch-mark t_s^i, the next synthesis pitch-mark can be obtained using Equation (3.19). An example of computation of synthesis pitch-marks for pitch modification by 1.5 is shown in Figure 3-4. The virtual time axis here in pitch-scale modification is the same as the synthesis time axis.

3.3.2.4 *Computation of synthesis pitch-marks for time-scale modification*

Let $\beta(t)$ be the time-scale modification factor at time instant t, the mapping $t_a^i \to t_s^i = D(t)$ will be as follow:

$$D(t) = \int_0^t \beta(\tau) d\tau \tag{3.20}$$

where $D(t)$ is called time-scale warping function. As mentioned before, time-scale modification should be done without affecting the pitch contour, so a stream of virtual pitch-marks, t_v^i, in the original signal is used. The relation between virtual and synthesis pitch-marks is:

$$t_s^i = D\left(t_v^i\right) \tag{3.21}$$

$$t_v^i = D^{-1}\left(t_s^i\right) \tag{3.22}$$

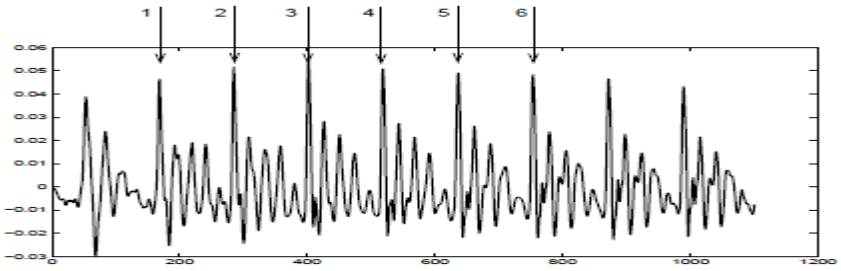
The synthesis pitch value over the interval $t_s^{i+1} - t_s^i$ equals the mean pitch value of the original signal over the interval $t_v^{i+1} - t_v^i$

$$t_s^{i+1} - t_s^i = \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} P(t)\, dt \tag{3.23}$$

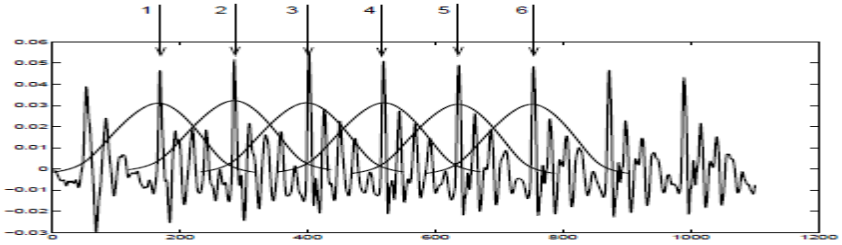
Thus, given the synthesis pitch-mark t_s^i, and the virtual pitch-mark t_v^i, the next synthesis pitch-mark, t_s^{i+1}, which preserves the original pitch value over the interval $t_v^{i+1} - t_v^i$, is obtained from the following relation:

$$t_s^{i+1} = t_s^i + \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} P(t)\, dt \tag{3.24}$$

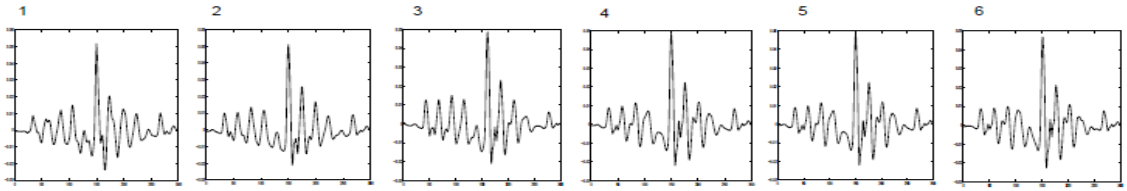
With $t_s^i = D\left(t_v^i\right)$. An example of computation of synthesis pitch-marks for time-scale modification by 1.5 is shown in Figure 3-5.



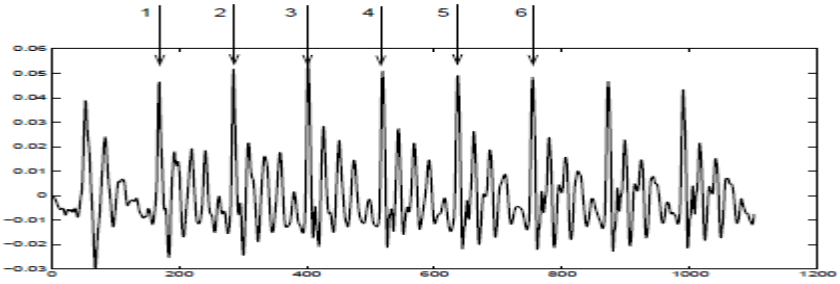
(a)



(b)



(c)



(d)

Figure 3-1 TD-PSOLA analysis-synthesis process without modification, (a) Original speech waveform with pitch-marks, (b)A Hanning window is centered on each pitch-mark, (c) Separate frames created by the Hanning window, and (e) synthesized speech waveform by overlap-add method [35].

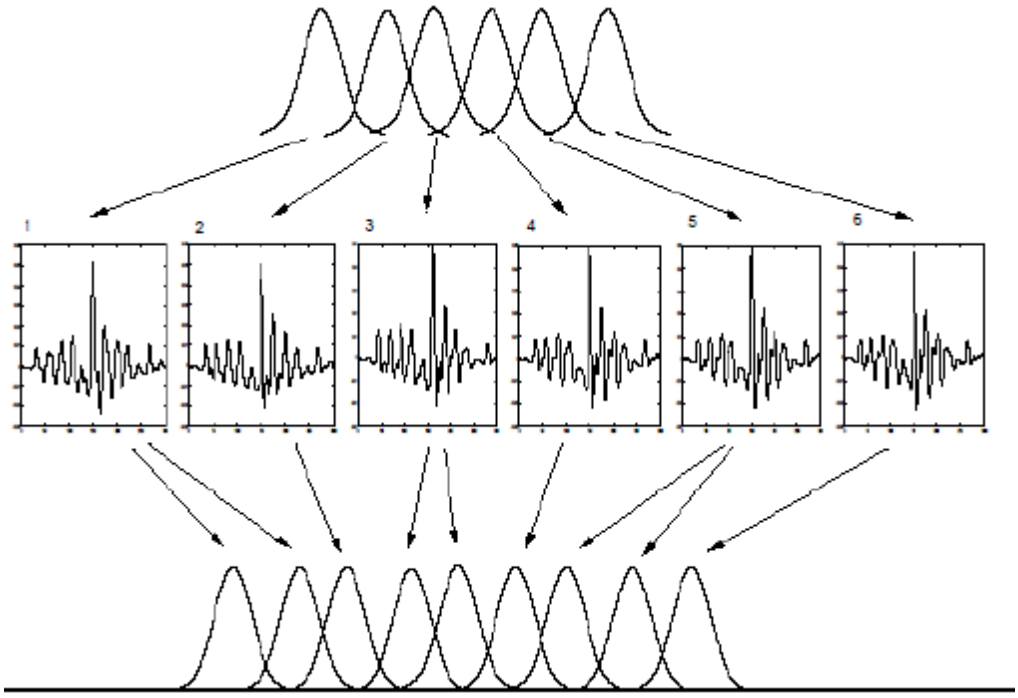


Figure 3-2 time-scaling (lengthening) using TD-PSOLA [35]

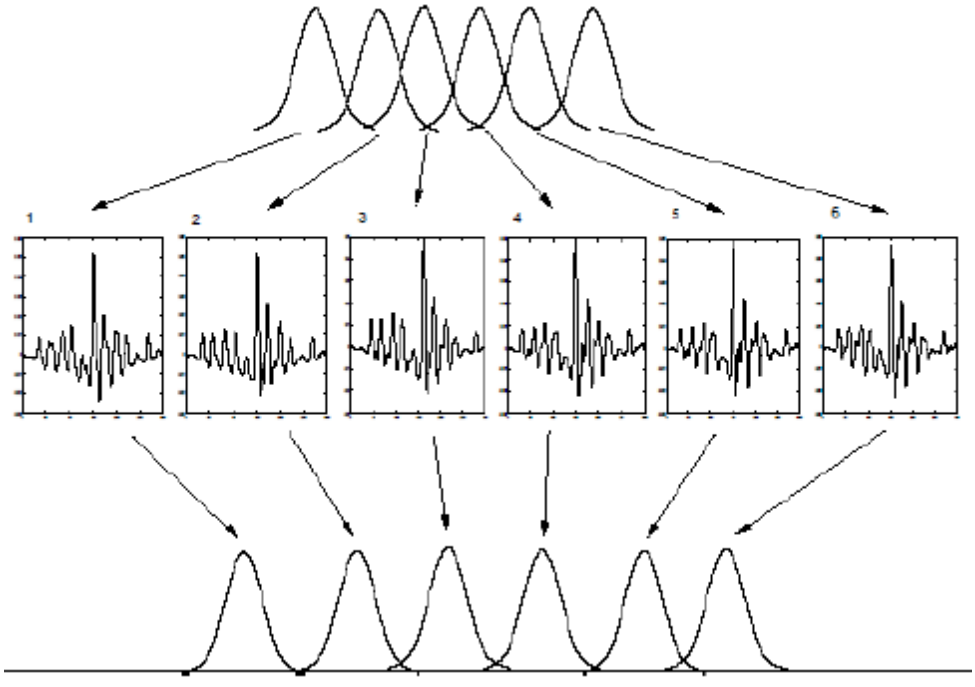


Figure 3-3 pitch scaling (lowering) using TD-PSOLA [35]

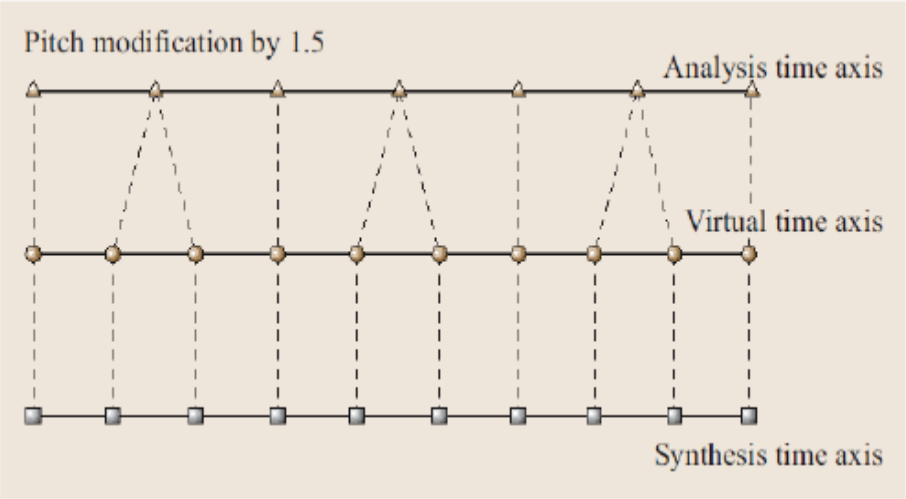


Figure 3-4 Computation of synthesis pitch-marks for pitch modification by 1.5 [47].

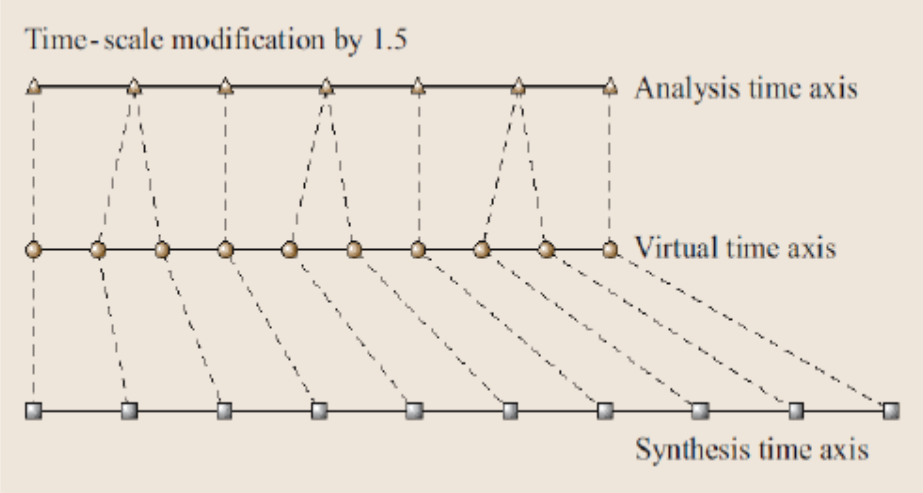


Figure 3-5 Computation of synthesis pitch-marks for time-scale modification by 1.5 [47]

3.3.2.5 *From the synthesis pitch-marks to the modified waveform*

After calculating the synthesis and virtual pitch-marks, the nearest analysis pitch-mark to the virtual pitch-mark is found (This process is illustrated in Figure 3-4, and Figure 3-5), then the frames which corresponds to the nearest analysis pitch-marks are centered on the synthesis pitch-marks. Finally, the overlapping regions are added together. An example of increasing the pitch in speech waveform is shown in Figure 3-6.

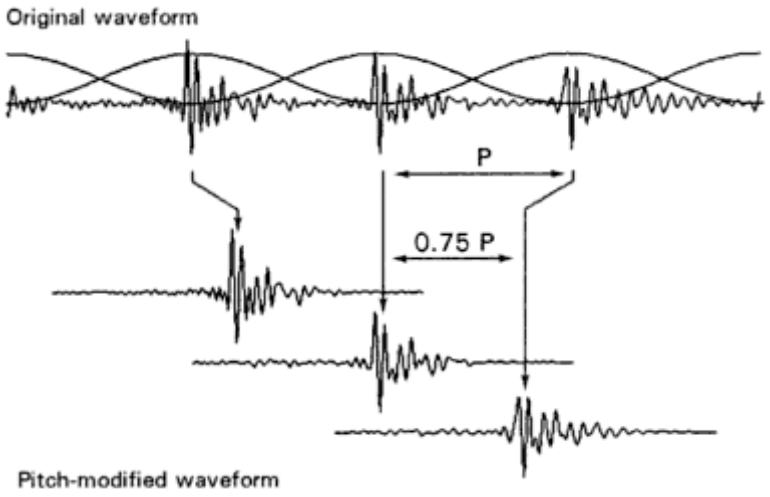


Figure 3-6 pitch raising in speech waveform using TD-PSOLA [47]

3.3.3 *Dynamic time warping*

Dynamic Time Warping (DTW) is a popular algorithm used in many applications such as speech recognition, data mining, and signature matching [49-51]. DTW can be used to find the best time-alignment between sequences of features of two signals by minimizing the distance between them. An example of alignment of two signals is shown in Figure 3-7. Given two time series X of length N, and Y of length M, where

$$X = x_1, x_2, \dots, x_i, \dots, x_N \tag{3.25}$$

$$Y = y_1, y_2, \dots, y_i, \dots, y_M \tag{3.26}$$

DTW can be used to align these sequence. First, an N-by-M matrix is constructed, the (i^{th}, j^{th}) element in this matrix equals the distance between the two points (x_i, y_j):

$$d(x_i, y_j) = \|x_i - y_j\| \tag{3.27}$$

Second, the cumulative cost matrix is constructed, where each element in this matrix, $C(i,j)$, represents the minimum cumulative cost of reaching this point:

$$C(i,j) = \begin{cases} \sum_{k=1}^{j} d(x_1, y_k) & , i = 1 \\ \sum_{k=1}^{j} d(x_k, y_1) & , j = 1 \\ d(x_i, y_j) + min\{C(i-1, j-1), C(i-1, j), C(i, j-1)\} & , other \end{cases} \tag{3.28}$$

The final step is to find the optimal warping (alignment) path W. The warping path is a set of matrix elements, $w_k = (i,j)_k$, which defines the mapping between the two sequences (X,Y). The optimal warping path contains a set of points which minimize the cumulative cost and achieve the following conditions:

- *Boundary conditions*: the warping path starts and finishes at start and end points of the two sequences $(X, and\ Y)$:

$$w_1 = (1,1), and\ w_K = (N,M)$$

- *Continuity*: the steps in the warping path is only to the adjacent cells:

$$w_k - w_{k-1} \in \{(1,1),(1,0),(0,1)\}$$

- *Monotonicity*: if $w_k = (a,b)$, and $w_{k-1} = (a',b')$, then,

$$a \geq a', \text{and } \ b \geq b'$$

An algorithm for calculating the optimal path is shown in algorithm 3.1, and Figure 3-8 shows an example of finding the optimal warping path.

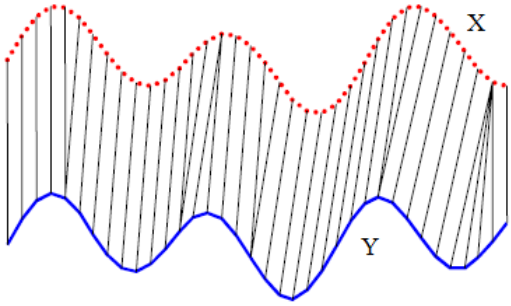


Figure 3-7 example of alignment of two signals [50]

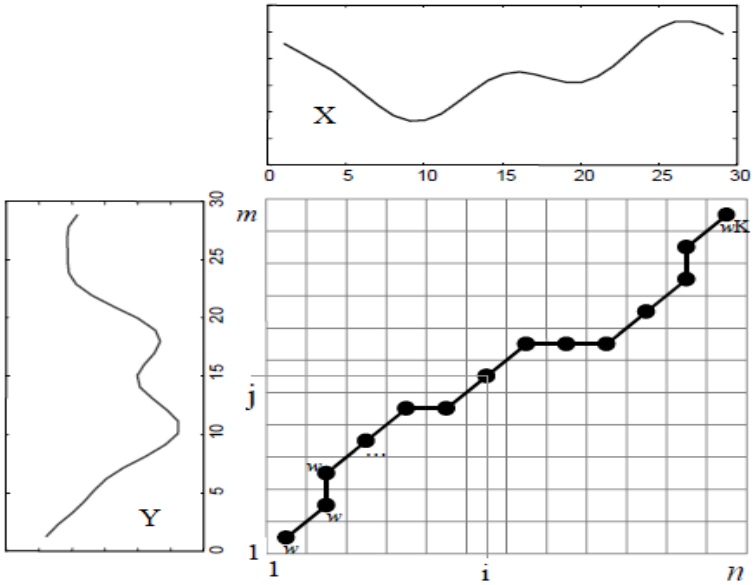


Figure 3-8 example of finding the optimal warping path [51]

Algorithm 3.1. OPTIMAL WARPING PATH (W) [49]

1: $path[] \leftarrow new\ array$

2: $i = N = rows(C)$

3: $j = M = columns(C)$

4: $\boldsymbol{while}(i > 1)\&(j > 1)\ \boldsymbol{do}$

5: $\boldsymbol{if}\ i == 1\ \boldsymbol{then}$

6: $j = j - 1$

7: $\boldsymbol{else\ if}\ j == 1\ \boldsymbol{then}$

8: $i = i - 1$

9: $\boldsymbol{else}$

10: $\boldsymbol{if}\ C(i-1,j) == min\{C(i-1,j); C(i,j-1); C(i-1,j-1)\}\ \boldsymbol{then}$

11: $i = i - 1$

12: $\boldsymbol{else\ if}\ C(i,j-1) == min\{C(i-1,j); C(i,j-1); C(i-1,j-1)\}\ \boldsymbol{then}$

13: $j = j - 1$

14: $\boldsymbol{else}$

15: $i = i - 1; j = j - 1$

16: $\boldsymbol{end\ if}$

17: $path.add((i,j))$

18: $\boldsymbol{end\ if}$

19: $\boldsymbol{end\ while}$

20: $\boldsymbol{return}\ path$

3.3.4 *Pitch detection*

Pitch detection of the speech signal is required for many applications such as speech enhancement, coding, and synthesis. Pitch detectors estimate the voiced or quasi-periodic frames of the speech signal (voiced/unvoiced (V/UV) estimation), and then calculate the pitch or fundamental frequency (F0) for voiced frames. Pitch detection can be done in either time-domain, frequency-domain, or both time-frequency-domain [52].

In time-domain pitch detection, the signal periodicity is computed from the temporal correlation between the signal samples. RAPT [53], PRAAT [54], and YIN [55] are well-known examples for pitch detectors which perform time-domain detection and give accurate pitch estimation for clean speech.

In frequency-domain pitch detection, the signal periodicity is computed from the short-time spectral representation of the speech signal by searching for strong harmonic peaks near integer multiples of a frequency value, this value would be the fundamental frequency, F0. SHR [56], and SWIPE' [57] are well-known examples for frequency domain pitch detectors.

Pitch detection in time-frequency-domain involves the decomposition of the speech signal into multiple frequency sub-bands, and then applying time-domain techniques on each sub-band. MBSC [52] is an example of time-frequency-domain detection with accurate estimation for both clean and noisy speech.

In the following subsections, we present the basic idea of operation of PRAAT, and MBSC pitch detectors which were used in our system.

3.3.4.1 *PRAAT pitch detector*

PRAAT pitch detector employs the autocorrelation method for pitch calculation [54]. The basic idea of the autocorrelation method is to estimate the pitch period of a periodic signal from the position of the maximum of the autocorrelation function of this signal. The autocorrelation of the signal x(t) is defined as:

$$r_x(\tau) = \int x(t)\, x(t+\tau) dt \tag{3.29}$$

$r_x(\tau)$ has a global maximum at $\tau = 0$, if there exist global maxima at $\tau = nT$, the signal $x(t)$ would be a periodic signal with period T. For quasi-periodic signals like the speech signals, these signals have periodic parts with different periods, if the autocorrelation of any segment of this signal has local maxima with enough height (autocorrelation values at these maxima are large enough), this segments would be periodic and its period is calculated from the location of these local maxima.

This simple calculation causes errors in pitch detection due to sampling and windowing of the speech segment, these errors were tackled in PRAAT.

In PRAAT, the acoustic signal is divided into frames of length equals three times of the minimum pitch period (the minimum value of F0, pitch estimation should be higher than this value), then each segment is multiplied by a Hanning window. Let $x(t)$ be an acoustic segment with duration T and centered around t_{mid}, and $w(t)$ is the window function, then the windowed speech segment is:

$$a(t) = \left(x\left(t_{mid} - \frac{1}{2}T + t\right) - \mu_x\right) w(t) \tag{3.30}$$

where μ_x is the segment mean, and $w(t)$ is symmetric around $t = \frac{1}{2}T$ and zero everywhere outside the time interval $[0, T]$. The normalized autocorrelation of the windowed segment is defined as:

$$r_a(\tau) = \frac{\int_0^{T-\tau} a(t)a(t+\tau)dt}{\int_0^T a^2(t)dt} \tag{3.31}$$

The autocorrelation of the speech segment is calculated as the ratio between the normalized autocorrelation of the windowed segment to the normalized autocorrelation of the window.

$$r_x(\tau) \approx \frac{r_a(\tau)}{r_w(\tau)} \tag{3.32}$$

The sampling and windowing problems of the simple autocorrelation pitch detection mentioned above were solved in PRAAT by using $\sin x/x$ interpolation of the sampled version of the autocorrelation function $r_x(\tau)$ and by the division by the autocorrelation of the window as in Equation (3.32), respectively. PRAAT algorithm is illustrated in [54].

3.3.4.2 *MBSC pitch detector*

Multi-band summary correlogram-based pitch detection [52] gives high accuracy for clean and noisy speech signals. MBSC employs several signal processing schemes in both time-domain and frequency-domain such as sub-band multi-channel comb-filtering, HSR - based channel-selection-and-weighting (HSR stands for Harmonic-to-subharmonic energy ratio), stream-reliability-weighting. The use of multiple signal processing algorithms, in both time-domain and frequency-domain, is to obtain high accuracy of pitch estimation, and V/UV detection.

3.3.5 *Gaussian Mixture Model (GMM)*

Gaussian Mixture Model [46] is used to model the probability density function of a random signal space. A mixture of a number of Gaussian probability density functions is used to fit the signal space. Figure 3-9, and Figure 3-10 show an example of modelling of two-dimensional random variable using a mixture of two bivariate Gaussian densities.

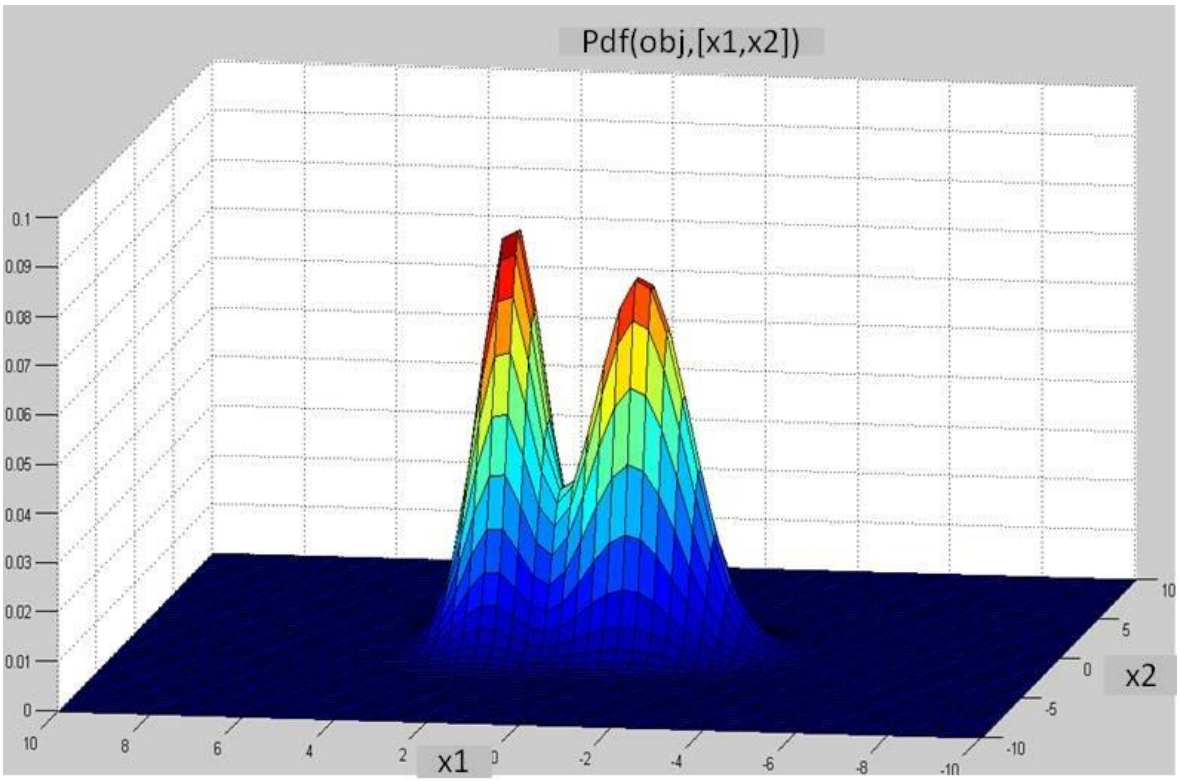


Figure 3-9 modelling of two-dimensional random variable using a mixture of two bivariate Gaussian densities.

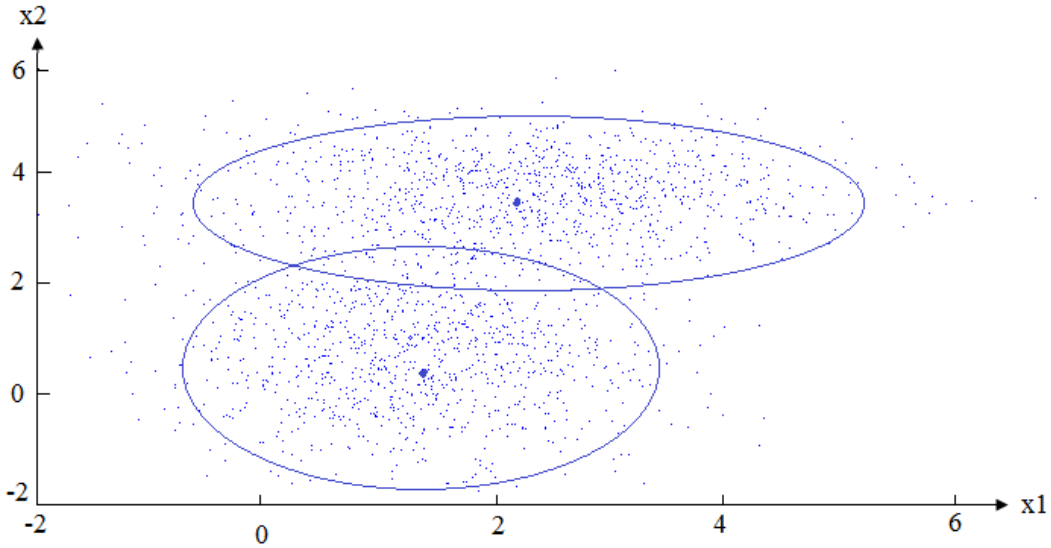


Figure 3-10 modelling of two-dimensional random variable using a mixture of two bivariate Gaussian densities (projection of the probability density functions on the variable plane).

A random process **X** can be fitted by a mixture of Gaussian components such that

$$f_x(x) = \sum_{k=1}^{K} P_k N_k(x; \mu_k, \Sigma_k) \tag{3.33}$$

where K is the number of Gaussian components used to fit the random process, $N_k(x; \mu_k, \Sigma_k)$ is the k^{th} component probability density function (pdf), with mean vector μ_k and covariance matrix Σ_k, and P_k is the prior probability of the k^{th} Gaussian component (the probability that the process **X** is associated with the k^{th} mixture).

A signal space can be fitted by an infinite number of different K-mixture Gaussian densities. Expectation Maximization (EM) can be used to estimate GMM parameters (P_k, μ_k, Σ_k).

3.3.5.1 *EM Estimation of Gaussian Mixture Model*

EM algorithm is an iterative maximum-likelihood (ML) estimation, and is used to estimate GMM parameters (P_k, μ_k, Σ_k) [46].

Given the observation vectors $[y(m), m = 0, \dots, N-1]$, we define the complete and incomplete datasets as follows:

- The incomplete data set: is the input observation vectors:

$$y(m), m = 0, \dots, N-1 \tag{3.34}$$

- The complete data set: is the observation vectors with a tag k attached to each vector to indicate the GMM component which generated the vector:

$$x(m) = [y(m), k] = y_k(m), m = 0, \dots, N-1, k \in (1, \dots, k) \tag{3.35}$$

  With this definition of the complete data set, if this data set is known, the calculation of the GMM parameters (P_k, μ_k, Σ_k) would be relatively simple.

Let $\Theta = \{\theta_k = [P_k, \mu_k, \Sigma_k], k = 1, \dots, K\}$ be the parameters of the GMM to be estimated, the EM algorithm starts with an initial estimate $\Theta_i = \{\theta_{k_i} = [P_{k_i}, \mu_{k_i}, \Sigma_{k_i}], k = 1, \dots, K\}$, then given the observations and a current estimate Θ_i, the expectation of the complete data would be as follows:

$$\begin{aligned} U(\Theta, \Theta_i) &= E\left[\ln f_{Y,K;\Theta}(y(m), k; \Theta) | y(m); \hat{\Theta}_i\right] \\ &= \sum_{m=0}^{N-1} \sum_{k=1}^{K} \frac{f_{Y,K|\Theta}\left(y(m), k|\hat{\Theta}_i\right)}{f_{Y|\Theta}\left(y(m)|\hat{\Theta}_i\right)} \ln f_{Y,K;\Theta}(y(m), k; \Theta) \end{aligned} \tag{3.36}$$

Now the joint pdf of $y(m)$ and the k^{th} GMM component is:

$$f_{Y,K|\Theta}\left(y(m), k|\hat{\Theta}_i\right) = P_{k_i} f_k\left(y(m)|\hat{\Theta}_{k_i}\right) = P_{k_i} N_k\left(y(m); \hat{\mu}_{k_i}, \hat{\Sigma}_{k_i}\right) \tag{3.37}$$

where $N_k\big(y(m);\hat{\mu}_k,\hat{\Sigma}_k\big)$ is a Gaussian density with mean vector μ_k and covariance matrix Σ_k:

$$N_k(y(m);\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(y(m)-\mu_k)^T\Sigma_k^{-1}(y(m)-\mu_k)\right) \quad (3.38)$$

The pdf of $y(m)$ as a mixture of k Gaussian densities is given by:

$$f_{Y|\Theta}\big(y(m)|\hat{\theta}_i\big) = N\big(y(m)|\hat{\theta}_i\big) = \sum_{k=1}^{K}\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big) \quad (3.39)$$

Substituting Equation (3.37) and Equation (3.39) in Equation (3.36) yields:

$$\begin{aligned} U\big[(\mu,\Sigma,P),\big(\hat{\mu}_i,\hat{\Sigma}_i,\hat{P}_i\big)\big] &= \sum_{m=0}^{N-1}\sum_{k=1}^{K}\frac{\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big)}{N\big(y(m)|\hat{\theta}_i\big)}\ln[P_kN_k(y(m);\mu_k,\Sigma_k)] \\ &= \sum_{m=0}^{N-1}\sum_{k=1}^{K}\left(\frac{\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big)}{N\big(y(m)|\hat{\theta}_i\big)}\ln P_k\right. \\ &\left. + \frac{\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big)}{N\big(y(m)|\hat{\theta}_i\big)}\ln N(y(m);\mu_k,\Sigma_k)\right) \end{aligned} \quad (3.40)$$

To find the next iteration (i+1), Equation (3.40) is maximized with respect to the GMM parameters $\{\theta_k = [P_k,\mu_k,\Sigma_k]\}$ each at a time, and given that $\sum P_k = 1$, yields:

$$\hat{P}_{k_{i+1}} = \arg\max_{P_k} U\big[(\mu,\Sigma,P),\big(\hat{\mu}_i,\hat{\Sigma}_i,\hat{P}_i\big)\big] = \frac{1}{N}\sum_{m=0}^{N-1}\frac{\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big)}{N\big(y(m)|\hat{\Theta}_i\big)} \quad (3.41)$$

,

$$\begin{aligned} \hat{\mu}_{k_{i+1}} &= \arg\max_{\hat{\mu}_k} U\big[(\mu,\Sigma,P),\big(\hat{\mu}_i,\hat{\Sigma}_i,\hat{P}_i\big)\big] \\ &= \frac{\sum_{m=0}^{N-1}\frac{\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big)}{N\big(y(m)|\hat{\Theta}_i\big)}y(m)}{\sum_{m=0}^{N-1}\frac{\hat{P}_{k_i}N_k\big(y(m);\hat{\mu}_{k_i},\hat{\Sigma}_{k_i}\big)}{N\big(y(m)|\hat{\Theta}_i\big)}} \end{aligned} \quad (3.42)$$

,and

$$\hat{\Sigma}_{k_{i+1}} = \arg\max_{\hat{\Sigma}_k} U\left[(\mu, \Sigma, P), \left(\hat{\mu}_i, \hat{\Sigma}_i, \hat{P}_i\right)\right]$$

$$= \frac{\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i} N_k\left(y(m); \hat{\mu}_{k_i}, \hat{\Sigma}_{k_i}\right)}{N\left(y(m)|\hat{\Theta}_i\right)} \left(y(m) - \hat{\mu}_{k_i}\right)\left(y(m) - \hat{\mu}_{k_i}\right)^T}{\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i} N_k\left(y(m); \hat{\mu}_{k_i}, \hat{\Sigma}_{k_i}\right)}{N\left(y(m)|\hat{\Theta}_i\right)}} \quad (3.43)$$

Equations (3.41)–(3.43) are the estimates of the GMM parameters at the $(i+1)^{th}$ iteration using the parameters' estimate at the i^{th} iteration. The iterative process continues until the parameters' estimate converge.

3.3.6 *CLASSIFICATION AND REGRESSION TREES (Decision trees)*

Classification and regression tree (CART) is an important machine learning technique used for a variety of applications [33]. To construct the CART from the training dataset ʒ; First a set of questions about the variables are found, then the tree is grown from the root node towards the leaf node using the best questions, finally the tree is pruned to a level which minimizes the misclassification rate of the new test data.

3.3.6.1 *Choice of Question Set*

Let *X* denotes the input sample to the CART

$$X = (x_1, x_2, \dots, x_N) \quad (3.44)$$

where the values of each variable x_i can be discrete in case of classification tree or continuous in case of regression tree. A standard set of questions Q for the CART will be as follows:

1. Each question in Q is simply about the value of one variable x_i.
2. For classification trees, if x_i has discrete values of the set $\{c_1, c_2, \dots, c_k\}$, then the question about its value will be in the form of:

   $$\{Is\ x_i \in S\} \quad (3.45)$$

   where *S* is any subset of $\{c_1, c_2, \dots, c_k\}$
3. For regression trees, the question about the value of the continuous variable x_i will be in the form of:

   $$\{Is\ x_i \leq c_n\} \qquad n = 1, 2, \dots, M \quad (3.46)$$

   $$c_n = \frac{v_{n-1} + v_n}{2}\ , v_0 = 0 \quad (3.47)$$

   where *M* is the number of training samples in ʒ, and v_n is the value of the variable x_i of the n^{th} sample.

3.3.6.2 *Splitting Criteria*

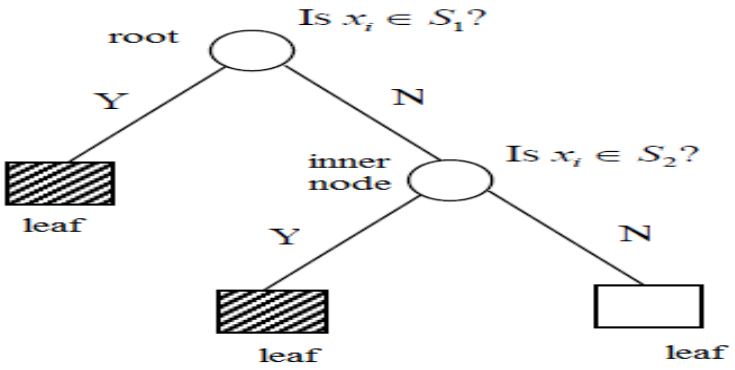


Figure 3-11 classification tree example

As shown in Figure 3-11, the tree grows from the root node towards the leaf nodes. Each node has a question which represents a split of the training samples. The task to find the splitting criteria is to find the best question set which split the training data at each node.

For an input sample X, let its classification decision output be the random variable Y. the weighted entropy for any node t can be defined as follows:

$$\bar{H}_t(Y) = H_t(Y)P(t) \tag{3.48}$$

$$H_t(Y) = -\sum_i P(\omega_i|t) \log P(\omega_i|t) \tag{3.49}$$

where $P(\omega_i|t)$ is the percentage of samples for class i in node t; and $P(t)$ is the prior probability of visiting node t.

Let the question q at node t splits the node t into leaves l and r, then the entropy reduction will be:

$$\Delta\bar{H}_t(q) = \bar{H}_t(Y) - \left(\bar{H}_l(Y)+\bar{H}_r(Y)\right) = \bar{H}_t(Y) - \bar{H}_t(Y|q) \tag{3.50}$$

Thus, the splitting criterion can be considered as finding the question which gives the greatest entropy reduction.

$$q^* = \underset{q}{\text{argmax}}\left(\Delta\bar{H}_t(q)\right) \tag{3.51}$$

The entropy for a tree, T, is defined as the sum of weighted entropies for all the terminal (leaf) nodes

$$\bar{H}(T) = \sum_{t\ is\ terminal} \bar{H}_t(Y) \tag{3.52}$$

It can be shown that the entropy of the tree is reduced during the tree-growing process.

3.3.6.3 *Growing the Tree*

The tree-growing algorithm begins from the root node and splits nodes using the best question set Q and the splitting criteria $\Delta\bar{H}_t(q)$. The algorithm continues for all nodes until reaching terminal nodes (leaf nodes) which satisfy one of the following conditions:

1. All the samples in the node are of the same class.

2. The maximum entropy reduction of best question is less than a pre-determined threshold β:

$$\max_{q\in Q} \Delta\bar{H}_t(q) < \beta \tag{3.53}$$

3. The number of samples in the node is less than a pre-determined threshold α which is known as (minimum leaf parameter).

3.3.6.4 *Tree Pruning*

CART may be designed precisely on the training data without generalization, so high classification error may occur for new test data. To minimize the classification error for new test data, a pruning approach is used. This approach merges leaves on the same tree branch until the misclassification rate for the new test data reaches its minimum. Using Minimum Cost-Complexity pruning algorithm [33], the weakest subtree (subtrees) for each pruning level k is determined. For $k = 1$, the weakest subtree T_{t_1} is determined using the minimum cost complexity algorithm. After pruning away T_{t_1} from T, the new pruned tree T_1 will be:

$$T_1 = T - T_{t_1} \tag{3.54}$$

Using the same process, the weakest subtree T_{t_2} in T_1 can be found. After pruning away T_{t_2} from T_1, the new pruned tree T_2 will be:

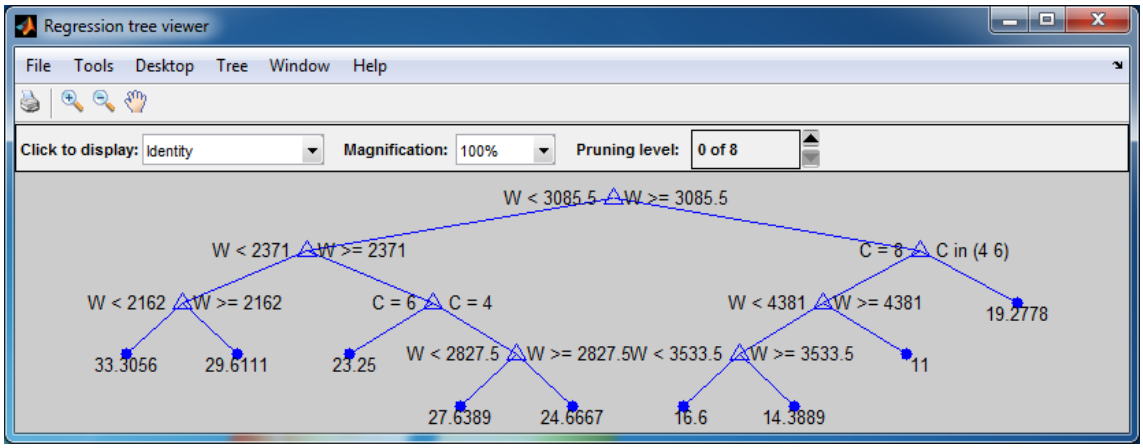
$$T_1 = T_1 - T_{t_2} = T - T_{t_1} - T_{t_2} \tag{3.56}$$

$$k = 2 \tag{3.57}$$

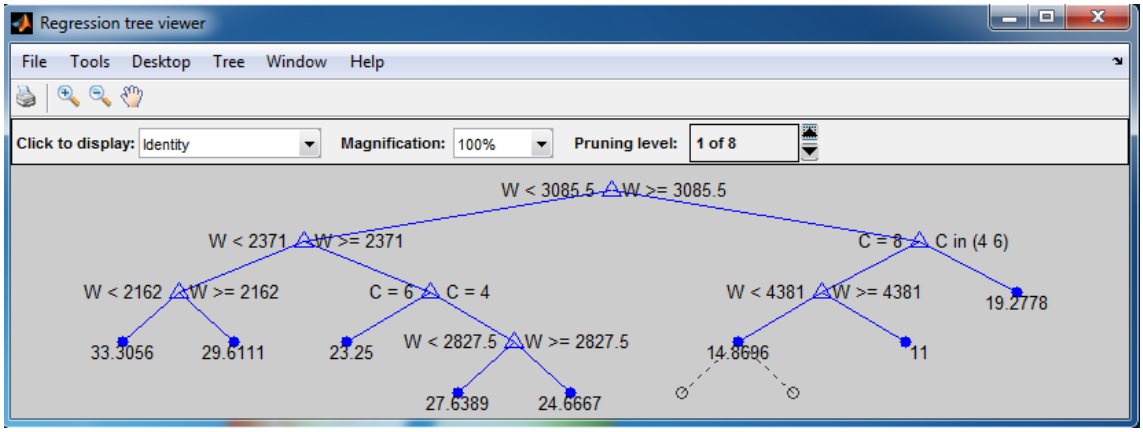
This process continues to find the next weakest subtree until obtaining a tree contains only the root node.

$$T \succ T_1 \succ T_2 \succ T_3 \ldots \succ \{r\} \tag{3.58}$$

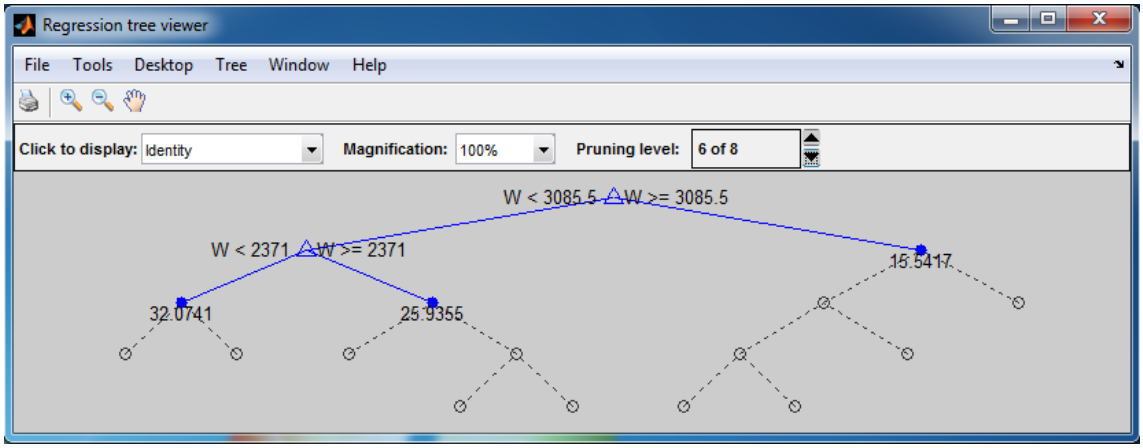
Figure 3-12 shows an example of regression tree which is pruned at different pruning levels.



(a)



(b)



(c)

Figure 3-12 regression tree example with different pruning levels, (a) the original tree without pruning, (b)the tree after pruning the weakest subtree (pruning level ,k=1), (c) the tree after pruning the 6 waekest subtrees (k=6)

3.3.6.5 *Cross-Validation*

In the previous section, we illustrated the importance of tree pruning to reduce the misclassification rate for any new test data, and the use of the Minimum Cost-Complexity pruning algorithm to find the weakest subtree for each pruning level. In this section we describe the use of *v-fold cross-validation* to find the best pruning level which minimizes the misclassification rate of the tree for any new test data.

In *v-fold cross-validation*, the training set $\mathfrak{Z}$ is divided into v separate subsets $\mathfrak{Z}_1, \mathfrak{Z}_2, \ldots, \mathfrak{Z}_v$ which contain the same number of training samples. The i^{th} training set $\mathfrak{Z}^i$ is defined as follows:

$$\mathfrak{Z}^i = \mathfrak{Z} - \mathfrak{Z}_i \qquad i = 1, 2, \ldots, v \tag{3.59}$$

v is usually selected to be high, like 10. The main tree T is grown on the training set $\mathfrak{Z}$, and v auxiliary trees are grown on the v training sets defined by Equation (3.59).

The misclassification rate of the new test data for the main tree can't be estimated directly since we don't have any test data, but this rate can be approximated via the test-set misclassification measure of the v auxiliary trees as follows:

$$R^{CV}(T) = \frac{1}{v}\sum_{i=1}^{v} R^*(T^i) \tag{3.60}$$

Where $R^{CV}(T)$ is the cross-validation estimate of the misclassification rate of the main tree T, and $R^*(T^i)$ is the misclassification rate of the ith tree, which was built on the training set $\mathfrak{Z}^i$ defined by Equation (3.59). The optimal *v*-fold cross-validation pruning level can be obtained through:

$$k^{CV} = \underset{k}{\operatorname{argmin}}\, R^{CV}(T_k) \tag{3.61}$$

where T_k is the pruned tree at pruning level k.

Chapter4: System overview

We propose a data-driven emotion conversion system to obtain an expressive Arabic TTS. Literature work on emotion conversion ensures that converting both spectral and prosodic parameters are necessary for emotion conversion. In our first trial to perform emotion conversion, we employed a phoneme-based spectral conversion using temporal decomposition and Gaussian mixture model, which has been proposed in [58] for voice conversion. In [25] a similar method is proposed for emotion conversion, and in addition to spectral conversion; prosody conversion is performed.

After the implementation of the proposed spectral conversion technique in [58]; we tested the system for questioning and found that no change in expressiveness occurred (i.e. the utterance remains neutral). We have made several efforts to modify this method including adding pitch value of every frame to the spectral parameter vector and mapping both spectrum and pitch contour [59], trying to add delta and delta-delta features [25], trying different machine learning techniques like neural networks, and finally changing the window size, LSF order, and the training data size; but all these efforts didn't improve the results.

Finally, we found two studies on emotion conversion; where both prosodic and spectral parameters are modified based on their linguistic context [23, 32], and their results are reasonable in terms of quality and expressiveness, so our proposed system is based on these two studies.

In this chapter, we first illustrate the Phoneme-based Spectral Conversion Using TD and GMM. Then, we give an overview of our proposed emotion conversion system to obtain expressive speech.

4.1 Emotion conversion based on "Phoneme-based Spectral Conversion Using Temporal Decomposition and Gaussian Mixture Model"

This method can be summarized as follows: a trained GMM is used to transform LSF parameters, which represent the speech spectrum, to the desired expression. The method employs STRAIGHT analysis-synthesis [60] to get a smoothed spectrum and temporal decomposition is used to remove the discontinuities between the converted frames which results from traditional GMM as will be illustrated in the following subsections. Two stages for the implementation of this system: training and transformation stages.

To train the GMM: for each parallel utterances (neutral-expressive utterances) in the training database; the input speech signal is decomposed into spectral envelopes, F0 (fundamental frequency) information, and aperiodic components (AP) using STRAIGHT. From the spectral envelopes; LSF parameters are calculated, and then decomposed into event targets and event functions using Modified Restricted Temporal Decomposition (MRTD). Event targets of parallel utterances are modeled using GMM. The training procedure is illustrated in Figure 4-1.

In the transformation phase: for any new utterance, spectral envelopes are calculated from the speech signal using STRAIGHT; then LSF parameters are calculated from these spectral envelopes; and then using MRTD, LSF parameters are decomposed into event targets and event functions. These event targets are transformed to their expressive values using the trained GMM. The reverse process is performed to obtain the synthesized converted speech signal: using MRTD synthesis, the modified event targets are re-synthesized with the original event functions (unmodified) to obtain the new LSF parameters; then spectral envelopes are obtained from LSF synthesis. Finally, STRAIGHT synthesis is used to obtain the converted speech. The transformation procedure is illustrated in Figure 4-2.

4.1.1 *Modified Restricted Temporal Decomposition*

Traditional GMM-based spectral conversion do not take into account the relationship between frames in both training and conversion procedures, which causes discontinuous in the converted spectral contours and these discontinuities are heard as clicks in the converted speech. The use of temporal decomposition is proposed in [58] to solve this problem.

Temporal decomposition (TD) is proposed by Atal [61] for efficient coding of speech. Speech parameters (LPC or other parameter) are decomposed into event functions and event targets. Let $y_i(n)$ be the i^{th} speech parameter at frame n, then $y_i(n)$ can be represented as:

$$\hat{y}_i(n) = \sum_{k=1}^{K} a_k \varphi_k(n) \qquad 1 \le n \le N \ , 1 \le i \le P \tag{4.1}$$

where $\hat{y}_i(n)$ is the approximate value of $y_i(n)$ resulted from TD synthesis. $\varphi_k(n)$ is the k^{th} event function (interpolation function) at frame n, and a_k is the k^{th} event target or the contribution of the k^{th} event function to the i^{th} speech parameter. The value of K refers to the number of event functions in the speech segment which contains N frames ($N \gg K$).

Event targets can be seen as an average values over a set of frames and are interpolated using their event functions to re-produce the original speech parameters. The matrix form of Equation (4.1) is:

$$\hat{Y} = A\Phi \qquad \hat{Y} \in R^{P \times N}, A \in R^{P \times K}, \Phi \in R^{K \times N} \tag{4.2}$$

where $P, N,$ and K are the order of the spectral parameters, the number of frames in the speech segment, and the number of event functions, respectively.

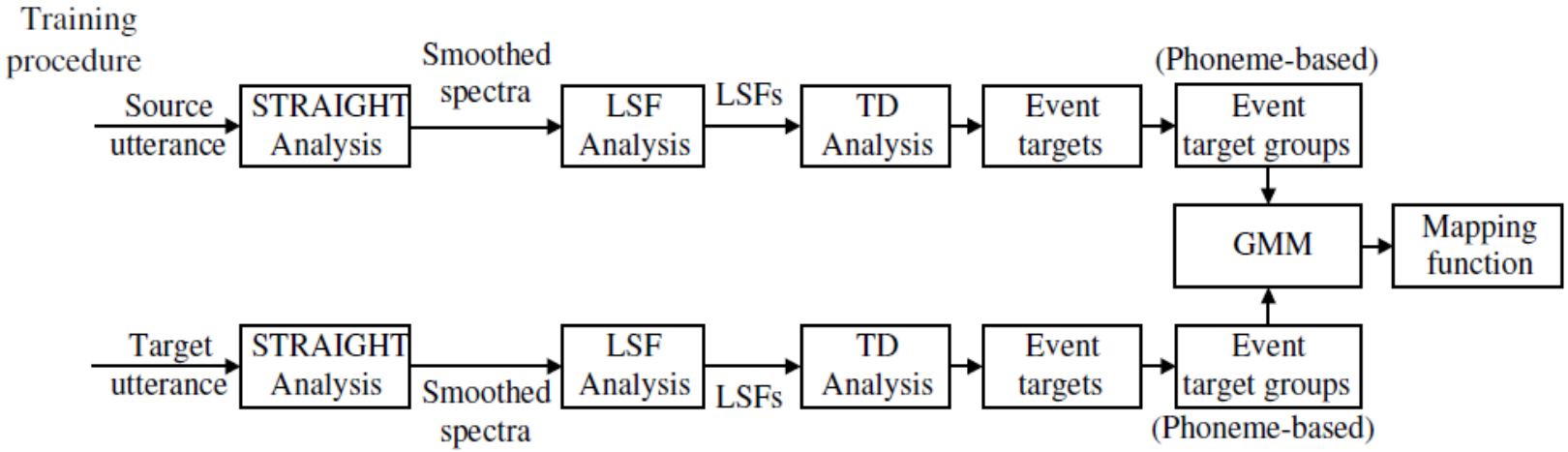


Figure 4-1 training procedure of Phoneme-based Spectral Conversion Using TD and GMM [58].

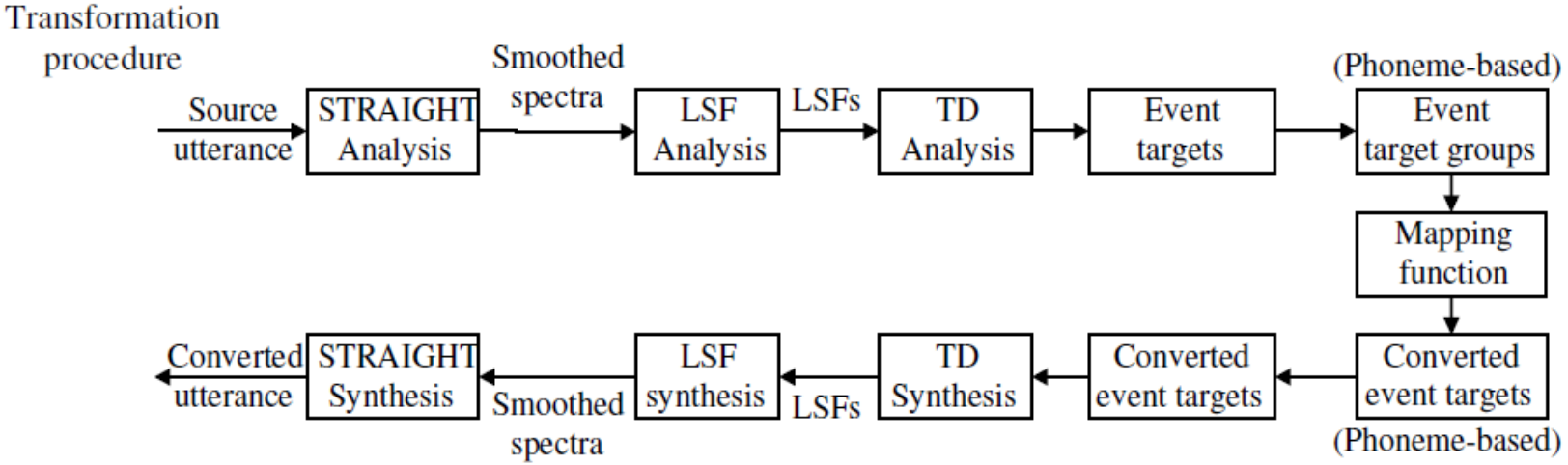


Figure 4-2 transformation procedure of Phoneme-based Spectral Conversion Using TD and GMM [58].

Modified Restricted Temporal Decomposition (MRTD) [62] is a modified form of TD, where only two event functions can overlap and sum up to one at any time, thus Equation (4.1) can be written as:

$$\hat{y}_i(n) = a_k\varphi_k(n) + a_{k+1}\varphi_{k+1}(n) \qquad n_k \le n \le n_{k+1} \tag{4.3}$$

where n_k, and n_{k+1} are the locations of event k and event $k+1$, respectively. Since $\varphi_k(n)$ and $\varphi_{k+1}(n)$ sum up to one in the interval $n_k \le n \le n_{k+1}$, the above equation can be written as follows:

$$\hat{y}_i(n) = a_k\varphi_k(n) + a_{k+1}\big(1 - \varphi_k(n)\big) \qquad n_k \le n \le n_{k+1} \tag{4.4}$$

The event function in MRTD is well-shaped which means that the event function has only one peak, this peak exists at the event location as shown in Figure 4-3. Event locations were set based on phoneme locations as proposed in [63]; where each phoneme is divided into four equal parts, and the event locations are set at the five points marking these parts. Phoneme locations are extracted using HTK-based forced alignment [64]. Calculation of event target and event function as illustrated in [62] are as follows:

$$\varphi_k(n) = \begin{cases} 1 - \varphi_{k-1}(n), & if\ n_{k-1} < n < n_k \\ 1, & if\ n = n_k \\ min\Big(\varphi_k(n-1), max\big(0, \hat{\varphi}_k(n)\big)\Big), & if\ n_k < n < n_{k+1} \\ 0, & otherwise \end{cases} \tag{4.5}$$

$$\hat{\varphi}_k(n) = \frac{\langle (y(n) - a_{k+1}), (a_k - a_{k+1}) \rangle}{\|a_k - a_{k+1}\|^2} \tag{4.6}$$

where $\langle .,. \rangle$ and $\|.\|$ are the inner product of two vectors and the norm of a vector, respectively.

The event targets are calculated from the event function using the least square sense as follows:

$$A = Y\Phi^T(\Phi\Phi^T)^{-1} \tag{4.7}$$

Using TD, speech parameters can be modified by modifying their event targets and event functions, instead of modifying these parameters frame by frame, this ensures the smoothness of the modified speech through the shape of the event functions.

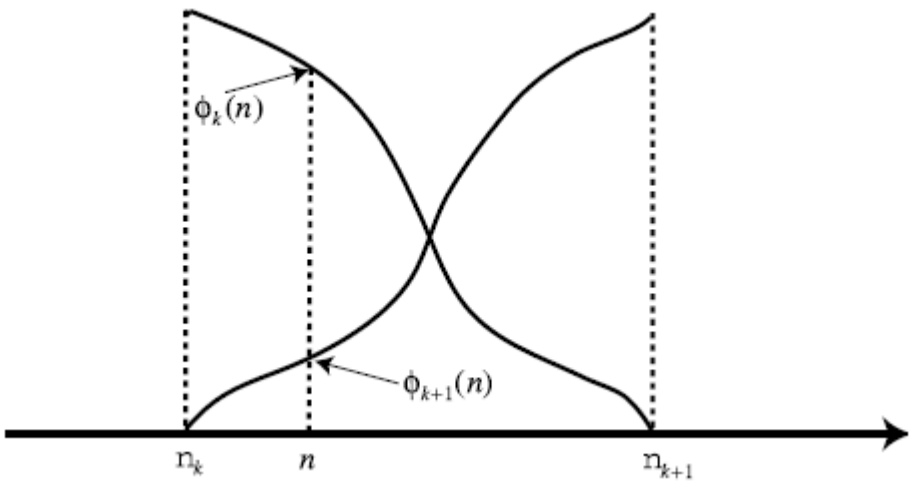


Figure 4-3 Two adjacent event functions in MRTD [62].

4.1.2 *GMM training and transformation function estimation*

As mentioned earlier, two parallel neutral and expressive utterances are introduced to STRAIGHT and LSF analysis, then MRTD decompose the LSF parameters into event targets and event functions. Let $\boldsymbol{X} = [X_1, X_2, \dots, X_N]$ be the event targets of the neutral utterance, where $X_i = [x_1, x_2, \dots, x_P]^T$, P, N are the LSF order, and the number of events in the utterance. The same for the expressive utterance, the event targets would be $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_N]$, and where $Y_i = [y_1, y_2, \dots, y_P]^T$. Note that $\boldsymbol{X}$, and $\boldsymbol{Y}$ are aligned and have the same number of events because event locations in both utterances are set based on phoneme locations.

We define the feature pair $\boldsymbol{Z} = [Z_1, Z_2, \dots, Z_N]$, where $Z_i = [X_i^T, Y_i^T]^T$. The joint distribution $\boldsymbol{Z}$ is modelled by M-mixtures GMM with $\alpha_m, \mu_m, \Sigma_m$ are the weight, mean, and variance of the m^{th} Gaussian component. The probability of any input joint feature vector z will be as follows:

$$p(z) = \sum_{m=1}^{M} \alpha_m N(z; \mu_m, \Sigma_m) \tag{4.8}$$

and

$$N(z; \mu_m, \Sigma_m) = \frac{1}{(2\pi)^{k/2}|\Sigma_m|^{1/2}} \exp\left(-\frac{1}{2}(z-\mu_m)^T \Sigma_m^{-1}(z-\mu_m)\right) \tag{4.9}$$

where $N(z; \mu_m \Sigma_m)$ is the Gaussian probability density function (pdf) of the m^{th} Gaussian component with mean, μ_m, and variance, Σ_m.

GMM parameters $(\alpha_k, \mu_k, \Sigma_k)$ are expected using the expectation–maximization (EM) algorithm illustrated in section 3.3.5.1. Matlab Statistics Toolbox implementation of GMM was used to estimate the model parameters.

4.1.3 *Transformation Procedure*

For any new neutral utterance, the speech waveform is introduced to STRAIGHT and LSF analysis, then MRTD decompose the LSF parameters into event targets and event functions. The output event targets are transformed to their expressive values using the trained GMM as follows:

$$F(x) = E(y|x) = \int_{M} yp(y|x)dy = \sum_{m=1}^{M} p_m(x)\left(\mu_m^y + \Sigma_m^{yx}(\Sigma_m^{xx})^{-1}(x - \mu_m^x)\right) \tag{4.10}$$

$$p_m(x) = \frac{\alpha_m N(x; \mu_x^x, \Sigma_m^{xx})}{\sum_{m=1}^{M} \alpha_m N(x; \mu_m^x, \Sigma_m^{xx})} \tag{4.11}$$

where $\mu_m = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix}$, $\Sigma_m = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix}$, and $\alpha_m(x)$ are the mean, variance, and weight of the m^{th} Gaussian component, respectively, and x, y, are the neutral and converted event targets.

After calculating the modified event targets, they are re-synthesized with the original event functions (unmodified) to obtain the modified LSFs, then spectral envelopes are obtained from LSF synthesis. Finally, STRAIGHT synthesis is used to get the converted speech.

4.2 The proposed system for emotion conversion

As proved in several literature studies [2-5], successful emotion conversion is achieved by converting both prosodic (pitch- duration- intensity), and spectral parameters. In this section, we start with copy-synthesis experiments to check the effect of different speech parameters on each expression, and then give an overview of our proposed emotion conversion system to obtain expressive speech.

4.2.1 *Copy-synthesis experiments*

Copy-Synthesis is a straight approach used to check the effect of prosodic and spectral parameters on each expression and to find the dominant and the ineffective parameters for each expression. An analysis-synthesis scheme is used, and before synthesis the parameter under check is changed to its expressive value [2].

In our copy-synthesis experiments, each of the prosodic and spectral parameters (pitch- duration- intensity- and spectrum) was checked separately, then we tested the effect of changing all of these parameters at one time. Changing one of the four parameters at a time plus changing all the parameters in 3 expressions (sadness- happiness- questioning); resulting in 15 (5×3) utterances in the experiment. Figure 4-4, Figure 4-5, and Figure 4-6 show copy-synthesis modules of spectrum, pitch, and duration. Only one utterance for each expression and one listener was asked to evaluate the output of these experiments; so results can't be generalized. These experiments are used only as indicator that changing these parameters will be useful to obtain expressive speech.

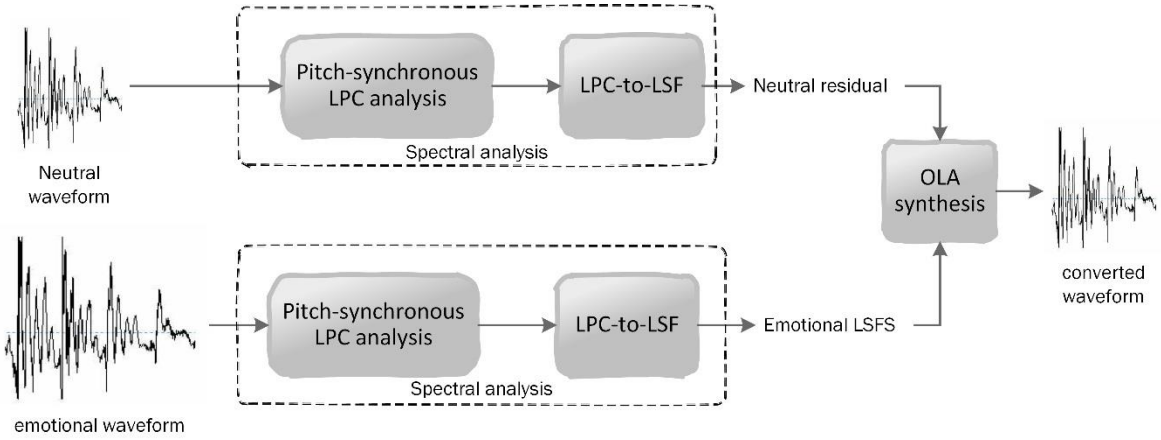


Figure 4-4 copy-synthesis of spectral parameters

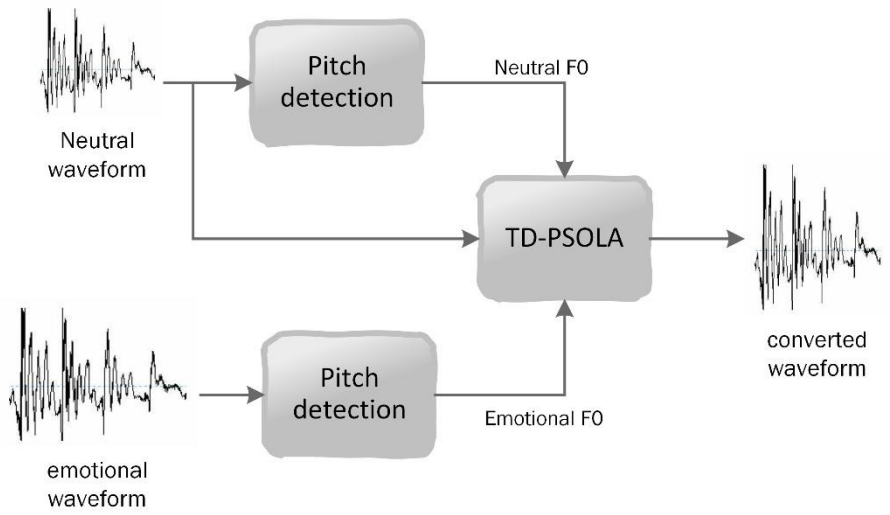


Figure 4-5 copy-synthesis of pitch contour

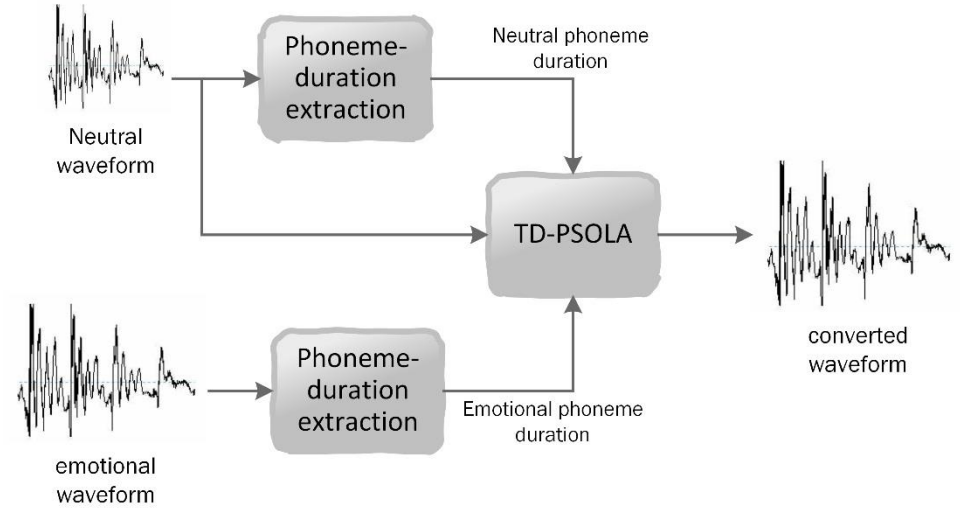


Figure 4-6 copy-synthesis of duration

Results of copy-synthesis experiments for the three expressions (sadness- happiness- questioning) are shown in Table 4-1. Results show that the dominant parameter in all expressions is the pitch variation, while the spectrum change has the lowest effect on all expressions.

Table 4-1 Results of copy-synthesis experiments for different expressions, these results show percentage of the desired expression in the output of each experiment

experiment	Effect on neutral to sad	Effect on neutral to happy	Effect on neutral to question
Change spectrum	10%	30%	0%
Change pitch	40%	60%	80%
Change duration	40%	0%	0%
Change (pitch + duration)	90%	60%	80%
Change (spectrum + pitch + duration)	100%	100%	90%

4.2.2 *System modules*

Our emotion conversion system consists of four modules: pitch conversion, duration conversion, spectral conversion, and energy conversion. The combination of these modules is shown in Figure 4-7, and a summary description for each module is illustrated in the following sub-sections.

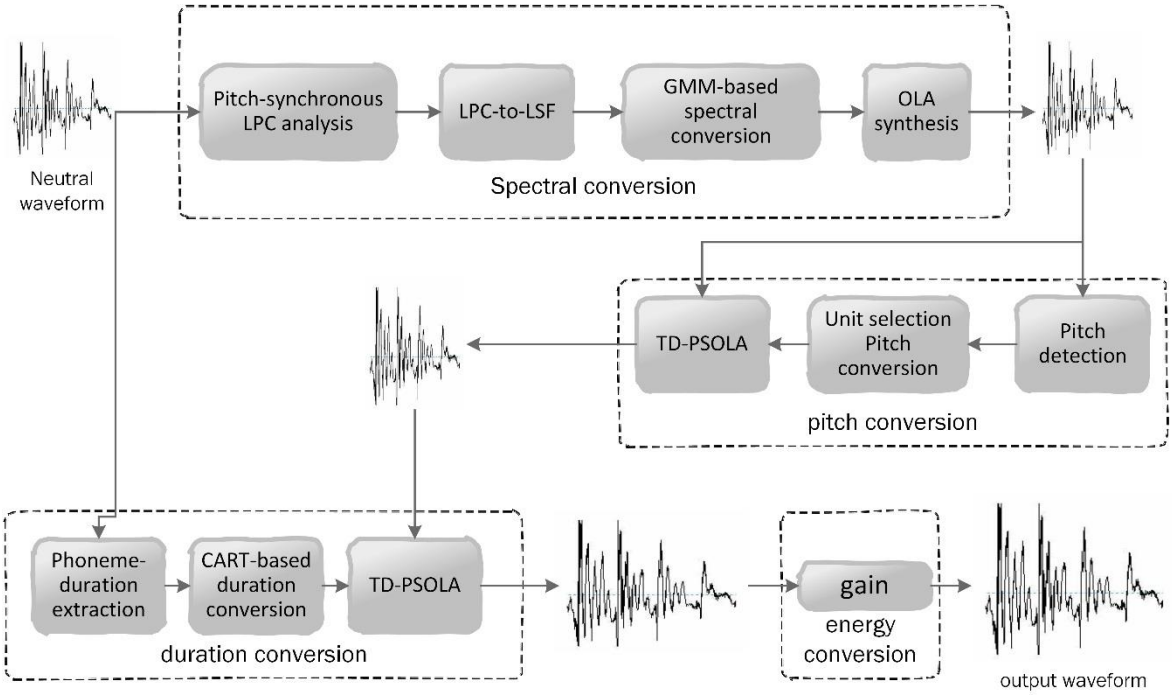


Figure 4-7 proposed emotion conversion system

4.2.2.1 *Spectral conversion*

A GMM framework Based on [23] is used to convert LSF parameters from the neutral to the target expression. 16-mixture GMM were used to fit the thirty-order LSF parameters which were extracted from each pitch-synchronous overlapping frame. DTW is used for time-alignment of the neutral and target expressions. Overlap and add approach is used to synthesize the converted LSFs with the unconverted residual signal.

4.2.2.2 *Pitch conversion*

Unit selection paradigm for pitch conversion using different intonation units and different pitch detectors. Pitch contour of the spectral-converted waveform is extracted; then the best expressive contour, which minimizes the total cost, is selected from a prepared database; and finally the neutral pitch contour is modified to the new converted contour by using Time Domain Pitch Synchronous Overlap Add (TD-PSOLA).

4.2.2.3 *Duration conversion*

Duration conversion is performed on the phoneme level using regression trees, neutral phone durations including pauses are transformed to their expressive durations and the duration is modified in the neutral waveform using Time Domain Pitch Synchronous Overlap Add (TD-PSOLA).

4.2.2.4 *Energy conversion*

The speech signal is multiplied by a factor to increase or decrease the energy level to the level of the target expression.

Chapter5: System implementation

As illustrated in the previous chapter that our emotion conversion system consists of four modules: pitch conversion, duration conversion, spectral conversion, and energy conversion. In this chapter, we illustrate each module in detail.

5.1 Spectral conversion

A GMM framework is used to convert LSF parameters from the neutral to the target expression. GMM is the most popular technique for spectral envelope modification in emotion conversion. A comparison between three different methods for transforming spectral envelopes from neutral to emotional speech is presented in [27] and results showed that weighted frame mapping and GMM based transformations are slightly better than the weighted codebook mapping.

16 Gaussian components were used to fit the thirty-order LSF parameters which were extracted from each pitch-synchronous frame. DTW is used for time-alignment of the LSF parameters of the neutral and the target expression. Overlap and add approach is used to synthesize the converted LSFs with the unconverted residual signal [23, 65]. Spectral conversion module is illustrated in Figure 5-1.

The choice of LSF parameters, to represent the speech spectrum for spectral conversion, is because they have the following characteristics [24]:

- Sensitivity (a bad estimation or modification of one coefficient affects only a port of frequency spectrum around this frequency)
- Efficiency (interpolation of LSFs causes low spectral distortion).
- Reliability (LSFs are reliably estimated).
- Relation to formant frequencies.

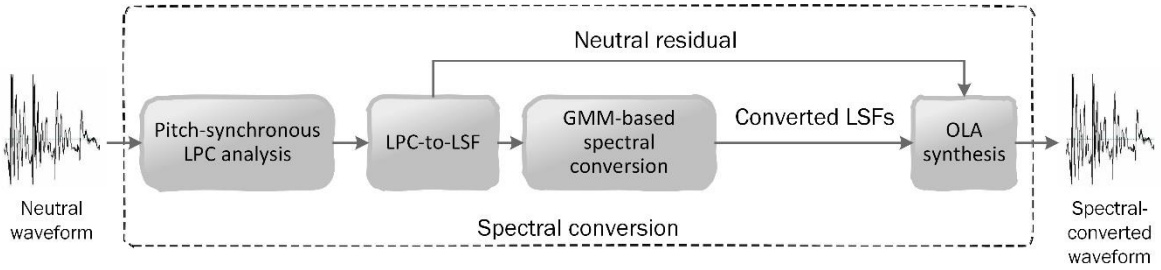


Figure 5-1 spectral conversion module

5.1.1 *Speech Analysis / Synthesis without modification*

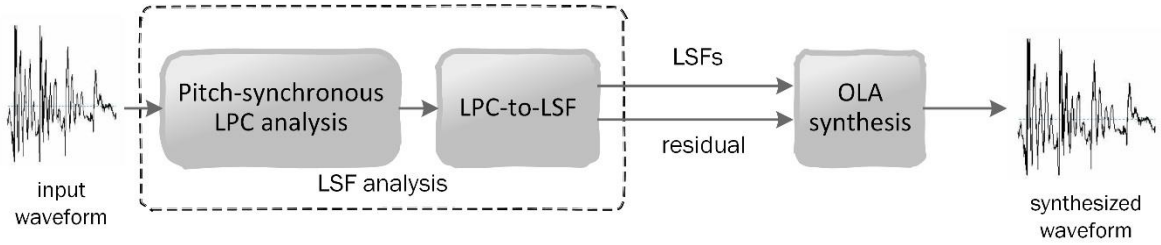


Figure 5-2 LSF analysis/synthesis module

Speech signal is divided into pitch-synchronous overlapped frames with frame length equals three times of the frame shift. Variable frame length of three pitch periods was used for voiced frames, and constant frame length of 10msec was used for unvoiced frames. After dividing the speech signal into frames, a hamming window is applied to each frame. Finally, LPC analysis is applied to each windowed frame, and then the LPC coefficients are converted to their LSF representation and the residual signal is calculated. A high LPC order of thirty was used in order to capture the spectral details.

The voice/voiceless decision of each frame is determined using the pitch marks calculated by PRAAT software [66] as follows: if the distance between two successive pitch marks is lower than a pre-determined threshold (50 Hz), the region between these pitch marks is considered unvoiced.

In the synthesis stage, the reverse operation is used to obtain the speech signal from the LSF parameters. First, LSF parameters are transformed back to their LPC representation. Second, these parameters are used as a filter parameters which filters the residual signal. Finally, Overlap and Add approach (OLA) is used to synthesize the overlapped frames; where the centers of these frames are placed back on their original pitch-marks, and the overlapping regions are added together, as illustrated in section 3.3.2.1.

5.1.2 *Spectral transformation*

A GMM framework is used to convert LSF parameters from the neutral to the target expression. GMM parameters and transformation matrix are estimated during the training phase to be used for any new utterance in the conversion phase.

5.1.2.1 *Training phase*

In the training phase of spectral conversion module, both the GMM parameters and the transformation matrix are estimated from the parallel training data using the following steps:

5.1.2.1.1 Analysis and feature extraction

For each sentence in the training data, LSF parameters of each pitch-synchronous frame are extracted, and then each frame gets a tag identifying the frame type (voiced / voiceless).

5.1.2.1.2 Time alignment

The number of frames in parallel utterances is different due to the difference in articulation and phoneme durations, so time alignment is used to align LSF vectors of parallel utterances (neutral-expressive pair). For more accurate alignment, feature alignment is performed on parallel phonemes in parallel utterances since phoneme locations can be extracted using HTK-based forced alignment [64] (these phoneme locations are then manually corrected to avoid mistakes). DTW is used for time alignment (section 3.3.3), and after alignment each frame in the neutral utterance corresponds to a frame in the expressive utterance.

Let n_t, and e_t denote the neutral LSF vector and its corresponding expressive LSF vector for frame t. They are $k \times 1$ dimensional vectors, where k is the LSF order. After feature extraction and time alignment of all the training data we will have N, E matrices for neutral-expressive utterances,

$$N = [n_1, n_2, n_3, \dots\dots, n_T] \,, \quad E = [e_1, e_2, e_3, \dots\dots, e_T] \tag{5.1}$$

where N, E are $k \times T$ matrices, and T is the total number of aligned frames in the training data. After the construction of N, E matrices, the unvoiced frames will be removed and the training will be done only on the voiced frames.

5.1.2.1.3 GMM fitting of the neutral speech

M-mixtures GMM is used to fit the neutral LSF parameters with $\alpha_m, \mu_m, \Sigma_m$ are the weight, mean, and variance of the m^{th} Gaussian component (see section 3.3.5 for more details). The probability of any input feature vector x will be as follows:

$$p(x) = \sum_{m=1}^{M} \alpha_m N(x; \mu_{m,} \Sigma_m) \tag{5.2}$$

And

$$N(x; \mu_m, \Sigma_m) = \frac{1}{(2\pi)^{k/2} |\Sigma_m|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_m)^T \Sigma_m^{-1} (x - \mu_m) \right) \tag{5.3}$$

where $N(x; \mu_m \Sigma_m)$ is the Gaussian probability density function (pdf) of the m^{th} Gaussian component with mean, μ_m, and variance, Σ_m.

GMM parameters $(\alpha_m, \mu_m, \Sigma_m)$ are expected using the expectation–maximization (EM) algorithm illustrated in section 3.3.5.1. Matlab Statistics Toolbox implementation of GMM is used to estimate the model parameters.

5.1.2.1.4 Transformation matrix estimation

The final step of the training phase is the estimation of the transformation matrix. Our transformation function is given by:

$$e_t = F(n_t) = \left(\sum_{m=1}^{M} P(c_m|n_t) W_m \right) \bar{n}_t \tag{5.4}$$

Where

$\bar{n}_t = [n_t \; 1]'$

n_t, and e_t are the neutral and expressive LSF vectors.

W_m is the transformation matrix, and $P(c_m|n_t)$ is the probability that vector n_t belongs to the mixture class c_m.

$$P(c_m|n_t) = \frac{\alpha_m N(n_t; \mu_m, \Sigma_m)}{\sum_{i=1}^{M} \alpha_i N(n_t; \mu_i, \Sigma_i)} \tag{5.5}$$

The matrix form of the transformation Equation (5.4) is:

$$E = W\psi(N) \tag{5.6}$$

Where

$$W = [W_1 \vdots W_2 \vdots W_3 \vdots \cdots \vdots W_4]_{k \times (M \times (k+1))} \tag{5.7}$$

$$\psi(n) = \begin{bmatrix} P(c_1|n)\bar{n} \\ \vdots \\ P(c_m|n)\bar{n} \\ \vdots \\ P(c_M|n)\bar{n} \end{bmatrix}_{(M \times (k+1)) \times 1} \tag{5.8}$$

$$\psi(N) = \begin{bmatrix} P(c_1|n_1)\bar{n}_1 & & P(c_1|n_t)\bar{n}_t & & P(c_1|n_T)\bar{n}_T \\ \vdots & & \vdots & & \vdots \\ P(c_m|n_1)\bar{n}_1 & \cdots & P(c_m|n_t)\bar{n}_t & \cdots & P(c_m|n_T)\bar{n}_T \\ \vdots & & \vdots & & \vdots \\ P(c_M|n_1)\bar{n}_1 & & P(c_M|n_t)\bar{n}_t & & P(c_M|n_T)\bar{n}_T \end{bmatrix}_{(M \times (k+1)) \times T} \tag{5.9}$$

To estimate the transformation matrix for each mixture component, W_m, Least Squares method (LSE) for Equation (5.6) is used, then the transformation matrix is given by the following equation

$$W = E\psi(N)'(\psi(N)\psi(N)')^{-1} \tag{5.10}$$

5.1.2.2 *Conversion phase*

Spectral conversion for any test utterance involves the following steps:

- Analysis and feature extraction: LSF parameters of each pitch-synchronous frame are extracted, and then each frame gets a tag identifying the frame type (voiced / voiceless).
- The transformation formula given by Equation (5.6) is applied on the voiced frames, and the unvoiced frames remains as in the neutral case.
- New LSF parameters are transformed to their LPC representation, and then LPC parameters are used as a filter parameters which filters the original residual signal. Finally Overlap and add approach (OLA) is used to synthesize the overlapping converted frames.

5.2 Pitch conversion

Unit selection paradigm for pitch conversion (selection of expressive units from a database according to a cost function) is presented in [23, 32] and gives good performance. Several studies used the syllables as the basic intonation unit for pitch conversion [23, 29, 67]. However, the proposed work in [32] shows that the choice of the best intonation unit depends on the language and the word unit is the best intonation unit for Basque language. In our work on Arabic, we applied unit selection paradigm for pitch conversion and tried two different intonation units (words, and syllables) with the appropriate linguistic features for each unit type. The effect of using different pitch detectors is also examined.

As shown in Figure 5-3, pitch conversion for any input neutral utterance is performed as follows: the pitch contour and the linguistic features of the utterance are extracted. Then using a prepared corpora and a trained cost function, the best expressive contours are chosen from the database. Finally, TD-PSOLA is used to convert the neutral pitch contour into the converted contour.

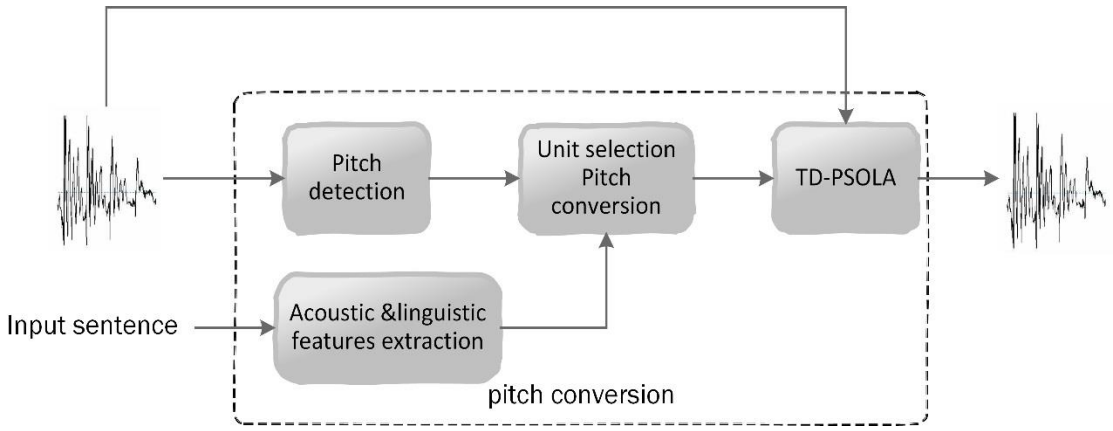


Figure 5-3 pitch conversion module

5.2.1 *Preparation of parallel corpora*

First for each parallel utterances in the input database (and using the text of the recorded sentence), a set of linguistic features and phoneme boundary information are extracted by HTK-based forced alignment; then using these information, it would be a very simple task to obtain the boundaries of our intonation unit (syllable or word). Second, pitch contours of parallel utterances in the training database are extracted, and the unvoiced segments in these contours are interpolated; then using intonation unit boundary information, these contours are chopped into small contours of each intonation unit. Finally neutral-expressive F_0 contours of each intonation unit in the training corpus with their corresponding linguistic features are stored to be used later for the emotional unit selection. Figure 5-4 shows an illustration to corpus preparation.

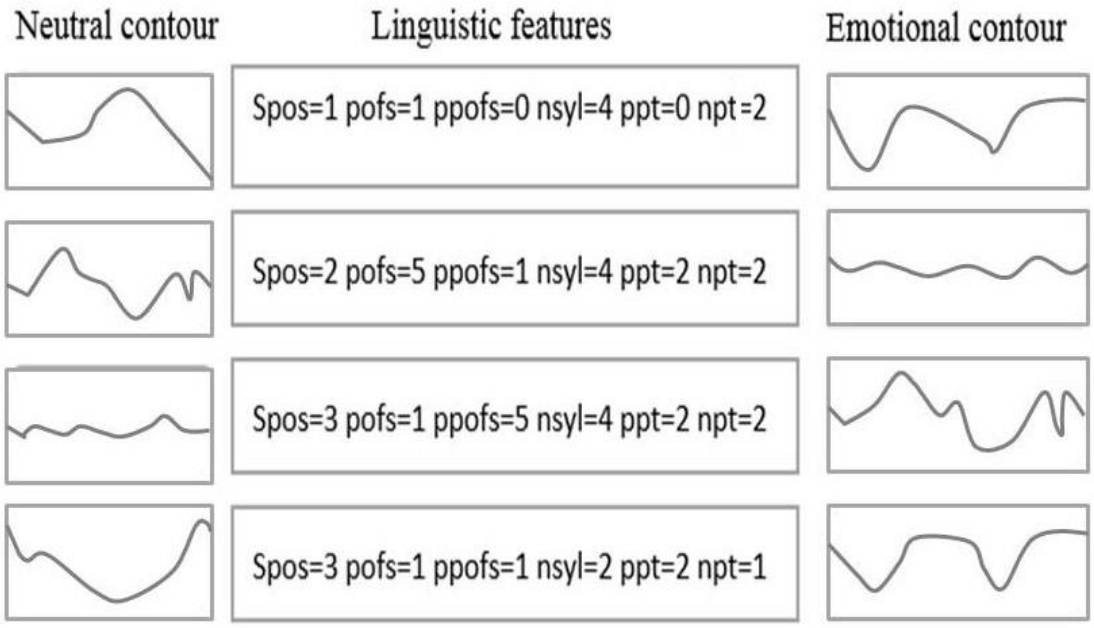


Figure 5-4 overview of word unit corpus (neutral contour- linguistic features- expressive contour).

5.2.1.1 *Linguistic features*

A set of linguistic features of Arabic speech which are thought to be responsible for different intonation are used for pitch conversion, the following features are used for both intonation units (word-syllable):

- *Sentence position, Spos*: identifies the position of the word in the sentence.
- *Part of speech, Pofs*: is the part of speech tag of the current word, thirteen tags are used for Arabic speech.
- *Previous part of speech tag, PPofs:* is the part of speech tag of the previous word.
- Number of syllables in the current word, Nsyl.
- *Previous pause type, Ppt*: is the type of the pause before the current word (no pause, full stop, or other).
- *Next pause type, Npt*: is the type of the pause after the current word (no pause, full stop, or other).

In the case of using syllables as the basic intonation units, a set of features related to syllables are added to the previous features:

- Syllable position in a word, *Wpos*.
- Structure of the syllable, *Sylstruct:* 5 syllable structures exist in Arabic, (CVV, CV, CVC, CVVC, and CVVC).
- Vowel identity within the syllable, *Sylvow*.

These features and their possible values are shown in Table 5-1, and Table 5-2.

Table 5-1 linguistic features for pitch conversion and their values

Feature Name	Symbol	Possible Values
Sentence position	Spos	· 1: for the first word in the sentence · 2: for the second word in the sentence · 3: for any n^{th} word in the middle of the sentence · 4: for the word before the last word in the sentence · 5: for the last word in the sentence
Part of speech / Previous part of speech	Pofs / Ppofs	· 1: noun (proper) [مصر – محمد] · 2: noun (number) [أحد – إثنان] · 3: noun (quantity) [كل – بعض] · 4: noun+ pronoun [موضعه – حياتنا] · 5: noun (other) [الحديث - الإنسان] · 6: adjective [المخلص – مفيد] · 7: comparative adjective [– أكثر أول] · 8: adverb [كلما – فوق] · 9: pronoun [هو – هي] · 10: verb [عاد - اشترت] · 11: particle [لا – لم] · 12: preposition [في – من] · 13: conjunction [و – إذا – ما – أن]
Number of syllables in the word	Nsyl	· 1, 2, 3, 4, …….
Previous pause type / Next pause type	Ppt / Npt	· 0: no pause · 1: full stop · 2: any other type of pauses

Table 5-2 additional linguistic features for syllable-based pitch conversion

Feature Name	Symbol	Possible Values
word position	Wpos	· 0: for single syllable word · 1: for the first syllable in the word · 2: for the middle syllables in the word · 3: for the last syllable in the word
Syllable structure	Sylstruct	· 1: if syllable structure is CV or CVV · 2: if syllable structure is CVC or CVVC · 3: if syllable structure is CVCC
Vowel identity in syllable	Sylvow	· 1: [فتحة مرققه] · 2: [فتحه مفخمه] · 3: [ألف مد مرققه] · 4: [ألف مد مفخمه] · 5: [ضمه] · 6: [واو مد] · 7: [كسره] · 8: [ياء مد]

5.2.2 *Conversion of pitch contour*

First, for any new input neutral utterance to be converted to the expressive case, the linguistic features and the pitch contour of each intonation unit is calculated as illustrated in section 5.2.1. Second a target cost between the input unit, *i*, and each candidate unit *n* is calculated (candidates are the stored units out of the parallel corpora preparation), then the best unit from the candidates which minimizes the target cost is determined, and this operation is done for all input units. Finally, the best candidates' expressive contours are assembled to obtain the desired pitch contour of the utterance, and then the pitch contour is modified in the input waveform using TD-PSOLA (illustrated in section 3.3.2).

Target cost T, between the input unit *i* and the candidate unit *n* is the sum of two sub-costs: acoustic sub-cost A, and linguistic sub-cost L.

$$\mathrm{T(i,n)} = \mathrm{A(i,n)} + \mathrm{L(i,n)} \tag{5.11}$$

The acoustic sub-cost is determined between the neutral contours of the two units, i, n, and it is a weighted sum of the contour distance, d_{cont}, mean log F0 level distance, d_{lev}, and length distance, d_{len} [32]. To calculate these distances, the durations of the two $F0$ contours are normalized to 1, and then their log F0 contours are represented in m^{th} order polynomials $P(t)$ (4^{th} order polynomial was proved to be enough to capture the log F0 contour [32]). The use of log F0 instead of $F0$ is because the frequency distances are perceived by the human ear in a logarithmic scale. The three distance values and the acoustic cost are calculated according to Equations (5.12 - 5.15)

$$d_{cont}(i,n) = \sqrt{\int_0^1 \left(\check{P}_i(t) - \check{P}_n(t)\right)^2 dt} \tag{5.12}$$

where

$$\check{P}(t) = P(t) - \mu_P\,, \quad \mu_P = \int_0^1 P(t)dt \tag{5.13}$$

and

$$d_{lev}(i,n) = \left|\mu_{P_i} - \mu_{P_n}\right| \tag{5.14}$$

$$d_{len}(i,n) = \left|\log\left(\frac{T_i}{T_n}\right)\right| \tag{5.15}$$

where T_i, T_n are the durations of the two contour, then the acoustic sub-cost is:

$$\mathrm{A(i,n)} = \mathrm{w_{cont}d_{cont}}(i,n) + \mathrm{w_{lev}d_{lev}}(i,n) + \mathrm{w_{len}d_{len}}(i,n) \tag{5.16}$$

The linguistic sub-cost between the two units, i, n, is a weighted sum of the distance between the linguistic features of the two units, $d_p(i,n)$, this distance is zero for matching features and one for mismatch.

$$\mathrm{L(i,n)} = \sum_{\mathrm{p=1}}^{\mathrm{P}} \mathrm{w_p d_p(i,n)} \tag{5.17}$$

where P is the number of linguistic features.

Substituting Equation (5.16), and Equation (5.17) in Equation (5.11), we will obtain the target cost as follows:

$$\mathrm{T(i,n)} = \mathrm{w_{cont} d_{cont}}(i,n) + \mathrm{w_{lev} d_{lev}}(i,n) + \mathrm{w_{len} d_{len}}(i,n) + \sum_{\mathrm{p=1}}^{\mathrm{P}} \mathrm{w_p d_p(i,n)} \tag{5.18}$$

The concatenation cost between units is neglected because the intonation unit is large enough to reduce the effect of this cost.

Pitch modification in the waveform is implemented using PRAAT software [66], only voiced parts are converted and the unvoiced parts remain unconverted, so interpolating these parts wouldn't affect the output. This interpolation only makes the selection of units more accurate.

5.2.3 *Estimation of the cost function weights*

During the conversion of pitch contours, the cost is a weighted sum of different distances; these weights are used to normalize distances and to rank the features by their impact. To calculate these weights four methods are proposed in [32]: (the Full Prediction Method, the Acoustic-Only Method, the Linguistic-Only Method, and the Classification& Prediction Method), in our system the second method gave the worst results and the remaining methods almost had the same effect; so we used the Full Prediction Method which can be illustrated as follows:

For a unit j in our prepared corpora, the target cost between the neutral part of this unit and the neutral part of another different unit k in the corpora is calculated as follows:

$$\begin{aligned}\mathrm{T(j,k)} = {} & \mathrm{w_{cont} d_{cont}}(N_j, N_k) + \mathrm{w_{lev} d_{lev}}(N_j, N_k) + \mathrm{w_{len} d_{len}}(N_j, N_k) \\ & + \sum_{\mathrm{p=1}}^{\mathrm{P}} \mathrm{w_p d_p(j,k)}\end{aligned} \tag{5.19}$$

where $\mathrm{w_{cont}}$, $\mathrm{w_{lev}}$, $\mathrm{w_{len}}$, and $\mathrm{w_p}$ are the weights of different sub-costs to be calculated, $\mathrm{d_{cont}}(N_j, N_k)$, $\mathrm{d_{lev}}(N_j, N_k)$, and $\mathrm{d_{len}}(N_j, N_k)$ are the contour distance, mean $\log \mathrm{F0}$ level distance, and length distance between the neutral contours of the two units j, k, respectively. $\mathrm{d_p(j,k)}$ is the distance between the linguistic features of the two units. The calculation of these distances is illustrated in section 5.2.2.

To estimate the weights of different sub-costs, the target cost between the neutral parts of two different units in the corpora is set equal to the acoustic distance between their corresponding expressive contours as follows:

$$\mathrm{T(j,k)} = 0.5 * \mathrm{d_{cont}}\left(E_j, E_k\right) + 0.5 * \mathrm{d_{lev}}\left(E_j, E_k\right) \tag{5.20}$$

Substituting from Equation (5.19) in Equation (5.20) we obtain the following equation:

$$\begin{aligned} &\mathrm{w_{cont}d_{cont}}\left(N_j, N_k\right) + \mathrm{w_{lev}d_{lev}}\left(N_j, N_k\right) + \mathrm{w_{len}d_{len}}\left(N_j, N_k\right) \\ &\quad + \sum_{\mathrm{p}=1}^{\mathrm{P}} \mathrm{w_p d_p(j,k)} = 0.5 * \mathrm{d_{cont}}\left(E_j, E_k\right) + 0.5 * \mathrm{d_{lev}}\left(E_j, E_k\right) \end{aligned} \tag{5.21}$$

Equation (5.21) is applied on each unit j in the parallel corpora with the remaining units k, to obtain a set of equations which can be solved using least squares to find the weights estimate.

5.2.4 *Choice of pitch detector*

The choice of pitch detector must first ensure high analysis-synthesis quality using PRAAT software (the software used for the conversion). Then it would be checked for the emotion conversion. In the implemented system, two well-known pitch detectors were examined (PRAAT (section3.3.4.1), and MBSC (section 3.3.4.2)) and their performance was compared for such application.

The selection of PRAAT pitch detectors is because pitch modification is used using PRAAT software, and the default pitch detector is PRAAT [54], while the selection of MBSC pitch detector [52] because it gives accurate results in both clean and noisy speech.

5.3 Duration conversion

Duration conversion is performed on the phoneme level. Regression trees are used to transform neutral phone durations including pauses to their expressive durations, then the neutral durations are modified to their new values in the speech waveform using TD-PSOLA [32]. As shown in Figure 5-5, duration conversion involves three basic blocks: phoneme-duration and linguistic features extraction, CART-based duration conversion, and waveform modification using TD-PSOLA.

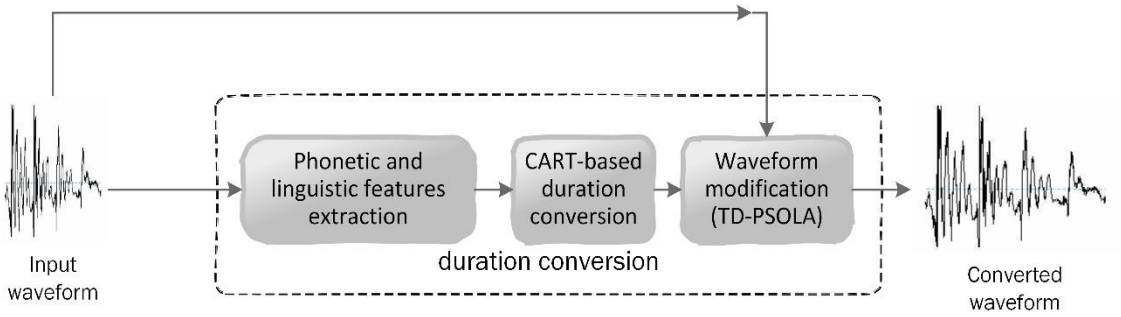


Figure 5-5 duration conversion module

5.3.1 *Phonetic and linguistic features extraction*

Phoneme duration in different expressions depends on different phonetic and linguistic parameters of the phoneme like: phonetic label, type (vowel/consonant), manner of articulation, voicing (voiced/unvoiced), label and manner of articulation of the previous and next phonemes, neutral phone duration, phoneme position within the syllable (first, middle, and last) , position of the syllable within the word (single syllable, first, middle, and last), position of the word within the sentence (initial, second, final, after pause, before pause, none), and part of speech tag of the word containing the phoneme and the previous word. These features are extracted using HTK-based forced alignment [64].Table 5-3, and Table 5-4 show values of phonetic and linguistic features for different Arabic phonemes.

5.3.2 *CART-based duration conversion*

Classification And Regression Trees (CART) (section 3.3.6) are used to transform neutral phone durations to their expressive values. Phonetic and linguistic features of the neutral phonemes are introduced as input to the regression tree and the output is the ratio between the phoneme duration of the target expression and of the neutral.

5.3.2.1 *Training of Regression Trees*

Matlab Statistics Toolbox implementation of classification and regression trees was used to train our trees as follows:

- Phonetic and linguistic features of each neutral phoneme in the training corpus and its corresponding expressive phoneme duration are extracted using HTK-based forced alignment.

- The optimal minimum leaf parameter which minimizes the error is calculated, this parameter specifies the minimum number of observations per tree leaf.
- The tree is built using this minimum leaf parameter.
- The optimal pruning level is calculated by minimizing the cross-validated error
- The tree is pruned to the optimal pruning level.

An example of regression tree for phoneme duration conversion from neutral to happy is shown in Figure 5-7.

5.3.3 *Waveform modification using TD-PSOLA*

As shown in chapter 2, TD-PSOLA is used for time scale modification, we used PRAAT software implementation of TD-PSOLA to perform phoneme duration modification. The duration tier is determined, duration tier is an object specifies the duration ratio between expressive and neutral phonemes at each time instant of a speech signal as shown in Figure 5-6, and this duration tier is presented to PRAAT software with the neutral waveform to perform duration modification.

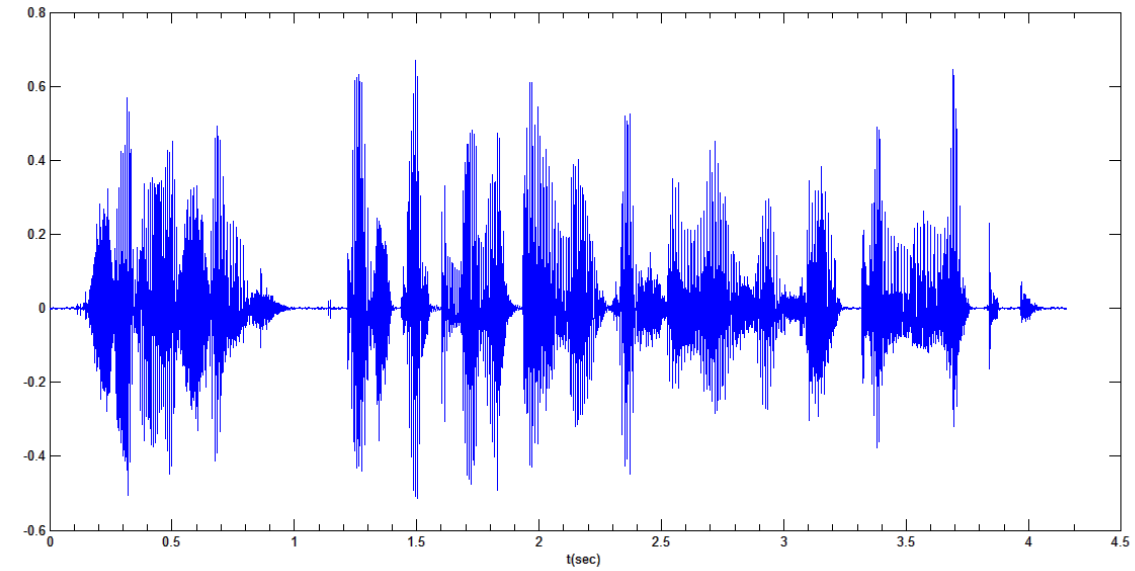
Table 5-3 phonetic parameters of Arabic phones

Feature Name	Symbol	Possible Values
Phonemic label / label of two previous and next phonemes	Phoneticlabel	· Phoneme id from 1 to 39
Vowel/consonant	Vowcon	· 1: vowel · 0: consonant
Manner of articulation / type of the two previous and next phonemes	Artmanner	· 1: Vowels · 2: Fricatives · 3: Glides · 4: laterals · 5: Trills · 6: Nasals · 7: Affricates · 8: Stops · 9: Pauses
Voiced/unvoiced	VUV	· 1: voiced · 0: unvoiced
position within the syllable	sylpos	· 1: first phoneme in the syllable · 2: middle phonemes in the syllable · 3: last phoneme in the syllable
position of the syllable within the word	wpos	· 0: single syllable word · 1: first syllable in the word · 2: middle syllables in the word · 3: last syllable in the word
position of the word within the sentence	Spos	· 1: first word in the sentence · 2: second word in the sentence · 3: previous to last word in the sentence · 4: last word in the sentence · 5: word after pause · 6: word before pause · 7: any other word

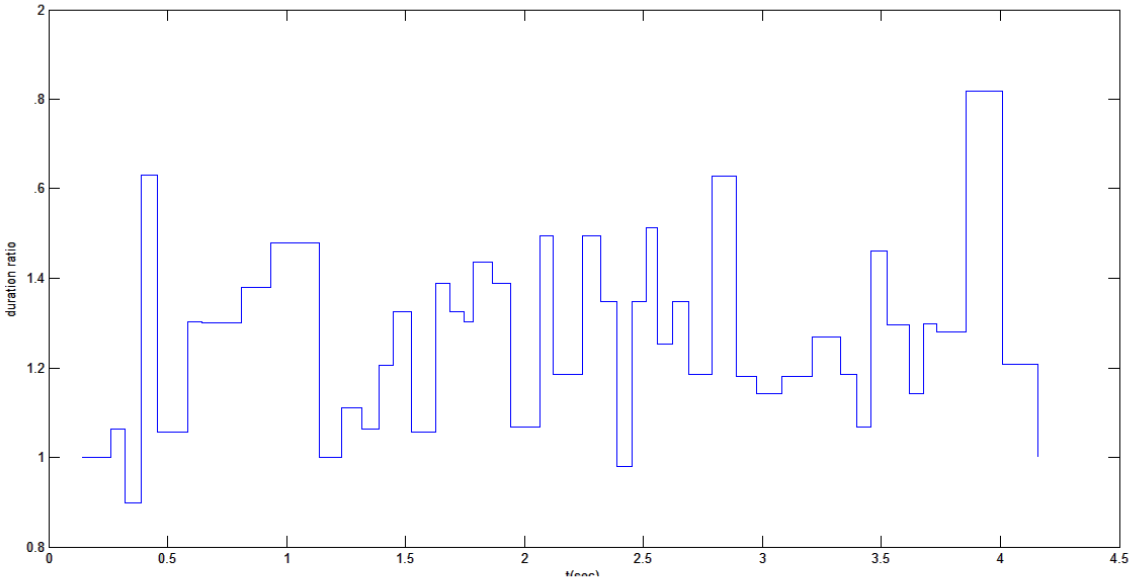
Table 5-4 Features of Arabic phonemes

Arabic phoneme	Phoneme ID	Vowel/Consonant	Voiced/Unvoiced	Manner of articulation
فتحه مرققه	1	V	V	Vowel
فتحه مفخمه	2	V	V	Vowel
ألف مد مرقق	3	V	V	Vowel
ألف مد مفخم	4	V	V	Vowel
ضمه	5	V	V	Vowel
واو مد	6	V	V	Vowel
كسره	7	V	V	Vowel
ياء مد	8	V	V	Vowel
همزة	9	C	UV	Fricative
ب	31	C	V	Stop
ت	32	C	UV	Stop
ث	10	C	UV	Fricative
ج	30	C	V	Affricate
ح	11	C	UV	Fricative
خ	12	C	UV	Fricative
د	33	C	V	Stop
ذ	13	C	V	Fricative
راء مرققة	26	C	V	Trill
راء مفخمة	27	C	V	Trill
ز	14	C	V	Fricative
س	15	C	UV	Fricative
ش	16	C	UV	Fricative
ص	17	C	UV	Fricative
ض	34	C	V	Stop

ط	35	C	UV	Stop
ظ	18	C	V	Fricative
ع	36	C	V	Stop
غ	19	C	V	Fricative
ف	20	C	UV	Fricative
ق	37	C	UV	Stop
ك	38	C	UV	Stop
ل مرققة	24	C	V	Lateral
ل مفخمة	25	C	V	Lateral
م	28	C	V	Nasal
ن	29	C	V	Nasal
هـ	21	C	UV	Fricative
و	22	C	V	Glide
ي	23	C	V	Glide
علامات الترقيم (سكوت)	39	-	-	pauses



speech waveform



duration tier

Figure 5-6 speech waveform of the sentence "شرم الشيخ تستقبل آلاف السياح في كل وقت" and its calculated duration tier for sadness.

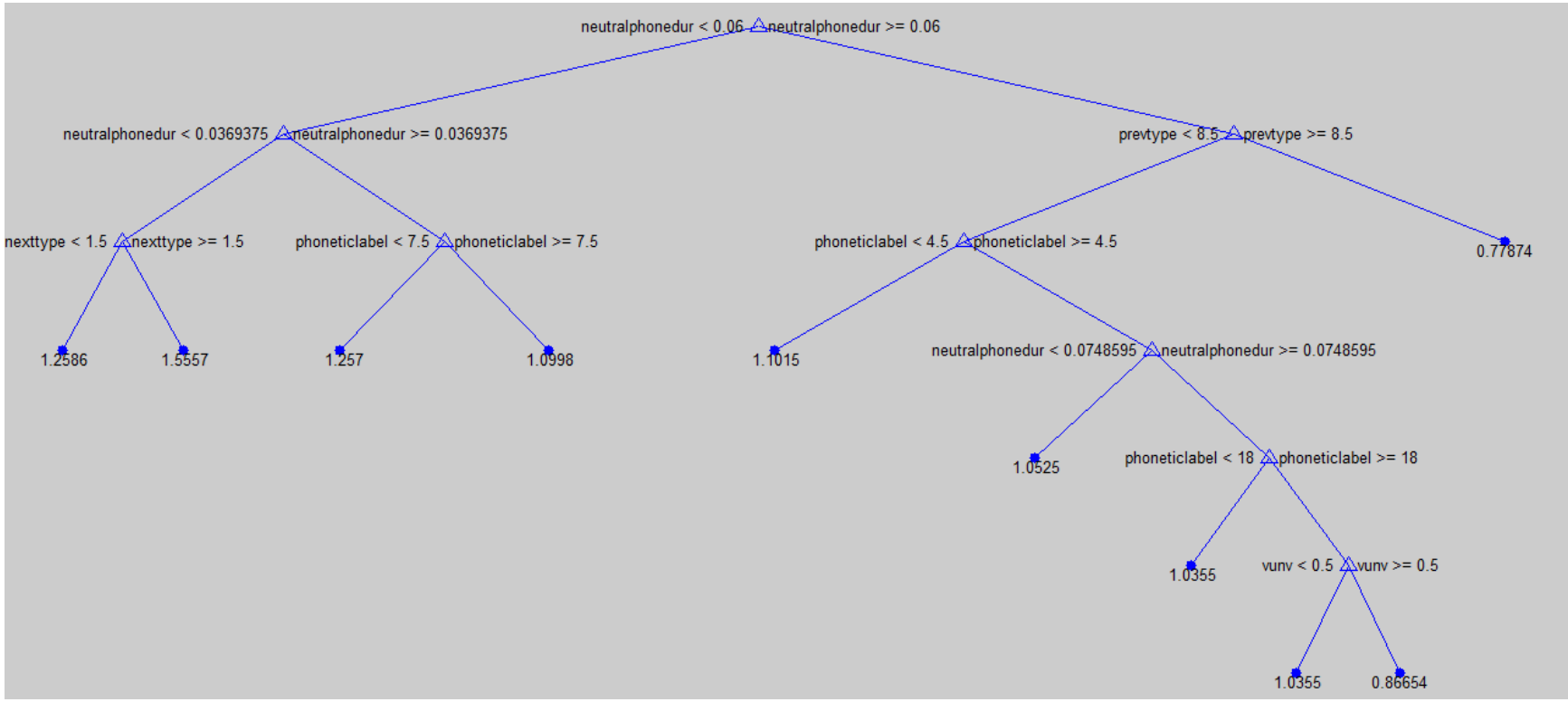


Figure 5-7 Regression tree for phoneme duration conversion from neutral to happy.

5.4 Energy conversion

Changing the energy level is required for some expressions like sadness and anger. It was observed from the analysis of parallel corpus that energy levels of sadness is lower than those of neutral. Energy conversion in our system is achieved by multiplying the converted signal by a factor. This multiplication factor is obtained from the training utterances by calculating the average power ratio of the target expressive to the neutral signals [2].

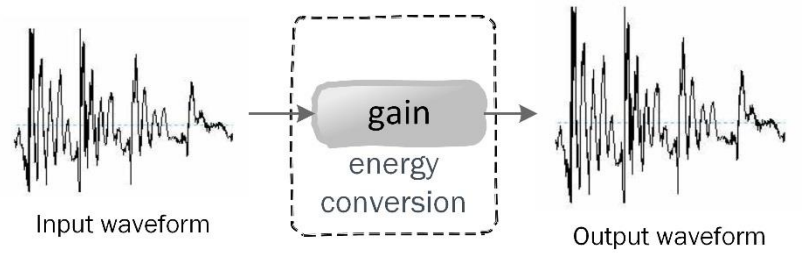


Figure 5-8 energy conversion module

Chapter6: System evaluation

To evaluate our system for emotion conversion in Arabic, first a speech corpus of neutral and other important expressions was recorded, then different experiments were carried out to evaluate pitch conversion using different intonation units and different pitch detectors, then evaluate the effect of different modules in our system, and to evaluate the quality and expressiveness of the overall system. Finally, subjective tests were used for the evaluation, where a reasonable number of volunteers were asked to evaluate the converted speech.

6.1 Expressive speech data collection

Sadness, happiness, and questioning expressions were chosen in our study because they are thought to be useful for expressive Arabic TTS.

Expressive speech data for sadness and happiness was recorded using a subset of unit selection concatenative TTS text corpus designed by the Research & Development International Company (RDI®), where a professional male speaker , who recorded the TTS corpus, was asked to record this subset in two expressions (sadness and happiness). The corpus was recorded at a sampling rate of 96 KHz in a high quality studio, then they were downsampled to 16 KHz. The corpus contains around 100 sentences, 80% of them were used for training, and 20% for test.

Since questioning is an important expression for Arabic TTS, another professional male speaker was asked to record a new text corpus in neutral and questioning expressions. At this time the text corpus was designed to contain sentences with different lengths, and can be uttered as a neutral utterance and as a question. The corpus was recorded at a sampling rate of 44 kHz, and also downsampled to 16 kHz. 280 sentences were recorded, and the best expressive set of them was chosen to be used for training resulting in 68 parallel utterances for training and 15 for test.

Table 6-1 parallel training corpus size for different expressions.

	Neutral-sad	Neutral-happy	Neutral-question
Number of sentences	77	75	68
Total number of words	575	567	295
Total number of syllables	1614	1555	845
Total duration of neutral utterances (min)	5:25	5:16	2:47

6.2 Experimental setup

Two experiments were carried out, the first one was to examine the effect of using different intonation units and different pitch detectors in pitch conversion, while the second experiment was to evaluate the effect of changing different speech parameters (pitch, duration, and spectrum) on both the expressiveness and the quality of the output converted speech. We can also conclude the expressiveness and the quality of the overall emotion conversion system from the second experiment. These experiments are illustrated in the following subsections and the summary of them is shown in Table 6-2.

6.2.1 *Experiment1: evaluation of pitch detection using different intonation units and pitch detectors*

During the first experiment, using only pitch conversion module, each utterance in the test corpus was converted into the target expression using MBSC and PRAAT pitch detectors and using words and syllables as the intonation units, mutually. This experiment results in three transformed voices for each test utterance:

1. Voice1: is the converted speech signal using word based pitch conversion and MBSC pitch detector.
2. Voice2: is the converted speech signal using word based pitch conversion and PRAAT pitch detector.
3. Voice3: is the converted speech signal using syllable based pitch conversion and MBSC pitch detector.

Subjective tests were used for the evaluation, where 38 volunteers participated in the system evaluation only 10 of them were familiar with speech technologies. The volunteers were divided into seven groups and two or three sentences were presented to each group.

An opinion score of five levels was used for judging the quality and expressiveness of the converted utterances [68]. The volunteers were asked to listen to the neutral and expressive test utterances then listen to the three transformed voices and evaluate their expressiveness and quality as illustrated in the evaluation sheet shown in Table 6-3.

6.2.2 *Experiment2: evaluation of conversion modules separately and the overall system performance*

During the second experiment, pitch conversion is performed using MBSC pitch detector and word as the intonation unit. Pitch, duration, and spectrum of each input utterance were transformed to the desired expression by changing: first pitch contours only, second pitch and duration, third prosodic parameters (pitch, duration, and energy), and finally prosodic and spectral parameters of each test utterance. This experiments results in four transformed voices for each test utterance:

1. Voice1: is the converted speech signal after converting only pitch contours.
2. Voice2: is the converted speech signal after converting pitch and duration.

3. Voice3: is the converted speech signal after converting prosodic parameters (pitch, duration, and energy).
4. Voice4: is the converted speech signal after converting all parameters (prosodic and spectral parameters).

Since the energy conversion is performed only in case of sadness, voice 2 and voice 3 are the same for happiness and questioning.

Subjective tests were used for the evaluation, where 17 volunteers participated in the system evaluation only seven of them were familiar with speech technologies. The volunteers were divided into four groups and four or five sentences were presented to each group.

As in experiment1, the volunteers were asked to listen to the neutral and expressive test utterances then listen to the three transformed voices and evaluate their expressiveness and quality as illustrated in the evaluation sheet shown in Table 6-3.

Table 6-2 summary of the experiments

Experiment	Voices	The purpose of the experiment	Number of listeners
Experiment1	· Voice1: is the converted speech signal using word based pitch conversion and MBSC pitch detector. · Voice2: is the converted speech signal using word based pitch conversion and PRAAT pitch detector. · Voice3: is the converted speech signal using syllable based pitch conversion and MBSC pitch detector.	Evaluation of pitch detection using different intonation units and pitch detectors.	38= (10 expert+28 non-expert)
Experiment2	· Voice1: is the converted speech signal after converting only pitch contours. · Voice2: is the converted speech signal after converting pitch and duration. · Voice3: is the converted speech signal after converting pitch, duration, and energy. · Voice4: is the converted speech signal after converting all parameters (pitch, duration, energy, and spectrum).	Evaluation of conversion modules separately and the overall system performance.	17= (7expert+10 non-expert)

6.3 Experimental results and evaluation

Subjective tests were carried out to evaluate our system. The most widely used method for subjective tests is the Mean Opinion Score (MOS) test, where a number of listeners are asked to listen and evaluate the test utterances using a five- point scale and the average of their scores is calculated and is referred as MOS score. This method is one of the subjective tests recommended by IEEE subcommittee and ITU. The minimum number of listeners who participate in the test is 10 if they are expert listeners or 20 in case of non-expert listeners [69].

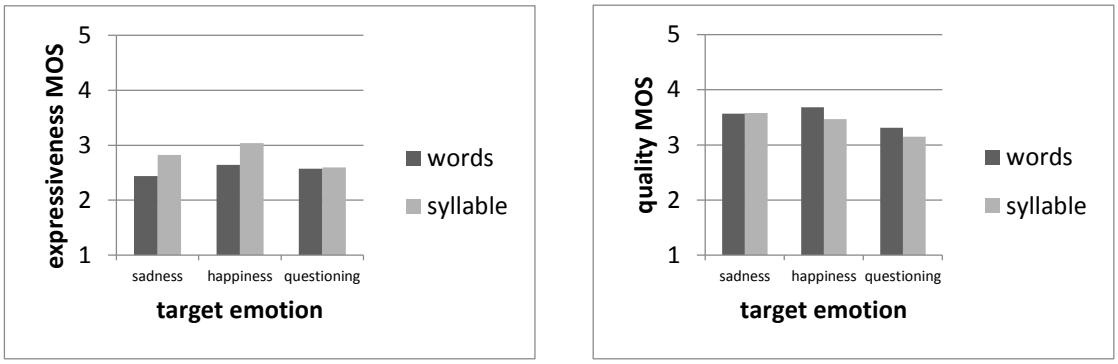
Statistical significance is used in experiments which are performed just on samples of the population to determine if their results can be generalized to the population or no. The p-value, the probability of rejecting the results (hypothesis) of an experiment, is used to indicate statistical significance. In most scientific literature, the results were significant if p-value is less than 0.05 [70]. In our system evaluation we use p-values resulted from t-test to verify MOS score results.

Table 6-3 evaluation sheet

Voice 1			Voice 2			Voice 3		
Expressiveness score			Expressiveness score			Expressiveness score		
sounds exactly like neutral	1		sounds exactly like neutral	1		sounds exactly like neutral	1	
sounds slightly different from neutral	2		sounds slightly different from neutral	2		sounds slightly different from neutral	2	
sounds different from neutral	3		sounds different from neutral	3		sounds different from neutral	3	
sounds more different from neutral	4		sounds more different from neutral	4		sounds more different from neutral	4	
sounds exactly like target	5		sounds exactly like target	5		sounds exactly like target	5	
Speech quality			Speech quality			Speech quality		
Very poor	1		Very poor	1		Very poor	1	
Poor	2		Poor	2		Poor	2	
Good	3		Good	3		Good	3	
Very Good	4		Very Good	4		Very Good	4	
Excellent	5		Excellent	5		Excellent	5	

6.3.1 *Results of using different intonation units for pitch conversion*

The effect of using different intonation units can be examined through a comparison between the evaluation of voice1 and voice3 in experiment 1, where different intonation units were used with the same pitch detector for pitch conversion, we can see that both syllables and words are effective for emotion conversion in Arabic speech as shown in Figure 6-1. The p-values of preferences for word-based and syllable-based pitch conversion are shown in Table 6-4. From the resulting MOS score and p-values, we can conclude that using syllables as the basic intonation unit significantly gives higher expressiveness for sadness and happiness ($p < 0.05$). The difference in quality for the three expressions weren't significant ($p > 0.05$), and the same for expressiveness in questioning ($p > 0.05$).



(a) *(b)*

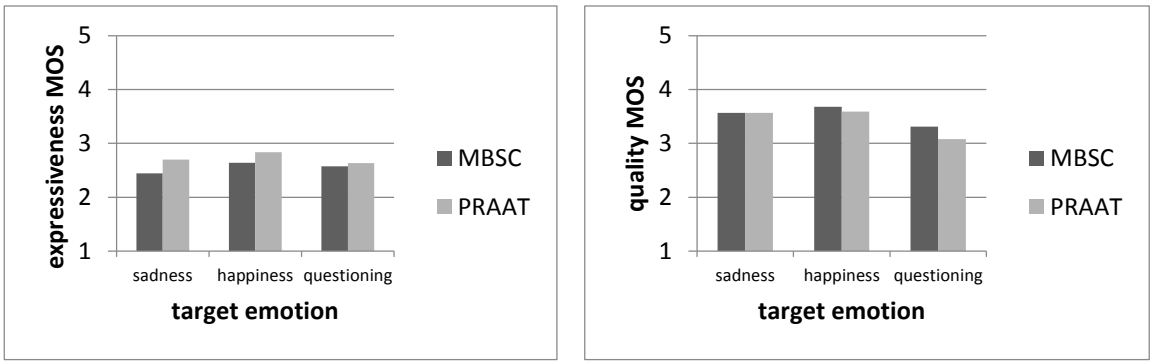
Figure 6-1 MOS of output speech of pitch conversion using two different intonation units (Word-syllable), (a) expressiveness MOS, (b) quality MOS

Table 6-4 the p-values performed on preferences for word-based and syllable-based pitch conversion.

	Word-based, syllable based pitch conversion (expressiveness)	Word-based, syllable based pitch conversion (quality)
Sadness	0.012888	0.45199
Happiness	0.013893	0.089904
questioning	0.443308	0.110547

6.3.2 *Results of using different pitch detectors for pitch conversion*

As in the previous test, to conclude the effect of using different pitch detectors, a comparison between the evaluation of voice1 and voice2 in experiment 1 was carried out. MOS scores of using different pitch detectors are shown in Figure 6-2, and p-values for those different detectors MOS preferences are shown in Table 6-5. We can conclude that MBSC pitch detector significantly gives higher quality than PRAAT for questioning, while no significant difference between using PRAAT and MBSC for happiness and sadness.



(a) *(b)*

Figure 6-2 MOS of output speech of pitch conversion using two different pitch detectors (MBSC-PRAAT), (a) expressiveness MOS, (b) quality MOS

Table 6-5 the p-values performed on preferences for pitch conversion using MBSC & PRAAT.

	pitch conversion using MBSC &PRAAT (expressiveness)	pitch conversion using MBSC & PRAAT (quality)
Sadness	0.063538	0.5
Happiness	0.123029	0.257965
questioning	0.346172	0.040194

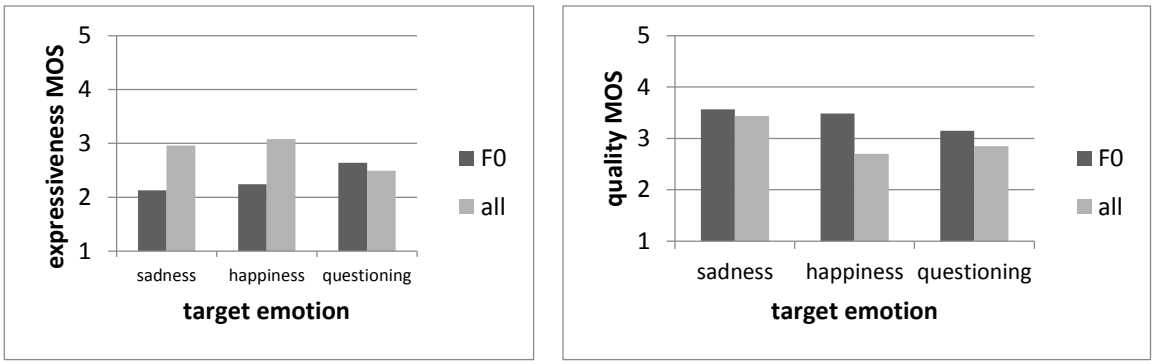
6.3.3 *Evaluation of system modules*

Our system consists of four modules, each of them is used to convert a speech parameter (pitch, duration, energy, and spectrum). In this section we used the results of experiment 2 to evaluate the effect of converting each parameter using our proposed modules. The evaluation of converting each parameter is illustrated in the following subsections and a summary of the effective parameters for different expressions is illustrated in Table 6-10.

6.3.3.1 *Evaluation of pitch conversion*

We evaluate the effect of pitch conversion in our system by comparing the output of converting only pitch contours and converting all speech parameters. This is done by comparing voice1 & voice4 in experiment 2.

We can see that converting the pitch contour, in our proposed system, is the dominant for expressiveness in happiness and questioning as illustrated in Figure 6-3. For questioning, we recommend using only pitch conversion since adding other parameters (duration, and spectrum) add no significant effect on expressiveness (p=0.09) but significantly affects the speech quality ($p \ll 0.05$).



(a) *(b)*

Figure 6-3 MOS of output speech of pitch conversion and all parameter conversion, (a) expressiveness MOS, (b) quality MOS

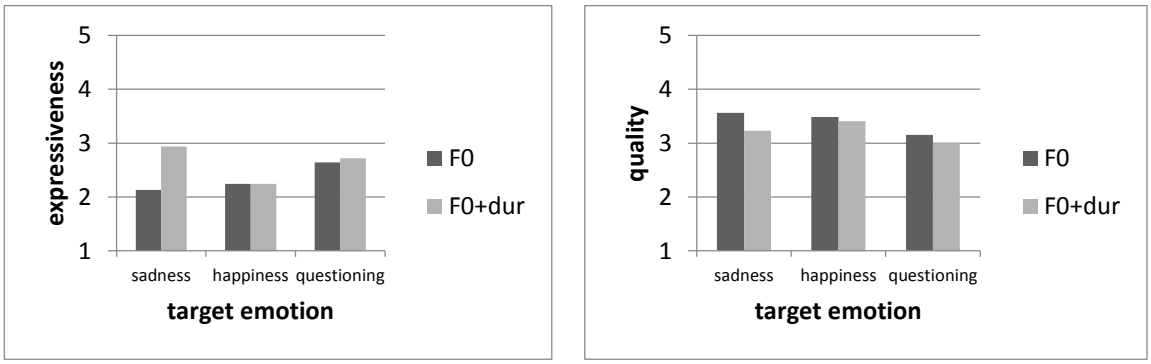
Table 6-6 the p-values performed on preferences for pitch conversion & the overall system conversion.

	Pitch conversion , all parameter conversion	
	expressiveness	quality
Sadness	5.92577E-14	2.68291E-06
Happiness	0.000863	0.000513
questioning	0.093665993	4.20258E-07

6.3.3.2 *Evaluation of duration conversion*

In this section we evaluate the contribution of duration conversion in our system to the expressiveness and quality of the output speech. This is can be done by comparing the output of converting only pitch contours and converting pitch and duration parameters in the three expressions. A comparison between voice1 & voice2 in experiment 2 is shown in Figure 6-4.

We can conclude that converting duration significantly improve the sadness in the converted speech ($p \ll 0.05$), however no significant effect of duration conversion was observed on neither expressiveness nor quality for happiness and questioning ($p > 0.05$).



(a) *(b)*

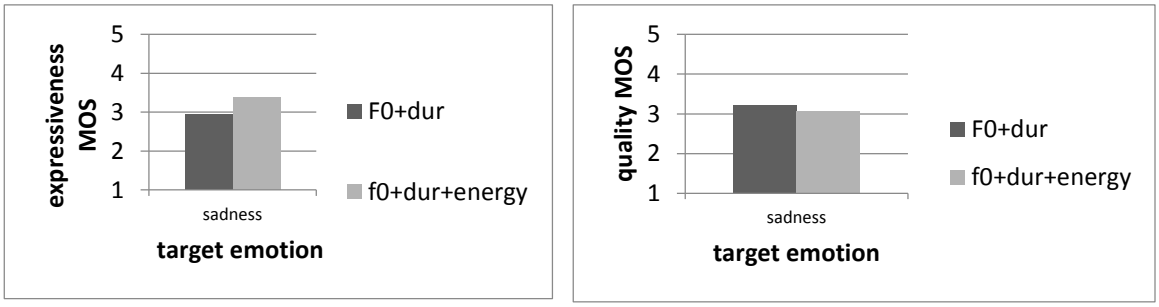
Figure 6-4 MOS of output speech of pitch conversion and pitch &duration conversion, (a) expressiveness MOS, (b) quality MOS

Table 6-7 the p-values performed on preferences for pitch conversion & pitch and duration conversion.

	Pitch conversion , pitch +duration conversion	
	expressiveness	quality
Sadness	1.61577E-08	0.00410488
Happiness	0.095994	0.215926
questioning	0.320704175	0.109373702

6.3.3.3 *Evaluation of energy conversion for sadness*

Energy conversion was proposed only for sadness, since the analysis of training data illustrated that the energy level decreases in sadness. To evaluate decreasing the energy level we compare voice 2 and voice 3 for sadness in experiment 2. Results show that decreasing the energy level has a significant effect on expressiveness ($p = 0.00158$) while the quality didn't significantly affected ($p = 0.10459$).



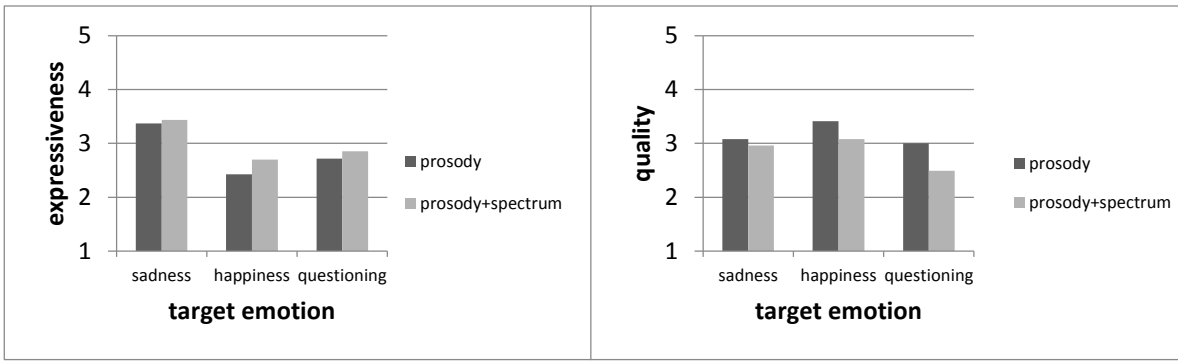
(a) *(b)*

Figure 6-5 MOS of output speech of adding energy conversion to pitch and duration conversion, (a) expressiveness MOS, (b) quality MOS

6.3.3.4 *Evaluation of spectral conversion*

Finally, to evaluate the effect of spectral conversion in our system; voice3 and voice4 are compared. Figure 6-6 shows the MOS scores of converted speech using all prosodic parameters (pitch, duration, and energy) with and without spectral conversion. The p-values for MOS score preferences are shown in Table 6-8.

From the MOS scores and p-values of adding spectral conversion, we can conclude that spectral conversion adds significant expressiveness to happiness only, and the quality of the converted speech with spectral conversion is significantly becomes worse in happiness and questioning. No significant effect was observed on the quality of sadness, this may be because of energy level decrease makes the effect of quality not observable.



(a) *(b)*

Figure 6-6 MOS of the output speech with and without spectral conversion, (a) expressiveness MOS, (b) quality MOS

Table 6-8 the p-values performed on preferences for speech conversion using all parameters with and without spectral conversion

	Prosody conversion , prosody +spectral conversion	
	expressiveness	quality
Sadness	0.3490	0.17900
Happiness	0.0300	0.00342
questioning	0.20021	$5*10^{-5}$

6.3.3.5 *Processing time estimation of different modules*

To estimate the processing time of different modules, we calculate the average ratio between the processing time of the module and the duration of the utterance for all test utterances. Results are shown in Table 6-9.

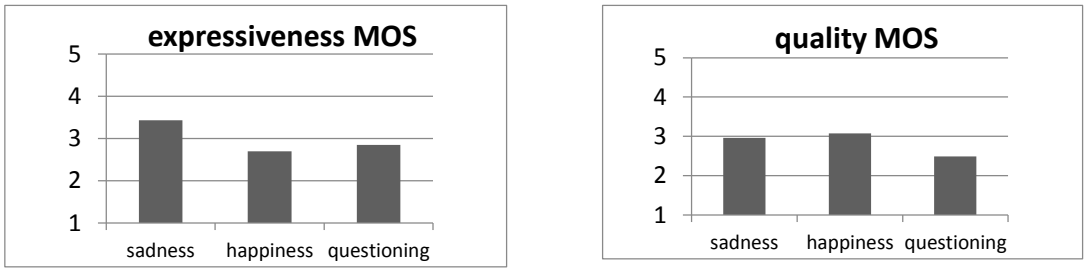
Table 6-9 the average ratio between the processing time of the module and the duration of the utterance.

Spectral conversion	Word-based pitch conversion		Duration conversion
	MBSC	PRAAT	
0.8906	7.79	6.3916	0.0392

6.3.4 *Overall emotion conversion system evaluation*

After transforming the four parameters of speech: pitch, duration, energy, and spectral envelope; we can see that the implemented system managed to add the expressiveness somewhat well with a good quality ($MOS \approx 3$) for sadness and happiness as shown in Figure 6-7. The questioning has the least quality due to the limited size of training data for this expression. It is recommended to change pitch contours only for the questioning, since converting duration and spectral parameters doesn't improve the expressiveness and

degrades the quality. If only pitch contours for questioning are converted, almost the same results of the two other expressions will be obtained. Some of the synthesized speech samples are available at the following link: http://rdi-eg.com/Technologies/Expressive_TTS.htm.



(a) *(b)*

Figure 6-7 MOS of the output speech from the overall system, (a) expressiveness MOS, (b) quality MOS

Table 6-10 summary of effective parameters in our system on different expressions.

	PITCH	DURATION	ENERGY	SPECTRUM
SADNESS	√	√	√	–
HAPPINESS	√	–	–	√
QUESTIONING	√	–	–	–

Chapter7: Conclusion and Future Work

7.1 Thesis contribution

This thesis presents an emotion conversion system for expressive Arabic speech based on linguistic context. The system combines prosody and spectral conversion. Unit selection was used for pitch conversion and the effect of using different intonation units and different pitch detectors was studied. Subjective tests were carried out to evaluate the pitch conversion module with different intonation units and pitch detectors, and to evaluate the effect of converting each speech parameter using our proposed system, and also to evaluate the overall emotion conversion system.

7.1.1 *Unit selection pitch conversion*

Conversion of pitch contours was performed through the selection of expressive units from a database according to a cost function and pitch is modified in the waveform using TD-PSOLA. Words and syllables were used as intonation units and different pitch detectors were checked.

We can conclude that both syllables and words are effective intonation units for emotion conversion in Arabic speech when the right linguistic features are used. Using syllables as the basic intonation unit significantly gives higher expressiveness for sadness and happiness ($p < 0.05$). No significance difference is observed on the quality of the converted speech using the two intonation units.

MBSC pitch detector significantly gives higher quality than PRAAT for questioning, while there isn't significant difference between using the two detectors for sadness and happiness.

7.1.2 *Emotion conversion system*

The proposed emotion conversion system for expressive Arabic speech combines spectral and prosodic parameter transformation. Pitch, duration, energy, and spectral envelope are transformed to the desired expression. Subjective tests were carried out to study the effect of converting each parameter using our system and also the effect of the overall system on both the expressiveness and the quality of the converted speech.

From subjective tests we can conclude that, pitch conversion in our system is the dominant for expressiveness in happiness and questioning (this conclusion also was observed from copy-synthesis experiments). Converting phoneme duration significantly improves the sadness, and makes no significant effect on happiness and questioning. Spectral conversion adds significant expressiveness to happiness.

For questioning, it is proposed to change only pitch contours, since duration conversion module has no effect on questioning and spectral conversion module adds degradation on the output quality without any observable improvement on the expressiveness.

Energy conversion was proposed only for sadness, since the analysis of training data illustrated that the energy level decreases in sadness, and results show that decreasing the energy level has a significant effect on expressiveness in sadness and makes the degradation due to spectral conversion unobservable.

Finally we can say that using a small training corpus, the overall emotion conversion system succeed to achieve an acceptable level of expressiveness with good quality for sadness and happiness, and the quality degradation in questioning can be solved by changing only pitch contours in this expression.

7.2 Future Research

Future work of our research for expressive Text To Speech can include the following proposed points:

- *Change the prosody parameters in the residual:*

To increase the system quality, prosody modifications can be performed on the residual signal of the LSF [29], since spectral modification of the speech signal causes quality degradation, and our proposed system performs spectral transformation and then modifies the pitch, duration and energy of the output of spectral modification.

- *Increase the size of training corpora and determine the minimum size to obtain good emotion conversion system.*
- *Try rule-based emotion conversion to obtain expressive speech and compare the output results with data-driven emotion conversion.*
- *Non-verbal vocalization*

It is thought that adding non-verbal laughs, speech grunts, and other small sounds which indicate a certain expression may increase the expressiveness of the output speech.

References

[1] G. Salvi, F. Tesser, E. Zovato, and P. Cosi, "Cluster analysis of differential spectral envelopes on emotional speech," in *INTERSPEECH*, 2010, pp. 322-325.

[2] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. M. Lee, S. Lee, *et al.*, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *INTERSPEECH*, 2005, pp. 801-804.

[3] Y. Shao, Z. Wang, J. Han, and T. Liu, "Modifying spectral envelope to synthetically adjust voice quality and articulation parameters for emotional speech synthesis," in *Affective Computing and Intelligent Interaction*, ed: Springer, 2005, pp. 334-341.

[4] R. Barra, J. M. Montero, J. Macias-Guarasa, J. Gutiérrez-Arriola, J. Ferreiros, and J. M. Pardo, "On the limitations of voice conversion techniques in emotion identification tasks," in *Proc. Interspeech*, 2007, pp. 2233-2236.

[5] O. Turk and M. Schroder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 965-973, 2010.

[6] I. Saratxaga, E. Navas, I. Hernáez, and I. Luengo, "Designing and recording an emotional speech database for corpus based synthesis in Basque," in *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, 2006, pp. 2126-2129.

[7] T. Kostoulas, T. Ganchev, I. Mporas, and N. Fakotakis, "A Real-World Emotional Speech Corpus for Modern Greek," in *LREC*, 2008.

[8] M. Schröder and M. Grice, "Expressing vocal effort in concatenative synthesis," in *Proc. 15th international conference of phonetic sciences*, 2003, pp. 2589-2592.

[9] A. Iida and N. Campbell, "Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders," *International Journal of Speech Technology*, vol. 6, pp. 379-392, 2003.

[10] W. L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, 2002, pp. 163-166.

[11] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1099-1108, 2006.

[12] I. Steiner, M. Schröder, M. Charfuelan, and A. Klepp, "Symbolic vs. acoustics-based style control for expressive unit selection," in *SSW*, 2010, pp. 114-119.

[13] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *INTERSPEECH*, 2003.

[14] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. ICSLP*, 2004.

[15] J. M .Montero, J. M. Gutierrez-Arriola, S. E. Palazuelos, E. Enriquez, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: from speech database to TTS," in *ICSLP*, 1998, pp. 923-926.

[16] H. M. Meral, H. K. Ekenel, and A. S. Ozsoy, "Role of Intonation Patterns in conveying emotion in Speech," 2002.

[17] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. M. Blanco, *et al.*, "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[18] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.2000 ,

[19] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, pp. 369-390, 1995.

[20] I. R. Murray, M. D. Edgington, D. Campion, and J. Lynn, "Rule-based emotion synthesis using concatenated speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[21] J. E. Cahn, "The Generation of A ect in Synthesized Speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1-19, 1990.

[22] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1145-1154, 2006.

[23] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, pp. 268-283, 2009.

[24] B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," *Acoustical science and technology*, vol. 30, pp. 170-179.2009 ,

[25] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, pp. 134-138, 2012.

[26] F. Tesser, E. Zovato, M. Nicolao, and P. Cosi, "Two vocoder techniques for neutral to emotional timbre conversion," in *SSW*, 2010, pp. 130-135.

[27] O. Türk and M. Schröder, "A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis," in *INTERSPEECH*, 200 ,8 pp. 2282-2285.

[28] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech communication*, vol. 16, pp. 175-205, 1995.

[29] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 972-980, 2006.

[30] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, pp. 1303-1306.

[31] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, pp. 187-207, 1999.

[32] D. Erro, E. Navas, I. Herndez, and I. Saratxaga, "Emotion conversion based on prosodic unit selection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 974-983, 2010.

[33] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing* vol. 18: Prentice Hall Englewood Cliffs, 2001.

[34] T. Sheridan, *Lectures on the Art of Reading*: printed for J. Dodsley [and others], 1792.

[35] P. Taylor, *Text-to-speech synthesis*: Cambridge University Press, 2009.

[36] A. Acero, "Formant analysis and synthesis using hidden Markov models," in *EUROSPEECH*, 1999, pp. 1047-1050.

[37] J .Yamagishi, "An introduction to hmm-based speech synthesis," Technical report, Tokyo Institute of Technology2006.

[38] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *SSW* ,2007 ,pp. 294-299.

[39] T. F. DE MÁSTER, "Design and test of an Expressive Speech Synthesis System based on Speaker Adaptation techniques".

[40] B. P. Nguyen, "Studies on spectral modification in voice transformation," 2009.

[41] S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez, "Real-time incremental speech-to-speech translation of dialogs," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 437-445.

[42] P. Gebhard, M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, *et al.*, "IDEAS4Games: building expressive virtual characters for computer games," in *Intelligent Virtual Agents*, 2008, pp. 426-440.

[43] W. M. Azmy ,S. Abdou, and M. Shoman, "Arabic Unit Selection Emotional Speech Synthesis using Blending Data Approach," *International Journal of Computer Applications*, vol. 81, pp. 22-28, 2013.

[44] R. Fernandez and B. Ramabhadran, "Automatic Exploration of Corpus-Specific Properties for Expressive Text-to-Speech: A Case Study in Emphasis," in *6th ISCA Workshop on Speech Synthesis*, 2007.

[45] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach*: CRC Press, 2003.

[46] S. V. Vaseghi, *Multimedia signal processing: Theory and applications in speech, music and communications*: John Wiley & Sons, 2007.

[47] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*: Springer, 2008.

[48] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication,* vol. 9, pp. 453-467, 1990.

[49] P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, pp. 1-23, 2008.

[50] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems,* vol. 7, pp. 358-386, 2005.

[51] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping to massive datasets," in *Principles of Data Mining and Knowledge Discovery*, ed: Springer, 1999, pp. 1-11.

[52] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech Communication,* vol. 55, pp.2013 ,856-841 .

[53] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis,* vol. 495, p. 518, 1995.

[54] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, 1993, pp. 97-110.

[55] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America,* vol. 111, pp. 1917-1930.2002 ,

[56] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, pp. I-333-I-336.

[57] A. Camacho and J. G. Harris" ,A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America,* vol. 124, pp. 1638-1652, 2008.

[58] B. P. Nguyen and M. Akagi, "Phoneme-based spectral voice conversion using temporal decomposition and Gaussian mixture model," in *Communications and Electronics, 2008. ICCE 2008. Second International Conference on*, 2008, pp. 224-229.

[59] Z. Hanzlíček and J. Matoušek, "F0 transformation within the voice conversion framework," 2007.

[60] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, pp. 349-353, 2006.

[61] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, 1983, pp. 81-84.

[62] P. C. Nguyen, O. Takao, and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," *IEICE TRANSACTIONS on Information and Systems*, vol. 86, pp. 397-405, 2003.

[63] T. Shibata and M. Akagi, "A study on voice conversion method for synthesizing stimuli to perform gender perception experiments of speech," in *Proceedings of the RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP*, 2008, pp. 180-183.

[64] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, *et al.*, "The HTK book (for HTK version 3.4)," *Cambridge university engineering department*, vol. 2, pp. 2-3, 2006.

[65] Z. Inanoglu, "Data Driven Parameter Generation For Emotional Speech Synthesis," PHD, Department of Engineering, University of Cambridge, St Edmund's College, 2008.

[66] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341-345, 2002.

[67] D. Lolive, N. Barbot, and O. Boeffard, "Pitch and duration transformation with non-parallel data," *Speech Prosody 2008*, pp. 111-114, 2008.

[68] D. Govind, S. M. Prasanna, and B. Yegnanarayana, "Neutral to Target Emotion Conversion Using Source and Suprasegmental Information," in *Interspeech*, 2011, pp. 2969-2972.

الملخص

تغيير العاطفة باستخدام مجموعات صوتية صغيرة هام جداً للحصول على نطق معبر للنص المكتوب. لقد تم تطبيق نموذج إختيار الوحدات لتغيير نغمة الصوت على نطاق واسع. لقد إستخدمت وحدات تنغيم مختلفة اعتمادًا على الخصائص اللغوية للغات المختلفة.

تقدم هذه الرسالة نظام لتغيير العاطفة للحصول على نطق عربي معبر. هذا النظام يجمع تحويل الطيف الصوتي والعوامل الخاصة بعلم العروض بناءً على الخواص اللغوية. يتم تحويل أربع معاملات خاصة بالصوت للحصول على التعبير المطلوب وهم : التردد الأساسي، مدة الكلام، شدة الصوت، وأخيرًا الطيف الصوتي. ويتم دراسة خصائص الكلام العربي المسؤولة عن تنغيم الكلمات بطرق مختلفة. يتم استخدام إختيار الوحدات لتغيير نغمة الصوت ودراسة تأثير إستخدام وحدات تنغيم مختلفة وإستخراج النغمة الأساسية للكلام بطرق مختلفة. كما يتم دراسة تأثير تحويل كل عامل من عوامل الصوت الأربعة – بإستخدام نظام تغيير العاطفة الذي قمنا بإقتراحه – على التعبير. وأخيرًا تم تقييم النظام الكامل لتغيير العاطفة على تعبيرات مختلفة.

تم عمل اختبارات من خلال تقييم الأشخاص لتقييم النظام على ثلاثة تعبيرات مختلفة: الحزن، السعادة، والإستفهام. لقد أظهرت النتائج أن استخدام المقاطع والكلمات كوحدات تنغيم أساسية لتغيير التردد الأساسي مؤثر بالرغم من أن المقاطع تعطي تعبير أعلى في حالة الحزن والسعادة. وأيضًا أظهرت النتائج أن تغيير النغمة الأساسية باستخدام نظامنا المقترح هو العامل الرئيسي في حالة السعادة والإستفهام وله تأثير عالي في حالة الحزن، بينما تغيير مدة الكلام يؤثر فقط على الحزن وتغيير الطيف الصوتي يؤثر فقط على السعادة.

لقد إقترحنا تقليل طاقة الصوت في حالة الحزن وذلك بعد تحليل مجموعات الصوت المستخدمة لتدريب النظام. وقد أظهرت النتائج أن تقليل طاقة الصوت مؤثر في حالة الحزن. أخيرًا، أظهر تقييم النظام الكامل لتغيير العاطفة أنه نجح على إضافة التعبير للكلام العربي بشكل مقبول مع جودة عالية للصوت ($MOS \approx 3$) في حالة الحزن والسعادة. يمكن الحصول على نفس النتائج في حالة الإستفهام إذا تم تحويل التردد الأساسي فقط، حيث أن تحويل الطيف الصوتي المستخدم يسبب قلة جودة الصوت الناتج بدون إضافة تعبير له وأيضًا تغيير مدة الكلام غير مؤثر في حالة الإستفهام.

مهندس: دعاء جمال مدني طايع
تاريخ الميلاد: 1987/11/16
الجنسية: مصري
تاريخ التسجيل: 2010/10/1
تاريخ المنح:

القسم: هندسة الإلكترونيات والإتصالات الكهربية
الدرجة: ماجستير
المشرفون: أ.د. محسن عبد الرازق رشوان
أ.م. حسام علي حسن فهمي
الممتحنون:

أ.م. خيري عبد النبي البربري (الممتحن الخارجي)
أ.د. شريف مهدي عبده (الممتحن الداخلي)
أ.د. محسن عبد الرازق رشوان (المشرف الرئيسي)
أ.م. حسام علي حسن فهمي (عضو)

عنوان الرسالة:

نحو تحويل النص العربي إلى خطاب معبر.

الكلمات الدالة:

تركيب كلام معبر؛ تغيير العاطفة؛ تغيير النبرة ؛ اختيار الوحدات.

ملخص الرسالة:

في هذه الرسالة، يتم اقتراح نظام لتغيير العاطفة للحصول على نص عربي معبر؛ هذا النظام يجمع تحويل الطيف الصوتي والعوامل الخاصة بعلم العروض. لقد تم تحويل أربع معاملات خاصة بالصوت للحصول على التعبير المطلوب وهم : التردد الأساسي، مدة الكلام، شدة الصوت ،و أخيرًا الطيف الصوتي. يتم دراسة تأثير كل عامل من العوامل الصوتية في نظامنا المقترح وتقييم أداء النظام الكامل لتغيير العاطفة. عند تغيير النغمة الأساسية يتم تقييم إستخدام وحدات تنغيم مختلفة (كلمات– مقاطع) وإستخراج النغمة الأساسية للكلام بطرق مختلفة.

نحو تحويل النص العربي إلى خطاب معبر

إعداد

دعاء جمال مدني طايع

رسالة مقدمة إلى كلية الهندسة – جامعة القاهرة

كجزء من متطلبات الحصول على درجة الماجستير

في

هندسة الإلكترونيات والإتصالات الكهربية

يعتمد من لجنة الممتحنين:

الأستاذ المساعد: خيري عبد النبي البربري	الممتحن الخارجي (جامعة قناة السويس)
الأستاذ الدكتور: شريف مهدي عبده	الممتحن الداخلي
الأستاذ الدكتور: محسن عبد الرازق رشوان	المشرف الرئيسي
الأستاذ المساعد: حسام علي حسن فهمي	عضو

كلية الهندسة – جامعة القاهرة

الجيزة – جمهورية مصر العربية

سنة 2014

نحو تحويل النص العربي إلى خطاب معبر

إعداد

دعاء جمال مدني طايع

رسالة مقدمة إلى كلية الهندسة – جامعة القاهرة

كجزء من متطلبات الحصول على درجة الماجستير

في

هندسة الإلكترونيات والإتصالات الكهربية

تحت إشراف

أ.د. محسن عبد الرازق رشوان

.....

أستاذ دكتور

قسم الإلكترونيات والإتصالات الكهربية

كلية الهندسة – جامعة القاهرة

أ.م. حسام علي حسن فهمي

.....

أستاذ مساعد

قسم الإلكترونيات والإتصالات الكهربية

كلية الهندسة – جامعة القاهرة

كلية الهندسة – جامعة القاهرة

الجيزة – جمهورية مصر العربية

سنة 2014

نحو تحويل النص العربي إلى خطاب معبر

إعداد

دعاء جمال مدني طايع

رسالة مقدمة إلى كلية الهندسة – جامعة القاهرة

كجزء من متطلبات الحصول على درجة الماجستير

في

هندسة الإلكترونيات والإتصالات الكهربية

كلية الهندسة – جامعة القاهرة

الجيزة – جمهورية مصر العربية

سنة 2014