



Vrije Universiteit Brussel

FACULTY OF APPLIED SCIENCES

COMMUNICATION OF EMOTIONS FOR E-CREATURES

Selma YILMAZYILDIZ

Academic Year: 2005- 2006

**Promoter: Prof. Dr. ir. Werner
Verhelst**

**Thesis submitted in partial fulfilment
of the requirements for the Master's
degree of Applied Computer Science.**

I would like to thank my promoter Prof. Dr. ir. Werner VERHELST, Mr. Ivan Hermans from Anty Foundation and all the other researchers in VUB ETRO-DSSP for their valuable contribution and support throughout this project.

Abstract

The goal of this project is to produce affective speech that can convey natural vocal emotions in *a non-existing and non-understandable language*.

The system uses a database with natural expressive speech samples from a professional speaker and a database with naturally spoken neutral speech samples from the same speaker. To produce nonsense utterance with a given desired emotion, the synthesizer randomly selects an expressive speech sample of the proper type from the expressive database and uses this as a prosodic template. Next, a nonsense carrier phrase is constructed that has the same segmental structure as the selected prosodic template. This is done by concatenating segments from the neutral database. Then, the same pitch and timing structure as found in the prosodic template is copied on the nonsense carrier phrase through prosody transplantation process which uses *PSOLA (Pitch Synchronous Overlap - Add)* method. Finally, voice quality structure of the template emotional utterance is applied to the artificially generated speech, which results as the final *synthesized emotional speech*.

Declaration

Here by I declare that this thesis was composed by me, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Selma Yilmazyildiz

Table of Contents

1. INTRODUCTION	7
1.1 OVERVIEW	8
1.2 ABOUT ANTY PROJECT	9
1.3 INTRODUCTION TO EMOTIONAL SPEECH SYNTHESIS	10
2. BACKGROUNDS	12
2.1 INTRODUCTION TO SPEECH SYNTHESIS	12
2.2 EMOTION IN SPEECH	14
2.3 EMOTIONAL SPEECH SYNTHESIS	16
2.4 OVERVIEW EMOTIONAL SPEECH SYNTHESIZERS	17
2.5 PSOLA METHOD	19
2.6 PROSODY TRANSPLANTATION	22
3. BUILDING EMOTIONAL VOICE	25
3.1 GENERAL SYSTEM OVERVIEW	25
3.2 DEVELOPING THE ACTUAL SYSTEM	26
3.2.1 <i>Speaker Selection</i>	26
3.2.2 <i>Emotion Selection</i>	27
3.2.3 <i>Database Collection</i>	28
3.2.3.1 <i>Data Collection Procedure</i>	28
3.2.3.2 <i>Recording Procedure</i>	29
3.2.3.3 <i>Evaluation Procedure and Final Database Construction</i>	30
3.2.4 <i>Producing Nonsense Language</i>	32
3.2.4.1 <i>Segmenting the Data</i>	33
3.2.4.2 <i>Generating Carrier Phrases by Segment Concatenation</i>	35
3.2.5 <i>Adding Prosody and Generating the Desired Emotions</i>	38
3.3 APPLICATIONS OF EMOTIONAL SPEECH SYNTHESIS	42
3.4 EVALUATIONS	45
3.4.1 <i>Method</i>	45
3.4.2 <i>Results</i>	46
4. CONCLUDING DISCUSSION	53
5. FUTURE WORKS	56
APPENDIX A - INITIAL DATA FOR THE DATABASES	57
APPENDIX B - UTTERANCE EVALUATION	60
APPENDIX C - FINAL DATABASES	62
REFERENCES	63

1
1

Chapter 1

1. Introduction

One of the most profound questions of engineering, arguably, is whether we will ever create human level consciousness in a machine. The best reason for believing that robots might some day become conscious is that we human beings are conscious, and we are a sort of robot ourselves. That is, we are extraordinarily complex self-controlling, self-sustaining physical mechanisms, designed over the eons by natural selection, and operating according to the same well-understood principles that govern all the other physical processes in living things: digestive and metabolic processes, self-repair and reproductive processes, for instance. It may be wildly over-ambitious to suppose that human artificers can repeat Nature's triumph, with variations in material, form, and design process, but this is not a deep objection. It is not as if a conscious machine contradicted any fundamental laws of nature, the way a perpetual motion machine does. Still, many sceptics believe--or in any event want to believe--that it will never be done. [01] Or at least today it sounds absurd.

“Only those who attempt the absurd...will achieve the impossible.” – M.C Escher

It is a continuous journey to reach a certain level of consciousness and emotional behaviour for artificially created creatures. The e-creatures we have today don't have feelings, but some of them can fake them amazingly well. This fake makes their interaction with human more embraceable. We don't yet have the organic tissue to make them look like flesh but we have the necessary technology to make them sound like real human. Human like speech requires naturalness, naturalness requires emotions.

This paper will focus on constructing a synthesizer that can convey natural vocal emotions in a non-existing and non-understandable language that is created for the actual system. To detail this study, in the first chapter introduction to the project and an overview is presented. The second chapter summarizes the past works in the field creating the fundamental background for this study. In chapter 3 the methodology and the system developed is discussed in details with results achieved. Also the potential implementation areas for the technology are presented in this chapter. Chapter 4 summarizes the conclusions while chapter 5 focuses on the future work to be performed.

1.1 Overview

When people speak, they naturally use acoustic effects in their voice to convey certain emotions, which are very important for signalling their feelings and their emotional state. The communication of emotions through vocalizations appears to be language independent to a large extent. Indeed, one can easily recognize these emotions even in a language which he doesn't understand. This makes the emotions being passed as a part of a speech one of the core components in human communication.ⁱ

This thesis concerns the process of building an emotional speech synthesizer through the *prosodic transplantation* [02], [03] of one recorded emotional utterance on to an artificially produced neutral carrier utterance. Transplantation is done by synchronizing the timing and the pitch patterns of the carrier utterance with the reference emotional utterance by using PSOLA (Pitch Synchronous Overlap-Add) method [04]. Also the voice quality of the template emotional utterance is applied to the PSOLA synthesized speech to generate the final synthesized emotional speech.

As described in more details in Section 3.3 - Applications of Emotional Speech Synthesis, one of the main application areas of emotional speech synthesis is e-creatures.

ⁱ <http://emotion-research.net>

When the right emotions are applied to the content, the communication of the e-creature becomes more natural and more believable. Especially for the e-creatures which communicate with children or with people at an emotionally sensitive state, the application of emotions is essential for success. The potential future implementation of the techniques developed in this study for *ANTY – An intelligent, autonomous huggy robot*, forms the main motivation of this thesis.



Figure 1 - ANTY

1.2 About ANTY project

Staying in a hospital is often a traumatic experience for children. They are separated from their parents and their friends for a while. They enter an unfamiliar environment with a lot of scary devices. The illness combined with pain makes their lives extra difficult. Anty wants to let these children forget their problems for a while. Anty is a non-profit organization. Founders of Anty foundation have aspired to realize a dream with the purpose of creating an intelligent huggable robot that is able to play with hospitalized children, thus making the hospital a livelier place. Anty uses some of today's most advanced technology to build one of the most up-to-date robots of its time. Anty also has a generous team of doctors and engineers working on it everyday. At this moment Anty Foundation is building its first prototype.ⁱⁱ

As Anty will be communicating primarily with hospitalized children, taking into account the sensitive emotional state of these children, it is essential to apply emotions to Anty's

ⁱⁱ <http://www.anty.org>

speech communication. The techniques developed in this study will potentially be implemented to this intelligent, autonomous huggy robot, making it a better experience for anybody communicates verbally with Anty.

1.3 Introduction to Emotional Speech Synthesis

Emotional speech synthesis can be described as; applying human like emotional effects to artificially produced speech. To understand the emotional speech synthesis, first “*Speech Synthesis*” and the “*Emotions in Speech*” should be well understood. In the upcoming chapter these topics will be discussed in detail.

The two characteristics that are used to describe the quality of synthesised speech are *naturalness* and *intelligibility*. The naturalness refers to how much the artificial output sounds like the speech of a real person.ⁱⁱⁱ Successful application of emotions to synthesised speech would result an increase in naturalness. Once the right content is delivered with the right emotions, the synthesised speech becomes more believable. This is the main goal of the emotional speech synthesis.

Considering that Anty would primarily communicate to children, having natural and believable synthesised output is essential for its success. This requirement is the main driver of this study.

There are many methods used in emotional speech synthesis. During this study the PSOLA (Pitch Synchronous Overlap and Add) method has been used. Also this method will be discussed in detail in the next chapter.

ⁱⁱⁱ http://en.wikipedia.org/wiki/Speech_synthesis#Synthesizer_technologies

2₂

Chapter 2

2. Backgrounds

2.1 Introduction to Speech Synthesis

“Speech is the mirror of the soul; as a man speaks, so is he” said Publilius Syrus a Roman author who lived in the 1st century B.C. Wikipedia describes speech as “The physical act of speaking, primarily through the use of vocal cords to produce voice”. Whether you prefer the way the Roman author defines speech or Wikipedia’s definition, speech has always been the primary way of communication of the modern human.

The easiest description of speech synthesis is the artificial production of speech. Even though the speech synthesis sounds like a recent concept, the desire of man to synthesise speech has started over two hundred years ago. In St. Petersburg, in 1779, Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made apparatus to produce them artificially. The research and experiments with mechanical and semi-electrical analogs of vocal system were made until 1960's, but with no remarkable success. [05]

The first ever electronic speech synthesis was done by the "*VODER (Voice Operating Demonstrator)*" of Homer Dudley in 1939 and it was demonstrated at the Worlds Fair. VODER was a humanly controlled version of the speech synthesizer. [06] [<http://www.cs.indiana.edu/rhythmsp/ASA/AUfiles/01.AU>]. This was a milestone in the history of speech

synthesis which convinced the scientific world that intelligible speech could be synthesised artificially.

PAT (Parametric Artificial Talker) and OVE (Orator Verbis Electris) followed VODER. In 1968 the first full text-to-speech system for English was developed in Japan. In late 1970's and early 1980's, many commercial text-to-speech and speech synthesis products were introduced. [05] Early electronic speech synthesizers sounded very robotic but the focus of the scientists and technology providers on speech synthesis never decreased. The quality of synthesized speech has steadily improved. Output from some speech synthesis systems became indistinguishable from actual human speech. By late 1990's text to speech engines became a standard add-in in personal computer operating systems, even a *Speech Synthesis Mark-up Language (SSML)* has been introduced by World Wide Web Consortium in 2000 to develop standards to enable access to the web using spoken interaction.^{iv}

There have been many significant advances in the technologies used for speech synthesis. Today with a click of a button one can reach tens of speech synthesis programs available on the internet.

The two characteristics that are used to describe the quality of a speech synthesis system are naturalness and intelligibility. The naturalness of a speech synthesizer refers to how much the output sounds like the speech of a real person. How easily the output can be understood is referred as the intelligibility of a speech synthesizer. The ideal speech synthesizer is both natural and intelligible, and each of the different synthesis technologies tries to maximize both of these characteristics.^v

Concatenative synthesis and formant synthesis are the two main technologies used for speech synthesis. The basis of the concatenative synthesis is the concatenation of

^{iv} <http://www.w3.org/TR/2000/WD-speech-synthesis-20000808>

^v http://en.wikipedia.org/wiki/Speech_synthesis#Synthesizer_technologies

segments of pre-recorded speech. In formant synthesis, the synthesized output is created using an acoustic model. It doesn't use any pre-recorded human speech samples.

Generally, concatenative synthesis gives natural sounding synthesized speech where as formant synthesized speech is usually reliably intelligible.

But isn't there something more in a human speech than linguistic information, something that a human naturally recognises in a speech, something one can feel. Like the identity, age, geographical origin, attitude, and emotional state of the speaker.

2.2 Emotion in Speech

Emotion is a complex word which has no single universally accepted definition. Emotions are mental states that arise spontaneously, rather than through conscious effort.

According to some researchers emotion is, to a high degree, dependent on social phenomena, expectations, norms, and conditioned behaviour of the group in which an individual lives. The influence of politics, religion, and socio-cultural customs can be sometimes traced or hypothesized.^{vi}

Contrary to this view, a study by Ekman P, Friesen W, and Ellsworth P in 1972 has shown that at least some facial expressions and their corresponding emotions are universal. Paul Ekman and his team categorized these basic universal emotions as anger, disgust, fear, joy, sadness and surprise. [07]

This study focuses on 4 of these basic emotions: Anger, fear, joy and sadness which provide the emotional basis of communication with children.

^{vi} <http://en.wikipedia.org/wiki/Emotions>

Speech contains acoustic features that vary with the speaker's emotional state. The effects of emotion in speech tend to alter pitch, timing, voice quality and articulation of the speech signal. [08] Many researchers suggest that grouping emotions based on high and low degrees of activation improve the performance of emotion recognition and synthesis. The high-activation emotions such as anger and joy have similar characteristics of greater loudness, higher pitch, and faster speed than low-activation emotions such as sadness. [09] The characteristics of each 4 emotions focused in this study are detailed below.

Anger: In a speech, emotional state of anger causes increased intensity with dynamic changes. The voice is breathy and has tense articulation with abrupt changes. The average pitch pattern is higher and there is a strong downward inflection at the end of the sentence. The pitch range and its variations are also wider than in normal speech and the average speech rate is also a little bit faster.

Happiness: Slightly increased intensity and articulation for content words are the characteristics of happiness in a speech. The voice is breathy and light without tension. Happiness also leads to increase in pitch and pitch range. In happiness, the peak values of pitch and the speech rate are the highest of basic emotions.

Fear: The intensity of speech is lower with no dynamic changes for fear. The average pitch and pitch range are slightly higher than in neutral speech. Articulation is precise and the voice is irregular. Energy at lower frequencies is reduced. The speech rate is slightly faster than in normal speech and contains pauses between words forming almost one third of the total speaking time.

Sadness: In speech sadness decreases the speech intensity and its dynamic changes. The average pitch is at the same level as in neutral speech, but there are almost no dynamic changes. The articulation precision and the speech rate are also decreased. High ratio of pauses to phonation time also occurs. [05]

During the recording sessions of this study, the basic characteristics of each 4 emotions focused are provided as input to the speaker. This will be discussed in more detail in Chapter 3.

2.3 Emotional Speech Synthesis

As mentioned earlier the main goal of speech synthesis is to produce synthesised speech which is both natural and intelligible. Modern speech synthesis systems are highly intelligible. However, most synthetic speech is easily identified as being "machine generated" as it sounds unnatural.

There are three main contributors to the naturalness of synthetic speech: intonation, voice quality and variability.

Intonation: One of the main factors in intelligibility of speech has been identified as the intonation contour of the utterance, and in particular the placement of word- and utterance-level accents; hence this is also of major importance for naturalness, as an incorrect intonation contour immediately suggests an unnatural voice.

Voice quality: The underlying "personality" of the voice is also a major contributor to naturalness. Systems based on recorded speech perform well in this respect, as the voice quality from the speaker comes through in the re-synthesized speech.

Variability: There are various variability elements in speech, but these are not included in most of the current synthetic speech systems - all speech is "neutral". However, if we wish synthetic speech to be natural, we must include some variability into the speech output. [10]

The acoustic effects caused by the feelings and the emotional state of a speaker are a part of a natural speech and offer the first opportunity to add variability to synthetic speech. A

successful implementation of the vocal emotion factors, such as alter in pitch, in timing, in voice quality and in articulation, to synthesised speech significantly improves the naturalness.[08] When the right emotions are applied to the content, the synthesised speech becomes more natural and more believable.

2.4 Overview Emotional Speech Synthesizers

The modelling of emotion in speech relies on a number of parameters like, level, pitch, timing, voice quality and articulation. Different synthesis techniques provide control over these parameters to very different degrees. [11] There is no unified categorization of emotional speech synthesizers. The most well known and most commonly used techniques are; Formant synthesis, diphone concatenation, unit selection and PSOLA method.

Formant Synthesis: Creates the acoustic speech data entirely through rules on the acoustic correlates of the various speech sounds. No human speech recordings are involved at run time. The synthesised speech sounds relatively unnatural and “robot-like”. But a large number of parameters can be varied quite freely, which is interesting for modelling emotional expressivity in speech.[11] Several researches have used Formant synthesis because of the high degree of control that it provides. One of the most well known examples of formant synthesizers is HAMLET (the Helpful Automatic Machine for Language and Emotional Talk) of Murray and Arnott. The HAMLET system is based around a series of rules which systematically alter the voice of the synthesiser in ways appropriate to the emotion being simulated. The parameters controlled by the system were the underlying voice quality of the synthesiser used (at the utterance level), together with the pitch and timing of individual phonemes within the utterance to maintain detailed control over the intonation contour. These alterations are applied above the basic intonation contour of the utterance (to retain accent information). [10]

Diphone Concatenation: In concatenative synthesis, recordings of a human speaker are concatenated in order to generate the synthetic speech. The use of diphones is common. Diphone recordings are usually carried out with a monotonous pitch. At synthesis time, the required F0 contour is generated through signal processing techniques. This introduces a certain amount of distortion, but the resulting speech quality is usually more natural than formant synthesis. In most diphone synthesis systems, only F0 and duration (and possibly intensity) can be controlled. In particular, it is usually impossible to control voice quality. To use diphone synthesis for emotion synthesis, the question to be answered is whether F0 and duration are sufficient to express emotions and whether voice quality is indispensable for emotion expression or not. Interestingly, very different results were obtained by different studies. [11]

Unit selection: This synthesis technique is usually accepted as being most natural. It is also referred as large database synthesis or speech re-sequencing synthesis. In this technique instead of a minimum speech data inventory as in diphone synthesis, a large inventory is used. Out of this large database, units of variable size are selected which best approximate a desired target utterance defined by a number of parameters. The same parameters as used in diphone synthesis, i.e. phoneme string, duration and F0 can be used. The weights assigned to the selection parameters influence which units are selected. If well-matching units are found in the database, no signal processing is necessary. While this synthesis method often gives very natural results, the results can be very bad when no appropriate units are found. [11]

PSOLA Method: The PSOLA (Pitch Synchronous Overlap Add) method was developed at France Telecom (CNET- Centre National d'Etudes des Télécommunications). Even though it was not aimed to be a synthesis method, it allows pre-recorded speech samples smoothly being concatenated. The good level of pitch and duration control it provides makes it a good technique for synthesis systems. There are several versions of the PSOLA algorithm. Time-domain version, TD-PSOLA, is the most commonly used due to its computational efficiency. [05] During this study TD-PSOLA has been used as the

underlying method. PSOLA method will be discussed in more detail in the next section of this chapter.

2.5 PSOLA Method

The word ‘*prosody*’ comes from ancient Greek, where it was used for a “*song sung with instrumental music*”. In later times the word was used for the “*science of versification*” and the “*laws of metre*”, governing the modulation of the human voice in reading poetry aloud. In modern phonetics, the word ‘prosody’ and its adjectival form ‘prosodic’ are most often used to refer to the properties of speech that cannot be derived from the segmental sequence of phonemes underlying human utterances. Examples of such properties are the controlled modulation of the voice pitch, the stretching and shrinking of segment and syllable durations, and the intentional fluctuations of overall loudness. [12]

In other words the term prosody refers to certain properties of the speech signal which are related to audible changes in *pitch, loudness and syllable length*. PSOLA (Pitch Synchronous Overlap Add) is a well-known technique for prosodic modification of speech signals, especially for pitch shifting and time scaling.

The basic PSOLA algorithm consists of three steps; analysis-modification-synthesis. The *PSOLA analysis-modification-synthesis method* belongs to the general class of *STFT (Short-Time Fourier Transform)* analysis-synthesis method. [13]

Analysis:

The analysis process consists of decomposing the speech waveform $x(n)$ into a stream of short-time analysis signals, $x(t_a(u), n)$. These short-time signals are obtained from the digital speech waveform $x(n)$ by a sequence of time-translated analysis windows.

$$x(t_a(u), n) = h_u(n)x(n - t_a(u))$$

$t_a(u)$ are referred as analysis time instants or more generally as pitch marks. They are set at a pitch synchronous rate on the voiced portions of speech and at a constant rate on the unvoiced portions. The analysis window is generally chosen to be a symmetrical Hanning window. The window length is chosen to be proportional to the local pitch period $P(s)$, for example $T = \mu P(s)$. The proportionality factor μ ranges from 2, for the standard time-domain PSOLA method, to $\mu = 4$, for the frequency-domain implementation. [13]

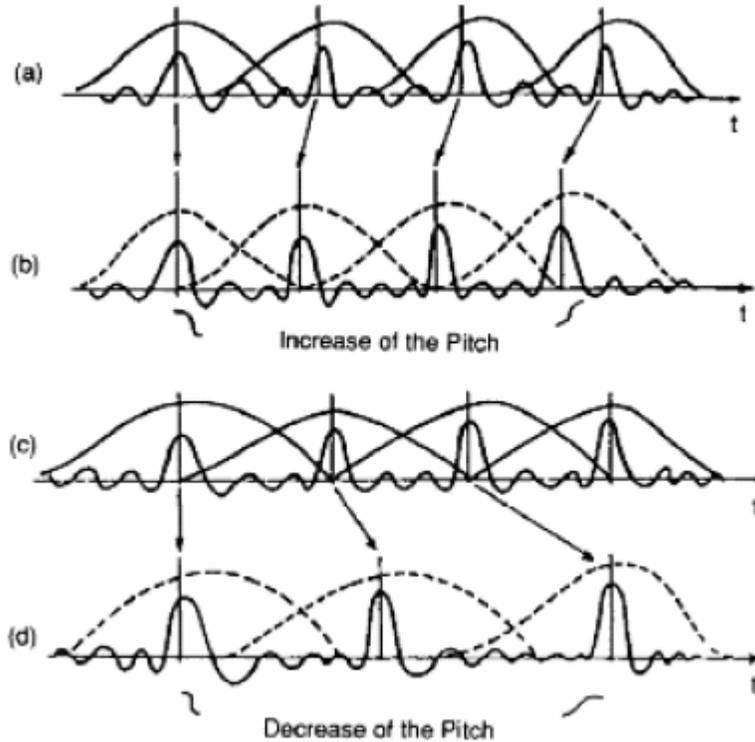


Figure 2 - Pitch Modification of a Voiced Speech Segment

Modification:

Simply put, modification process is the modification of each analysis signal to synthesis signal. Transformation consists of transforming the stream of short-time analysis signals into a stream of short-time synthesis signals, synchronized on a new set of synthesis time instants $t_s(u)$. According to the desired pitch-scale and time-scale modification, the synthesis time instants $t_s(u)$ are determined from the analysis time instants $t_a(s)$. [13]

Synthesis:

In the final step, the synthetic signal $y(n)$ is obtained by combining the synthesis waveforms synchronized on the stream of synthesis time instants $t_s(u)$. Least-square overlap-add synthesis procedure may be used for this purpose. In the TD-PSOLA algorithm, the synthesis window $f_u(n)$ is equal to the analysis window associated with the analysis time instant $t_a(s)$ mapped with the synthesis time instant $t_s(u)$. [13]

$$y(n) = \frac{\sum_k h(n - uR)x(uR.n)}{\sum_k h^2(n - uR)}$$

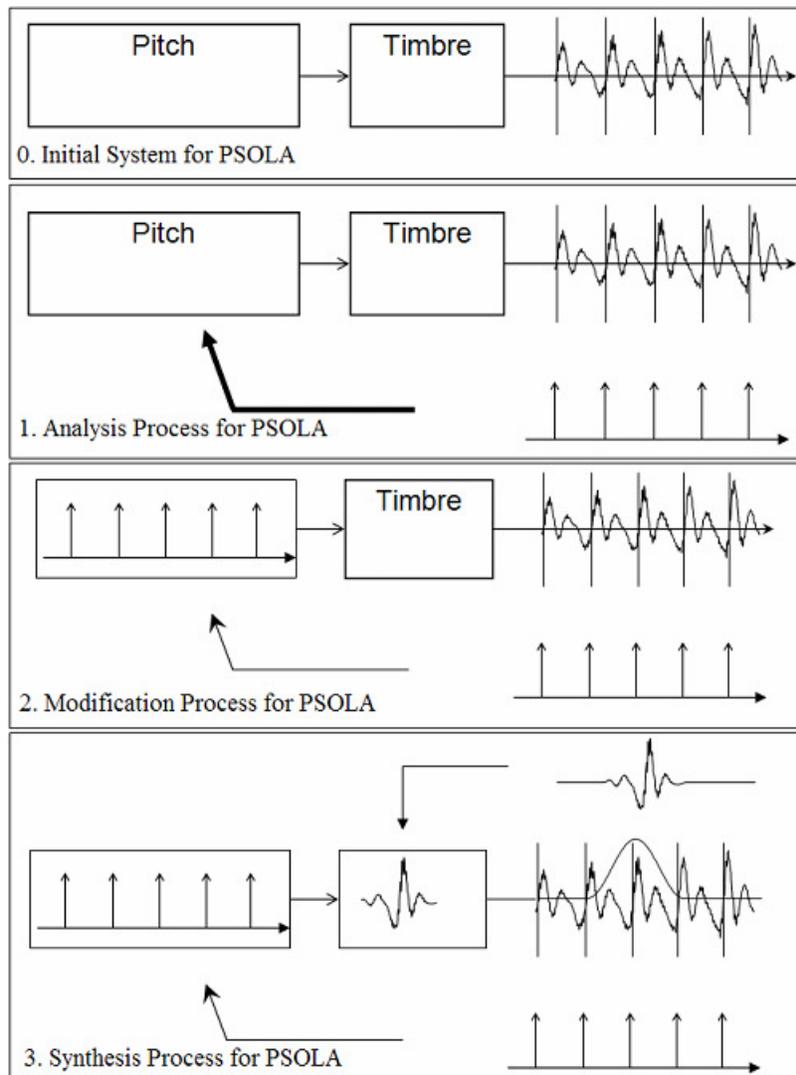


Figure 3 - PSOLA Process Overview

The quality of the modification results can be very high, but critically depends on the determination of the individual pitch periods (epochs) in the speech signal. The sound quality of TD-PSOLA (Time Domain PSOLA) modified speech is very sensitive to a proper positioning of the pitch marks that delimit the individual pitch epochs. As speech is a naturally produced signal, the robust estimation of pitch epochs thus becomes extremely important for TD-PSOLA applications, especially where real-time operation is required. [14]

The idea behind Prosody Transplantation is copying intonation and duration values from a recorded donor message to a recipient message. As TD-PSOLA is used for prosodic modification of speech signals, it can be used as a technique to implement Prosody Transplantation.

2.6 Prosody Transplantation

Prosody transplantation can be implemented to both single speaker and multiple speaker systems. As this study focused on single speaker systems (more details can be found in Chapter 3) in the next paragraph the prosody transplantation for single speaker systems will be described.

In a single speaker prosodic transplantation system, a trained speaker reads a sentence twice to produce two utterances. These two utterances, i.e. U_1 and U_2 , should have intonations which are clearly distinctive from each other. Waveform analysis is applied to both utterances to obtain the parameter sets which represent prosodic properties of the utterances. To determine the articulatory timing relationship between U_1 and U_2 , a pattern matching technique called *DTW (Dynamic Time Warping)* is applied. [15]

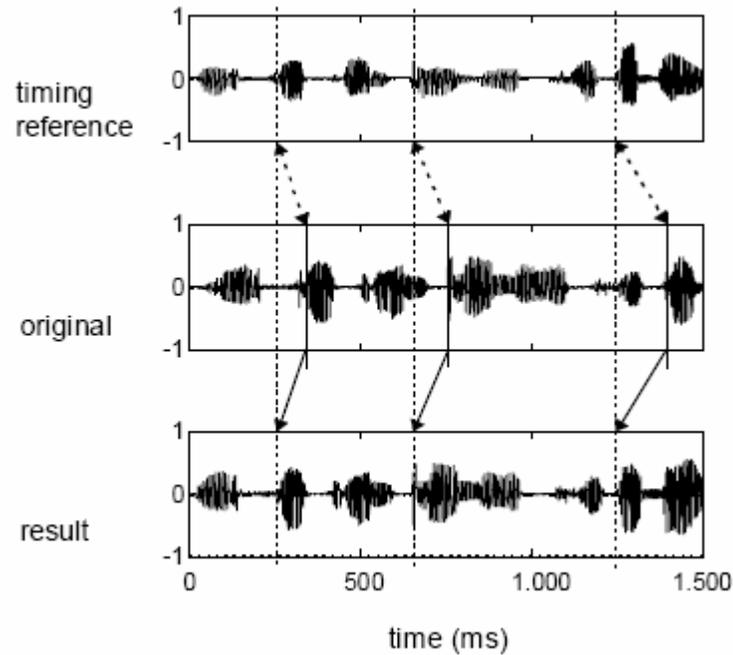


Figure 4 - After analyzing its relation to the timing [16]

DTW is an algorithm for measuring similarity between two sequences which may vary in time or speed. DTW is used to compute the timing relationship (time-warping path) between the original utterance and the utterance that serves as timing reference. [03]

The time scaling needs to be applied to the parameter sets obtained from U_1 before moving further. Now the selected prosodic parameters of U_2 can be substituted into the corresponding contours of U_1 . One can construct a synthesised utterance U_x by using waveform synthesis with the already modified parameters. The synthesised utterance U_x would have timing corresponding to the timing of U_2 . The prosodic properties of U_x , which are pitch, loudness and articulatory contours, correspond to those of either U_1 or U_2 . This can be summarized as transplantation of prosodic features from a donor utterance U_2 to a receiving utterance U_1 . [15]

In this study, PSOLA method has been used for prosody transplantation to synthesise emotional speech. More details on the implementation of the methods described will follow in the next chapter.

3
3

Chapter 3

3. Building Emotional Voice

This chapter describes the system that produces emotional voice. Firstly in 3.1, the overall system will be summarized. In 3.2, the actual system development steps will be explained, afterwards the possible applications of developed emotional speech synthesizer will be discussed in 3.3 and finally in 3.4, the evaluation method will be explained and the results of the evaluation will be stated.

3.1 General System Overview

The developed system produces an affective speech by firstly generating nonsense language from a neutral database and then transplanting the prosodic features of the templates from expressive database onto the nonsense phrases of that language.

The system uses two databases. One of them is constructed from the naturally spoken utterances of a professional speaker and the second one is constructed from expressive utterances of the same speaker. The details of these databases will be explained in 3.2.3.

Firstly, an expressive template is selected from the emotional database. To produce a nonsense carrier phrase in the same segmental structure with the template, the system chooses the segments from the first database in a somewhat clever way to provide inter-segment compatibility and then concatenates them one after the other. In 3.2.4, the segmenting procedure of the neutral utterances and the joining procedure of them, so called producing nonsense language, are explained in detail.

Then to transform this neutral nonsense carrier phrase into emotional phrase the prosody transplantation [17], which is explained in previous chapter, is applied. This is done through altering the timing, pitch and voice quality features of the carrier sentence in accordance with the template.

Both the prosodic modification and concatenation is done by using pitch synchronized overlap-add (PSOLA) [03] method which was described previously in Chapter 2. A text-to-speech diphone synthesizer [18], which was developed based on PSOLA technique, is optimized and used as the underlying synthesizer of this system. The overview of the general system can be seen from Figure 5.

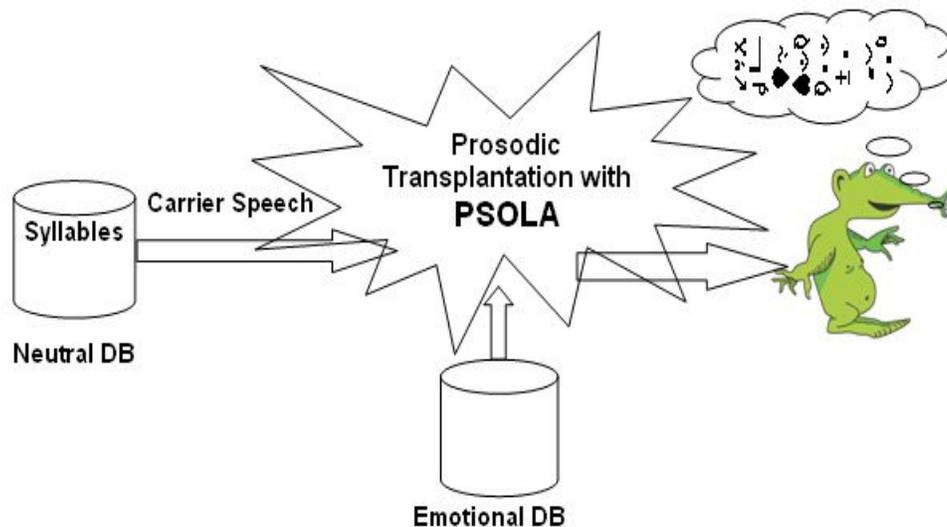


Figure 5 – High Level System Overview

3.2 Developing the Actual System

In this section, all the parts in developing the system will be explained step by step.

3.2.1 Speaker Selection

In order to produce the expressive speech as close as possible to the natural emotional speech, the quality of the prosodic templates was an important parameter, which was also mostly related with the speaker. The most important criteria for the speaker selection were:

1. ***Ability to keep the voice in the same tone:*** While recording the neutral text, the speaker should be able to keep the voice in the same tone for a long period which was hard and tiring ability and so required one who was familiar with such readings.
2. ***Ability to portray the desired emotions:*** On the other hand, the speaker should be able to express the desired emotions convincible since those emotional utterances would be used as prosodic templates and so would affect the overall system quality.
3. ***Suitability to a cartoon friend for children:*** As the produced voice would be a possible voice of a robot friend for hospitalized children, the voice should be acceptable for children.
4. ***Amateurship/Professionalism in speech:*** Since the above criteria would be hardly covered by an amateur speaker, the familiarity to the speech/acting was an important factor. Also the speaker should feel comfortable in studio environments where the recording would take place.
5. ***Easy access to the speaker when needed:*** As the recordings could be done repetitively, easy access to the speaker was also another important criterion. And the speaker should give a high attention to the work.

By taking into account all those above criteria; a professional speaker who was familiar with children for many years is selected among four speaker candidates.

3.2.2 Emotion Selection

As can be seen from a quick review of Chapter 2 of emotion in speech, there are many emotion categories. Among all those categories some of them should be decided to use in the research. Since the scope of this project is children, four primary human emotions

including anger, joy, sadness and fear are found enough to provide a basic communication with children. Those emotions on the other hand were easy to portray by the speaker.

During the recording sessions, fundamental prosodic characteristics of these basic emotions like pitch, speech rate and loudness/intensity are provided as the input to the speaker before each utterance recording and sometimes when the speaker needs to switch between the emotions, reference expressive sentences are made listened to the speaker to get him back into desired portraying mood.

3.2.3 Database Collection

Two databases were needed for the overall system. One of them was to produce neutral nonsense carrier sentences and the other one was for emotional templates. This section describes the steps of constructing those databases.

3.2.3.1 Data Collection Procedure

The term *neutral* for the first database corresponded to *emotionally unbiased*, but at the same time since that neutral database would be used to construct the nonsense language, the carrier sentences of that language should sound human-like. Thus, for the data to construct that database, the essential part was to having the naturalness of human conversation in the data. So, not articles to be read but screenshots of daily human talk are selected for the first database. Finally five sentences from those screenshots and a half page long text from a children story book is elected for recordings.

The second database is constructed according to two strategies. Firstly the five emotionally unbiased sentences from the first database are rerecorded by portraying them emotionally for each desired expressive category this time, like done in researches such as [19]^{vii}. The second strategy was recording utterances including expressive contents on themselves so to help the speaker easily get the mood of desired emotion, and then

^{vii} <http://sail.usc.edu/publications/BulutNarayananSyrdal.pdf>

extracting emotional highlights from it, like the method in constructing available emotion databases such as [20]. In the survey evaluations of recognizing the portrayed expressions which will be explained in the following sections, the second strategy is found more successful.

For the both two databases many storybooks, especially the ones for children as they are written in more expressive way and websites of plays are reviewed to find the emotionally unbiased and the correct expressive sentences. All these data can be found in Appendix A.

3.2.3.2 Recording Procedure

The recordings are done in an anechoic chamber. Thus, the background noise is minimized. Also all the equipments except that are needed for recordings are put outside the room.

The selected samples are converted into prompts^{viii} to provide them on the computer screen to the speaker. This was on one hand to prevent the unwanted noise of the papers added on the recorded utterances and on the other hand it was easier for the speaker to read the samples from the screen with big fonts rather than reading from paper.

The speaker was asked to sit always the same distance from the microphone in the same position of the room and in the same environment^{ix}. Also a headphone was provided to the speaker to hear his voice, so that he felt more comfortable.

Since the produced affective voice would be the possible voice of a robotic friend for children, it was decided to make recordings in two different coloured voice of the same speaker,-one was his natural colour (human-like) and the other one was cartoon-like colour-. The choice about which of the colours to work with would be evaluated afterwards. Firstly human-like recordings are completed.

^{viii} Prompts were a part of recording software that was used in the anechoic chamber.

^{ix} *Same environment* meant that the same position of the equipments in the room.

A sign with hand was given to the speaker to start reading the script. The utterances were directly recorded to the computer and the recording was repeated if necessary. Firstly, the neutral samples are recorded since that would be easier to start and a good practise for the speaker to get used to the studio environment. Secondly, the expressive samples are recorded emotion by emotion. The speaker could decide if he wanted to read either anger, sad, joy or fear first. Then the acoustic characteristics -like pitch, timing and voice quality- of the decided emotion label were reminded to the speaker and sometimes when needed, reference affective samples were provided before each script.

The recording was done in two days in four main sessions with small breaks. It was very tiring for the speaker especially during the expressive database recordings with cartoon-like coloured voice. Because he needed to change his voice and try to keep that voice in the same colour for a long time which forced and hurt his throat and larynx.

In between the third and fourth sessions all the neutral and expressive utterances were listened and if necessary some of them were rerecorded when the researchers could not detect any emotional colouring or detect opposite emotional colouring than intended. Finally there were 97 utterances including a few versions of some samples. The next step was to evaluate those utterances and constructing the databases.

3.2.3.3 Evaluation Procedure and Final Database Construction

The voice and the utterances to use in constructing the databases are selected through an evaluation process. Four criteria are determined to evaluate the quality of the utterances:

- Colour of the voice (human-like/cartoon-like),
- The emotion felt in the utterance,
- The closeness to the desired emotion,
- The quality of the portraying (imitated/real).

Ideally the utterances should be rated by a number of children since the application is aimed to serve them. But as the prepared survey was too long for them, keeping their focus for such a long time would be difficult. So the utterances are rated according to the

above criteria via an electronic survey^x by four researchers who are familiar with speech processing and one amateur musician.

Since the voice should sound as natural as possible a very important criteria was the *quality of the portraying (imitated/real)*. The produced affective speech should not give the feeling that it was imitated / faked. Thus, even though the cartoon-like voice was found more successful in portraying desired emotion, since on the other hand human-like voice was found sounding more real, it is decided to work with the human-like voice. Figure 6 and Figure 7 illustrate the survey results. It is easy to see that the cartoon-like voice was found more imitated which means not natural.

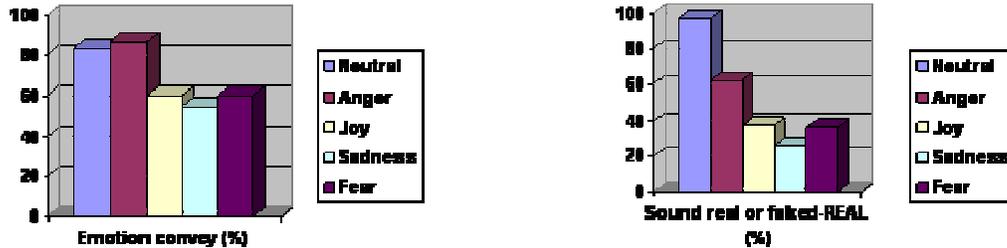


Figure 6 Statistics of Human-Like coloured voice

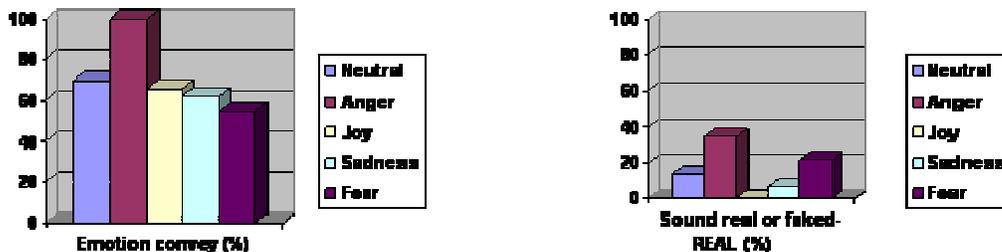


Figure 7 Statistics of *Cartoon-Like* coloured voice

After the voice colour decision, the samples are evaluated one by one for the databases. The important criteria were again the realness rate and emotion conveying rate. First, for each expressive category, the utterances which have the rate above 80 % or more for the *recognition level of desired emotion* criterion are selected. Then, the other criterion, *real or imitated*, is evaluated and the utterances which have the rate above 80 % or more for

^x The template of the survey can be found in the Appendix B.

that criterion are selected and put in the final database. If there were not enough samples for the examined expressive category, then the threshold value was decreased from 80 % to 60 % for the realness criterion. The threshold for the conveying the desired emotion criterion was always kept at 80 % and was not decreased, since this criterion was the core criterion for the quality of the prosodic templates.

The selected emotional speech samples are then put in emotional database to use as prosodic templates and the selected neutral speech samples are put in neutral database to construct nonsense carrier phrases. Table 1 and Table 2 illustrate the data quality rates of final Neutral and Emotional Databases. It was found that portraying *sadness* was the most difficult expressive category to act, so that the evaluators detected the correct emotion sufficiently only in one sample. Most of the evaluators thought the sadness samples were closest to *boredom* or *neutral*.

Desired Emotion	Number of Final Utterances	Emotion Convey (%)	Sound real or faked-REAL (%)
Neutral	7	91.4	97.1

Table 1 Final Neutral Data Quality

Desired Emotion	Number of Final Utterances	Emotion Convey (%)	Sound real or faked-REAL (%)
Anger	10	100	80
Joy	4	95	65
Sadness	1	80	60
Fear	8	90	60

Table 2 Final Expressive Data Quality

Next, waveforms of some long sentences of the neutral utterances are cut into smaller sub sentences. The smaller sentences were giving more precise and correct results with segmentation software. The final databases can be found in Appendix C.

3.2.4 Producing Nonsense Language

After the database collection, the next step was to construct the nonsense language to carry expressive states. This is done through firstly segmenting the speech samples from

neutral database into small segments and then concatenating some randomly selected pieces. In this section the methods of these sub steps will be explained.

3.2.4.1 Segmenting the Data

An interactive segmentation tool is developed based on the MEL-scale spectral distances. The tool works as follows.

First, the utterance to be segmented is needed to be selected and loaded. Then, for each sample point in the signal, MEL-cepstral distance is calculated between the Hanning windowed speech frame on the right and the frame on the left of the sample point, so that a distance array is calculated:

$$d(i, j) = \sqrt{\sum_{n=1}^L (r_i(n) - s_j(n))^2}$$

L = number of coefficients, which is used as $L=24^{\text{xi}}$,
 i, j = frame indexes s.

For the MEL-cepstral distance calculation 30ms Hanning window, with 99% overlap, is used. Because of that 99% overlap and the repetition of the calculation for each sample point in signal, this local distance array calculation is time consuming.

Large MEL-scale cepstral distances are an indication of a phone transition and small distances indicate stationary speech parts. The tool plots these distances and, given the desired number of segments, the user can specify the appropriate cut-points. These cut-points are then stored in a meta-data file that can be used for constructing the nonsense carrier utterances by concatenating randomly selected speech segments from the first database. Normally the meta-data file could be written with XML [21], but since there were small amount of data there was no need for that.

^{xi} The zeroth coefficient which only corresponds to the frame energy is asked to user as input for the inclusion into the calculation of local distances or not.

Figure 8 illustrates this segmentation process. The first panel shows the speech utterance ‘not ring’. After getting the number of segments as input from user, the tool allows specifying the cut points/point –which is one in that case- on the figure. The bottom panel shows the MEL-scale cepstral distances and the user-specified one cut-point.

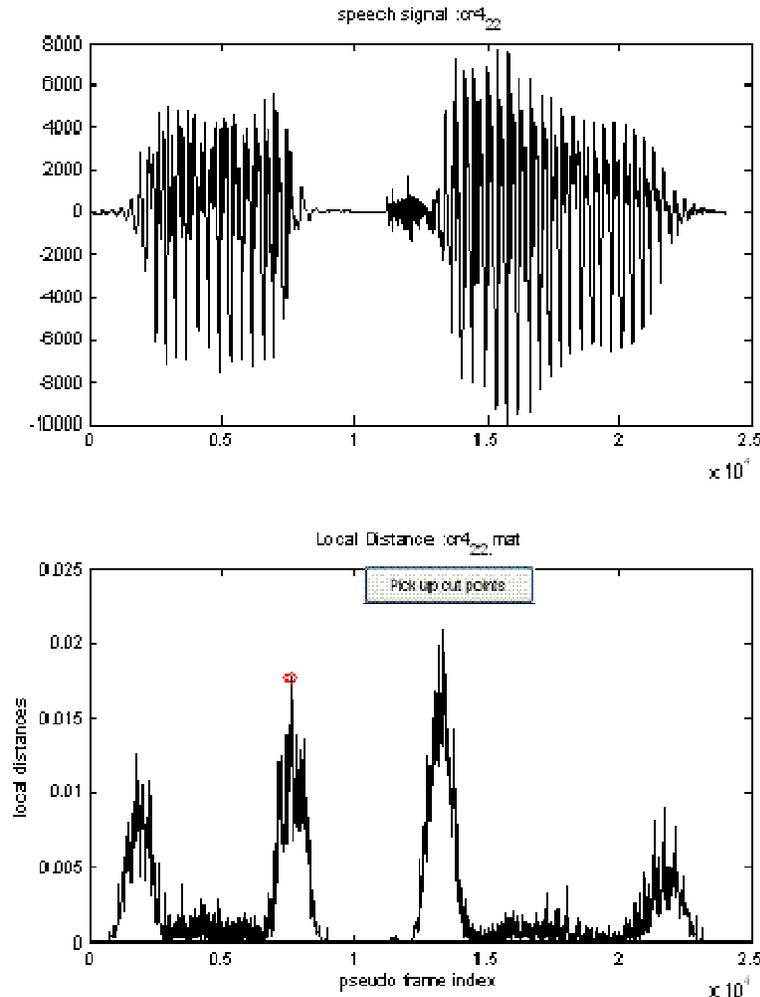


Figure 8 MEL-cepstral distance based manual segmentation. The upper panel shows the speech utterance ‘not ring’. The bottom panel show the MEL-scale cepstral distances.

By using the above tool, each utterance in the neutral database and in the expressive database is segmented and the cut-points are stored in the correct meta-data files. Neutral samples are segmented to construct the nonsense carrier phrases, expressive samples are segmented to analyse their syllabic structures. Finally there were 34 segments in the first database for concatenation, which was very limited source.

Also with that tool, the start and the end of original speech in the overall signal is marked^{xii} and stored in a separate file together with the segment information. A sample of that file can be seen in Figure 9.

cr1.wav	5	1645	43395
cr2.wav	3	705	28533
cr3.wav	5	1495	36316
cr4.wav	5	1262	60984
cr5.wav	3	7682	39571
cr6.wav	6	1704	80590
cr7.wav	7	6680	78675

Figure 9 All_carriers table. First column contains the names of the all neutral samples. Second column corresponds to number of segments in the file. Third and last columns correspond to start and end point of the speech in the overall signal respectively.

Once having these meta-data files and so the segments, the next step was the concatenation of these segments to complete the construction of nonsense carriers.

3.2.4.2 Generating Carrier Phrases by Segment Concatenation

Given the input about which expressive state desired to be transplanted, the tool selects a random emotional template from the corresponding expressive database. Then it reads the syllabic structure files which are stored previously and contain the number of segments and the duration of each segment. Next, the tool selects the segments and realizes the concatenation one by one for each segment.

Segment Selection

The carrier phrase needs to be constructed in the same syllabic structure with the desired expressive template. That means that the phrase and the template should contain the same number of segments and the segments of carrier and template could be closer to each other in their duration manner within a range. Thus, the segments should be selected somewhat in a clever way.

^{xii} The overall signal can contain silences at the beginning and at the end. It is needed not to take into account those silences, since inclusion would result in wrong duration calculation of segments.

Since the segments directly affect the synthesis results, this selection procedure is very important. Let's explain that on an example. Assume that anger is selected as the prosodic template. As it is known from Chapter 2, it has higher speech rate than neutral. That means the segments are shortened by a factor in time when one is angry. Now, assume that the randomly selected segments are shorter than the template segments. That means those segments must be lengthened by a factor in time to adapt it to the template phrase. As a result, the final synthesized nonsense phrase will be far from the desired, and so, far from the human nature. This is the lower limit to be assured for the selected segments. There is also an upper limit. Now if we assume that the selected segments are too long compared to template segments, this time the conversion in time will result in a non-noticed sound which cannot be pronounced by human. That also will bring the final synthesized nonsense phrase far from the human nature. This limits will be opposite in the case of sadness.

Thus, for each expressive state, one upper and one lower limit are used in the segment selection procedure. The first limits are easily calculated by using the databases. As can be remembered from Database Collection section, the expressive database was constructed according to two strategies. In the first strategy, the emotionally unbiased sentences from the first database were rerecorded and stored in the second database by portraying them emotionally for each desired expressive category. Thus, we had the same sentence in neutral and in all different emotions. By comparing these phrases the conversion rates and so the first limits are calculated. Ideally these calculated numbers should be used, but since there was a very limited segment database, sometimes the suitable segments were not found in the database. Thus, these numbers are adapted to the available system. The second coefficients on the other hand are determined by experiments on the system. So they are also adapted to the available system.

Concatenation

Once the expressive template is selected, the syllabic structure (like number of segments, and durations of each segment) is read from corresponding files, the first two segments to

construct the carrier phrase are selected, the next step is cutting these segments one by one from their original files in an appropriate way and concatenating them.

In that part text-to-speech phoneme concatenation tool, which was developed under our department, [18] is optimized to segment concatenation. For each segment, firstly the corresponding origin file and its previously stored pitch marks are read. Next, the nearest pitch marks to cut points are found and the segment is cut. Then all the pitch marks, and segment transition points –which refer to the end of each segment-, are adapted to the cut segment. Finally, the prepared segment is concatenated with the previously joined segments.

In the concatenation procedure, by choosing pitch markers as the segments cut-points to concatenate, the periodicity of the speech signal is assured not to be disrupted. On top of that, the tool finds the best cut-markers. This is done through, for each join, by selecting a pitch marker from the first and from the second segment in such a way that the transition will occur where there is as much similarity between the two segments as possible [08]. This similarity is found according to MEL-scale spectral distances.

After determining the cut marks, the pitch synchronisation of the transition is applied. This is to avoid one of the concatenation problems, which is, the introduction of irregular pitch periods at the transition point and which causes audible concatenation artefacts. For that synchronisation firstly, the tool firstly selects a number of pitch periods^{xiii} from the beginning cut-marker of the first and second segment, respectively. These are called *joinframe1* and *joinframe2*, respectively. Then, the pitch of these join frames is altered using the PSOLA technique, which will result in two signals having exactly the same pitch.

Finally, the two signals are cross-faded using a Hanning-function. This is to avoid another concatenation problem, which is, the appearance of abruptness at the transition between the two signals and which causes the concatenated speech would not sound fluent, but chopped. Thus, the segment transition is smoothed among a few pitch periods.

^{xiii} This number of pitch periods is typically 5.

Figure 10 illustrates that concatenation method by joining two voiced speech segments. As can be noted the first segment has a pitch value which is higher than that of the second segment. That can be seen in the upper panel of the figure. The middle panel shows the pitch-synchronisation of the extracted pitch periods, and the bottom panel shows the final concatenated speech, on which the segment transition is smoothed among a few pitch periods.

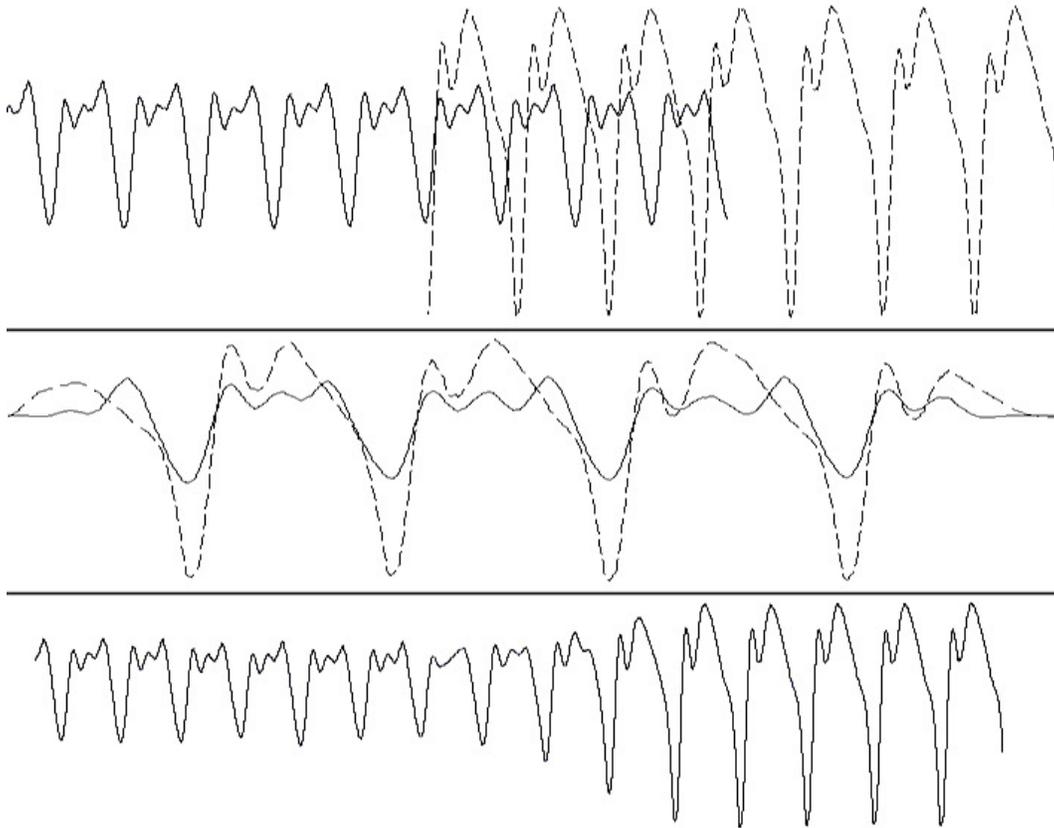


Figure 10 Pitch-synchronous concatenation.

As the last step, the position of the pitch marks and the segment transition points in the concatenated signal is calculated.

3.2.5 Adding Prosody and Generating the Desired Emotions

After producing the nonsense language and the phrases of it in the same syllabic structure with the expressive template, the next step was to transplant the prosodic features - pitch, timing and voice quality - of the template onto the carrier to generate emotional speech from that language.

Pitch and Timing Transplantation

As mentioned before, PSOLA technique is used to transplant pitch and timing features of donor onto recipient. At this point two inputs needed to be given to the system. First one is the information of prosodic structure of the template. That is the duration and average pitch of each segment. This information was previously stored in the database. Second one is the pitch markers of the nonsense speech, which was computed in the previous step from the pitch markers of the concatenated segments.

For altering the pitch and timing, these pitch markers of the nonsense speech is used as analysis-pitch markers for the PSOLA technique. Since the quality of the output signal greatly depends on the correctness of these markers an efficient and robust algorithm is used [22].

As we know by using the transition points, which indicates the segment change point, the synthesizer calculates the inherent length of each segment present in the concatenated signal. Using these lengths and the lengths of the segments present in the template – which is read from the database-, the amount of time-stretching that is necessary to provide the output speech with the correct timing properties is computed. Then, the PSOLA algorithm synthesizes the output signal by using a time varying time-stretch value going from segment to segment. Finally, the synthesis pitch markers are calculated by the PSOLA operation based on the pitch-parameters of the prosodic template [23]. These synthesis-pitch markers determine the pitch of the final output and thus, the imposed intonation curve ensured to be correctly assigned to the final speech signal.

Voice Quality Transplantation

After pitch and timing modification, the last prosodic feature to be transplanted is voice quality. In that step, the energy available in each segment of template (E_{temp}) and carrier (E_{carr}) is calculated. Next, every sample point in the segment of carrier phrase is multiplied by a factor of square root of E_{temp} / E_{carr} .

$$\sum_{m=0}^{N-1} s(m)^2 = E_{temp}$$

$$\sum_{m=0}^{L-1} w(m)^2 = E_{carr}$$

$$f(m) = w(m) \sqrt{\frac{E_{temp}}{E_{carr}}}$$

This operation is applied to each segment present in the carrier speech and so the final loudness of the donor is transplanted onto the recipient.

Figure 11 illustrates the affect of that complete prosody transplantation. The emotional template - which is ‘anger’ in that case- can be seen in the upper panel. There are 9 segments present in the signal. In the middle panel, the concatenated nonsense carrier phrase, which also contains 9 segments, is illustrated. As can be seen, initially the length of the carrier is more than the template. After prosodic transplantation with PSOLA this length is adapted to the length of the expressive template and that is shown in the last panel.

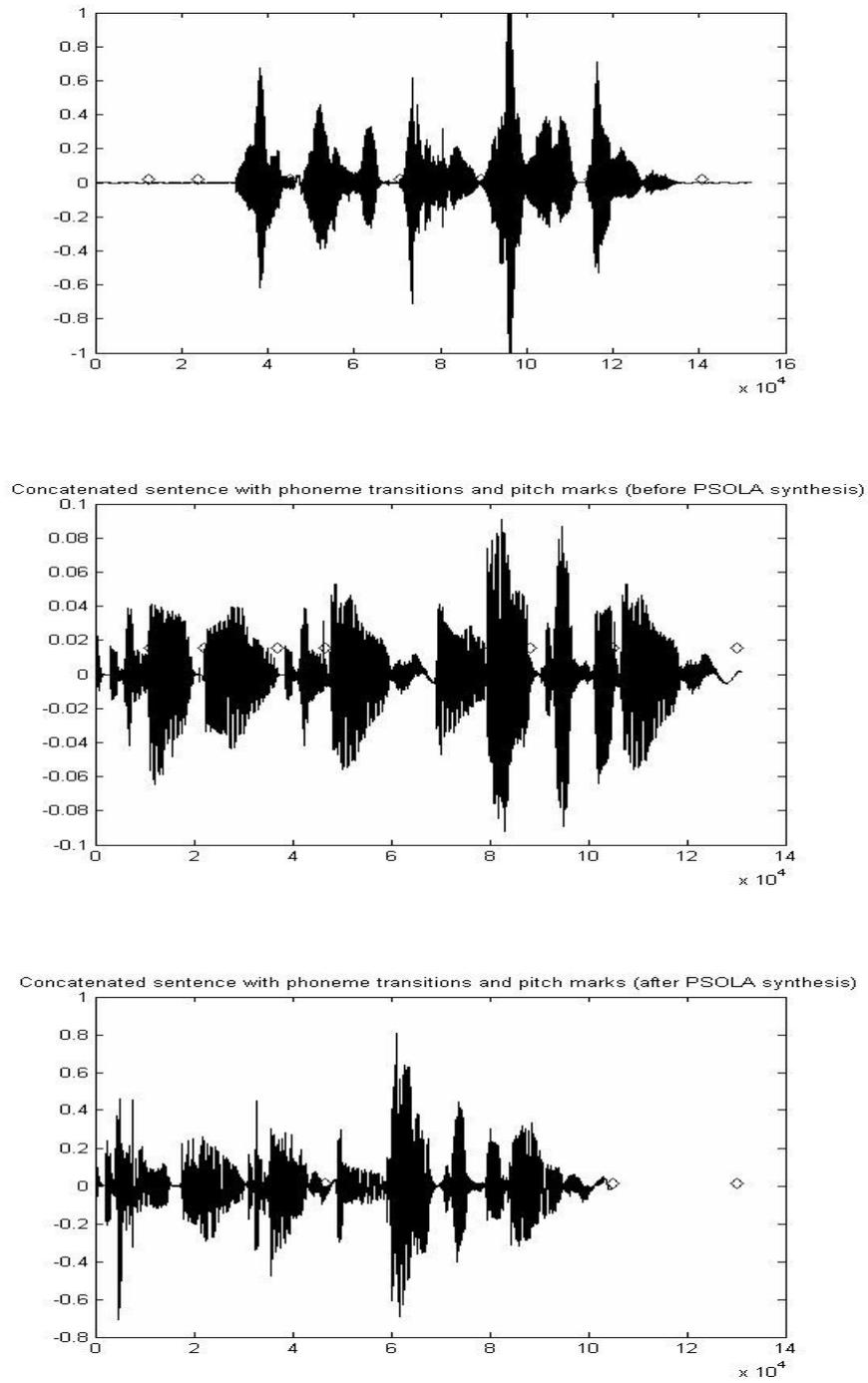


Figure 11 Prosody transplantation

3.3 Applications of Emotional Speech Synthesis

Mankind likes to communicate. Instead of pulling the lever or pressing the button, man prefers to talk. Instead of getting a printed paper or reading the text on the screen, man prefers to hear. Man always searched for the intelligence in the machine so that he can communicate, however, never liked the cold mechanic comments and desired that the interaction has emotions so it is easy to embrace. Basically man always searched for the soul in the machine.

The vision and desire of the mankind for an intelligent and emotional machine have been one of the mostly used concepts in the filming history. Sometimes it was working for the good of mankind like Sonny in “I, Robot” or being the cold blooded killer like HAL in “2001: A Space Odyssey”^{xiv}.

“2001: A Space Odyssey” is an influential 1968 science fiction film directed by Stanley Kubrick. The screenplay, written by Kubrick and Arthur C. Clarke, deals with themes of human evolution and technology, artificial intelligence, and extra-terrestrial life. The film is notable for its scientific realism, and its predictions of the future; some were accurate, while some were not.

Accurate predictions for 2001 include; ubiquitous computers^{xv}, small portable flat-screen computer monitors and televisions, glass cockpits in spacecraft and biometric identification.

Some of the film's inaccurate predictions of the future turned out to be colonies on the moon, routine commercial space flights, hotels orbiting 2000 ft above earth, manned missions to other planets, technology of placing humans to "suspended animation", and HAL's speech, understanding and self-determining abilities.

^{xiv} www.imbd.com

^{xv} Ubiquitous computing (ubicomputing) integrates computation into the environment, rather than having computers which are distinct objects. Other terms for ubiquitous computing include pervasive computing, calm technology, things that think and every ware. Promoters of this idea hope that embedding computation into the environment and everyday objects would enable people to interact with information-processing devices more naturally and casually than they currently do, and in whatever location or circumstance they find themselves. (www.wikipedia.org)

Maybe today we do not have advance technology to have HAL or Sonny, but the developing technology of emotional speech synthesis and emotion recognition have many application areas. This section focuses on the existing and potential implementation areas for the emotional speech synthesis.

One of the most used implementation areas of text to speech technology is to support the visually impaired people. Traditionally, live book reading sessions are held or pre-recorded readings are available for the visually impaired people. These methods require the extensive utilization of resources, and huge effort is needed to keep the existing materials up to date. With the introduction of the text to speech technology, now any text is available for the handicapped people to review any document at any time. With the application of emotional speech synthesis, the quality of the experience can be increased significantly.

Intelligence and imagination development of the small children are highly effected by story hearing^{xvi}. There are special courses available to parents, in which the parents get the training on using their voices appropriately while reading the story books so that each individual character and their emotions in the story are presented distinctly to the listening children. By utilizing the emotional speech synthesis, it will be possible to improve the story hearing experience for the little children and thus the benefits of the book reading.

The same logic of text reading for visually impaired people and small children can be applied to academics. With the proper set up, e-courses can be designed where actually the course is given by an e-creature, with the proper quality of teaching experience. This could be a huge opportunity for academic learning and personal development.

It is also important for the children and old people to have a companion. In the non-existence of a real person, robots or other artificial creatures can be used for accompanying them. Having someone you can talk, share your feelings with, getting proper responses and actually having an emotional interaction is crucial for the

^{xvi} http://www.trelease-on-reading.com/rah_chpt1_p2.html

psychology of person not have them feel lonely. There is a big application opportunity of emotionally speaking creatures to accompany the people.

Other industrial application areas of emotional speaking creatures could be the automotive industry. The interface of cars controlling navigation, climate, and car diagnosis – supplying the feedback for the current condition of the car, like the fuel level – can utilize the emotional speech synthesis to express the right seriousness level.

The application areas of the emotional speech synthesis are limitless on the PBX telephone systems and customer service. Understanding the feeling of the customer and giving the right responses with the right attitude is highly important for the service quality. By predetermining problem as well as the mood of the customer, businesses can increase their employees' efficiency and decrease the amount of time receptionists and office managers spend answering phones, while promoting a high level of customer service and professionalism.

Entertainment industry is maybe the most financially attractive application area of the emotional speech synthesis. Computer games are getting more and more interactive each day and instead of spending money for long hours of recording with professional speakers, utilization of emotional speech synthesis can be a significant cost saving for the proper voice generation in the computer games. Animation movies can also benefit from the same technology and the related cost savings.

As described above, the implementation areas of the emotional speech synthesis is endless. The other potential implementation areas are related to linguistics and tourism. Foreign language learning and improving the diction quality can utilize the emotional speech synthesis. It is also possible to design self-guided tour information to the travelling tourist.

The technology has a big potential for cost and effort saving, while improving the quality of the service and the experience received by the users. Any implementation involving an artificial creature will be a focus for emotional speech synthesis' applications.

3.4 Evaluations

The accuracy of the developed emotional synthesis was assessed in a perceptual test. By using the same evaluation criteria as in the voice and utterance selection procedure for the final database construction, sample outputs of the system are rated by a number expert group. In this section, firstly the evaluation method will be explained, and then the results will be analyzed.

3.4.1 Method

To evaluate the system, three templates for each expressive category, except sadness, are selected to be used. For sadness, since there was one template in the database, only that one template could be used for it, but the carrier phrases are selected more than done for the other categories. These templates are chosen by looking at their previously database evaluation results. The most important criterion here was ‘*the emotion convey*’ values of them. The selected expressive templates and their corresponding ‘*the emotion convey*’ and ‘*sound real or faked-REAL*’ values can be seen from Table 3.

Emotion Category	Template	Emotion Convey (%)	Sound real or faked-REAL (%)
Anger	ang1	100	80
	ang2	100	80
	ang8	100	80
Joy	joy2	100	60
	joy3	100	60
	joy4	100	60
Sadness	sad1	80	60
Fear	fea1	80	60
	fea3	80	60
	fea7	100	60

Table 3 Selected emotion templates and their corresponding statistical values.

Two sets of nonsense carriers are determined to produce for each template to see the affect of the carrier. Only for sadness four carriers are generated.

The experiment had 7 participants from which 4 were researchers working on speech processing and 2 were amateur musicians. 5 of these 7 participants had also evaluated the previous database evaluation process. All of them were male having the average age of 26.85.

Again four criteria –three of them are the same as the ones in database evaluation- are used to evaluate the outputs. These are:

- Artificiality recognition (human-like/machine-like),
- The emotion felt in the utterance,
- The closeness to the desired emotion,
- Emotion persuasiveness (imitated/real).

The survey consisted of two parts. In the first part, which will be referred as uncontrolled, only synthesized outputs are asked to be rated without giving the corresponding neutral carrier. In the second part, which will be referred as controlled, evaluations are done by firstly listening the carriers and then the outputs.

3.4.2 Results

When the results are analyzed, it is seen that there was a big difference in the first part and in the second part. This comparison can be seen from Figure 12. The synthesized outputs are found more successful and easier to recognise the emotions on them, when the neutral initial carrier utterances are first listened. The reason of that was, by listening firstly the carrier, the structure of the language was perceived so that the output could be evaluated more correctly. Thus, the statistics of the second part are more realistic in evaluating the overall system quality.

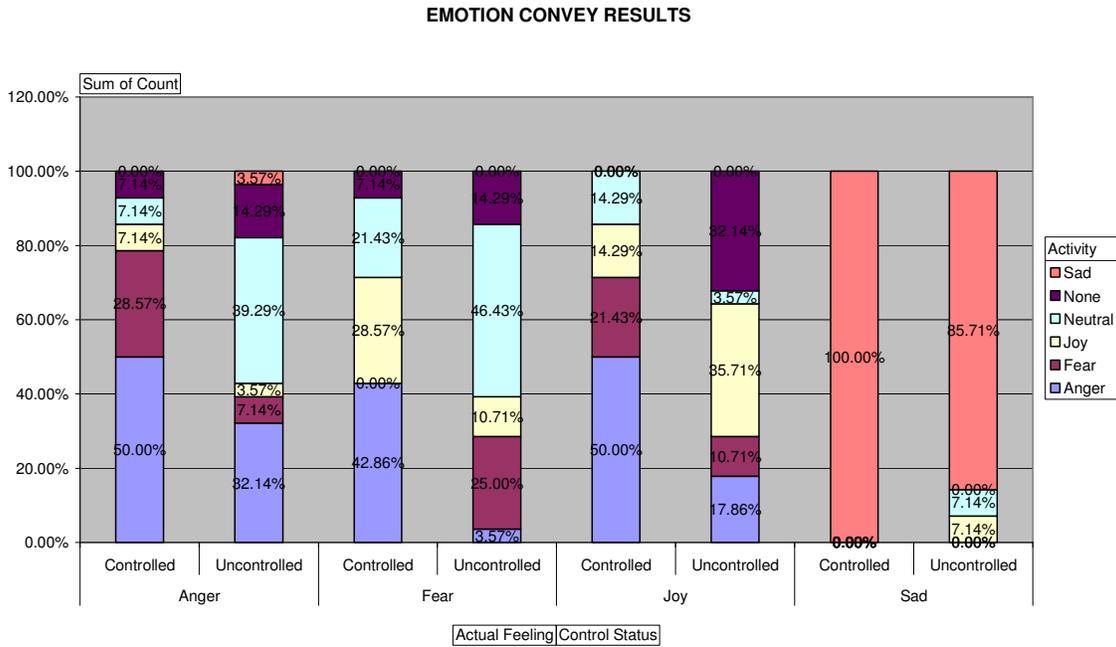


Figure 12 Emotion convey results

It is seen that always sadness was the easiest recognized emotion with a rate of 85.71%, 100%, and 90.48% in the controlled part, uncontrolled part and totally, respectively. The other expressive states are not perceived clearly, and even in uncontrolled evaluation sometimes an opposite expressive state is felt more than the original. On the other hand while it is easily recognized, 62.5% of them are found imitated as illustrated in Figure 14.

When looked at each expressive category in template base, big differences are noticed. Although they are in the same expressive category, the recognisability is varied. As can be seen from Figure 13, each template has a different rate. Even sometimes, like in the case of Fear 7, the correct emotion is recognized by any of the evaluators.

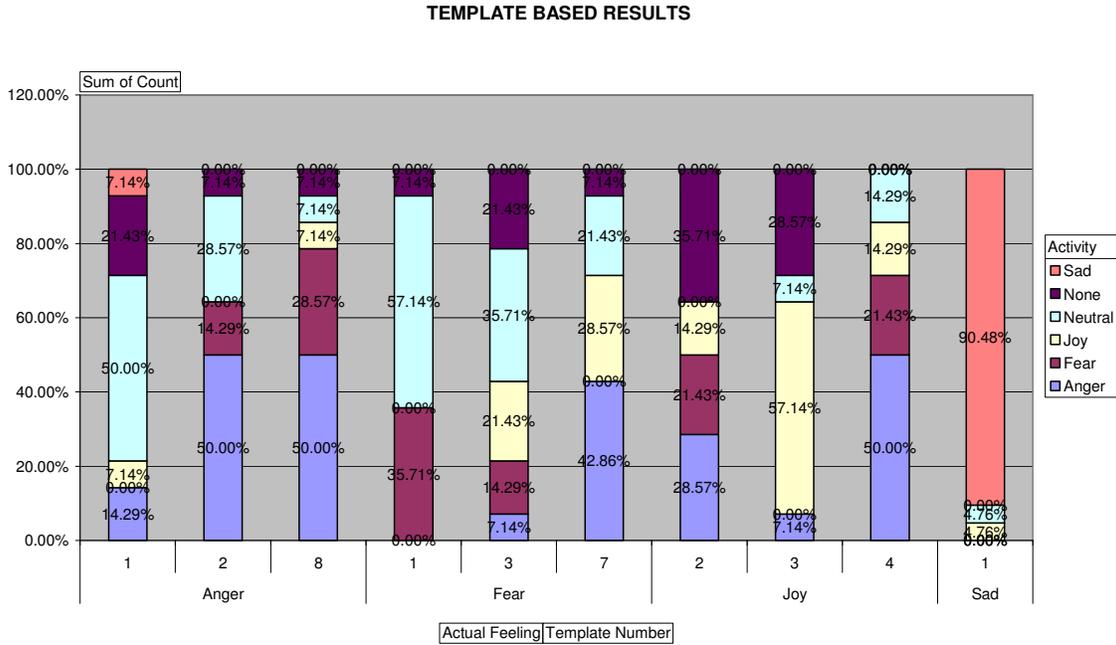


Figure 13 Template Based Results

The quality of the outputs for each template also varies in terms of different carrier sentences as shown in Table 4. This illustrates that the results are also depends on the produced carrier phrases. When looked at the sad as an example, since the same template was used for the four different carriers each time different results are achieved, like 86%, 57%, 100%, 100%, respectively.

The results for emotion persuasiveness, which is illustrated in Figure 14, show that the 51.73% of the synthesized phrases are found real while 48.27% of them sounded imitated. The initial template emotion persuasiveness was 66% real.

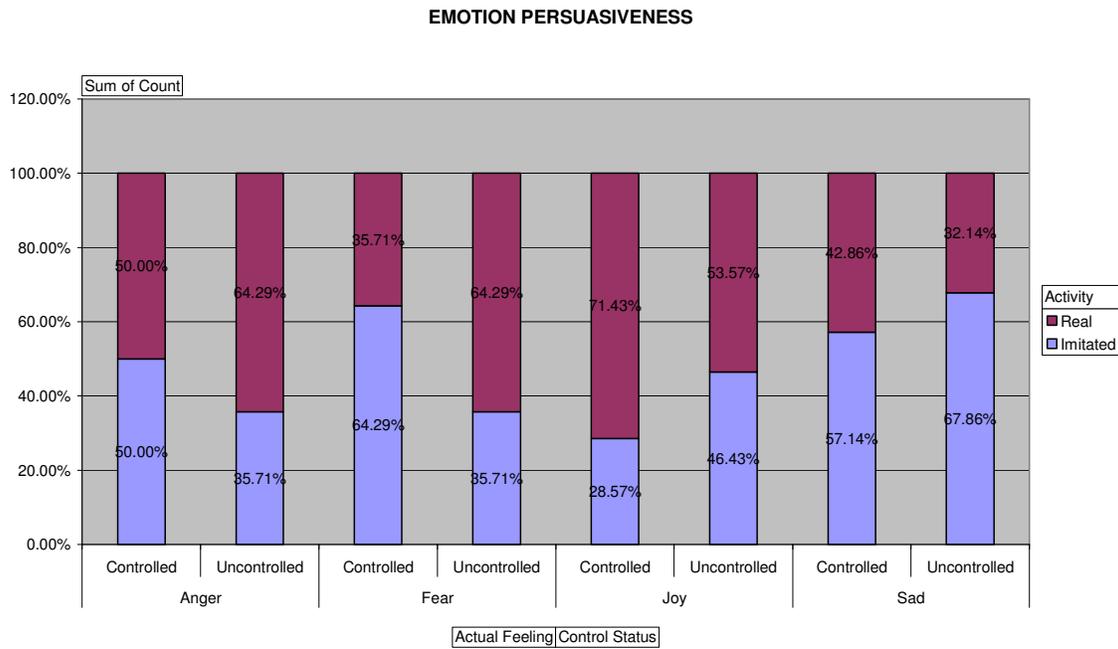


Figure 14 Emotion Persuasiveness Results

The total 46.62% overall synthesized outputs are found human-like and 58.38% machine-like. The distribution of these results on each emotion category can be seen from Figure 15.

Finally, all the statistical results are illustrated in Table 4.

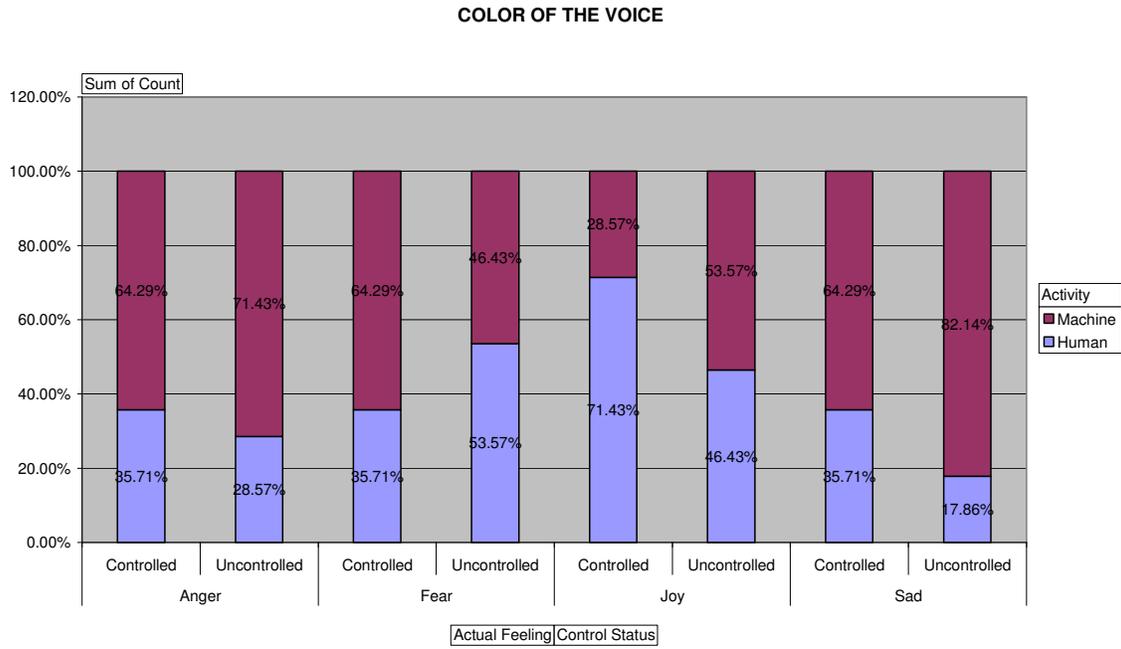


Figure 15 Machine-Like / Human-Like Distributions

EMOTION	TEMPLATE INDEX	CARRIER INDEX	CONTROL STATUS	ARTIFICIALITY RECOGNITION		EMOTION CONVEY						EMOTION PERSUASIVENESS	
				Human	Machine	Neutral	Anger	Joy	Sad	Fear	None	Real	Imitated
Anger	1	1	Uncontrolled	14%	86%	43%	14%	14%	0%	0%	29%	57%	43%
Anger	1	2	Uncontrolled	29%	71%	57%	14%	0%	14%	0%	14%	86%	14%
Fear	1	1	Uncontrolled	57%	43%	57%	0%	0%	0%	43%	0%	86%	14%
Fear	1	2	Uncontrolled	86%	14%	57%	0%	0%	0%	29%	14%	71%	29%
Joy	2	1	Uncontrolled	57%	43%	0%	43%	29%	0%	0%	29%	57%	43%
Joy	2	2	Uncontrolled	29%	71%	0%	14%	0%	0%	43%	43%	71%	29%
Sad	1	1	Uncontrolled	14%	86%	0%	0%	14%	86%	0%	0%	43%	57%
Sad	1	2	Uncontrolled	14%	86%	29%	0%	14%	57%	0%	0%	29%	71%
Anger	2	1	Uncontrolled	43%	57%	29%	71%	0%	0%	0%	0%	71%	29%
Anger	2	2	Uncontrolled	29%	71%	29%	29%	0%	0%	29%	14%	43%	57%
Joy	3	1	Uncontrolled	71%	29%	14%	0%	57%	0%	0%	29%	57%	43%
Joy	3	2	Uncontrolled	29%	71%	0%	14%	57%	0%	0%	29%	29%	71%
Fear	3	1	Uncontrolled	43%	57%	43%	14%	29%	0%	0%	14%	57%	43%
Fear	3	2	Uncontrolled	29%	71%	29%	0%	14%	0%	29%	29%	43%	57%
Sad	1	1	Uncontrolled	14%	86%	0%	0%	0%	100%	0%	0%	29%	71%
Sad	1	2	Uncontrolled	29%	71%	0%	0%	0%	100%	0%	0%	29%	71%
Anger	8	1	Controlled	57%	43%	0%	43%	0%	0%	57%	0%	29%	71%
Anger	8	2	Controlled	14%	86%	14%	57%	14%	0%	0%	14%	71%	29%
Fear	7	1	Controlled	29%	71%	43%	29%	29%	0%	0%	0%	43%	57%
Fear	7	2	Controlled	43%	57%	0%	57%	29%	0%	0%	14%	29%	71%
Joy	4	1	Controlled	71%	29%	14%	43%	14%	0%	29%	0%	57%	43%
Joy	4	2	Controlled	71%	29%	14%	57%	14%	0%	14%	0%	86%	14%
Sad	1	1	Controlled	29%	71%	0%	0%	0%	100%	0%	0%	43%	57%
Sad	1	2	Controlled	43%	57%	0%	0%	0%	100%	0%	0%	43%	57%

Table 4 the Overall Statistical Results

4
4

Chapter 4

4. Concluding Discussion

The expert survey results and the analysis made on these results have driven the following conclusions concerning the final results of this study.

When the expert survey results are analysed, it's been concluded that if the listener is familiar with the speakers nonsense neutral utterances, the emotions synthesised are recognised at a much higher rate. This can be a valuable input during the implementation of the emotional speech synthesis for Anty as the main objective of this study is to synthesize emotional speech for affective robot child communication. Our study suggests that the communication level of Anty with children with emotional nonsense speech should gradually grow by time. This would allow the listener children to understand the general characteristics of the nonsense language as well as the sound of the neutral utterances of Anty. Of course any psychological effect of this should be analysed by the area experts.

Also as can be easily seen from the analysis of the expert survey results, the emotional effect of sadness in synthesised speech was very successfully recognised by the listeners. Where as for the other 3 emotion types, the emotion recognition by the listeners were at a random level. The main reason for this difference in the recognition rate is the characteristics of the acoustic effects caused by the feelings. Sadness has unique characteristics such as lower pitch, slower time and lower loudness which make it easily recognisable among the other emotions. All the other 3 emotions share the characteristic of faster time and higher pitch, which makes them difficult to distinguish from each other.

The quality of the synthesized emotional speech relies on two important factors: The quality of the carrier phrases and the quality of the expressive template. Any inefficiency in one of these causes the synthesized emotion to be either robotic or un-recognizable by the listeners.

The quality of the carrier phrases is mainly having the carrier phrase neutral in terms of the emotional characteristics. If the initial carrier phrases include emotional effects, the transplantation results are ineffective. Also as the database used in this study included limited number of segments, during the random selection it is very likely to get a non-neutral segment in the concatenated carrier. Briefly there are two steps required to get higher success in this area. First one is, having nearly 100% neutral carrier phrases in the initial database and second one is, expanding the total number of segments in the database.

The quality of the expressive template can be described as the successful portraying of the desired emotion. If the acoustic characteristics of the desired emotion are not successfully presented on the expressive template by the speaker, the synthesised emotional speech can not reflect the desired emotion's characteristics.

Limited database not only caused the probability of having non neutral segments in the concatenated carrier, but also depending on the template sentence segmental structure, it caused some of the segments being repeated multiple types within the concatenated phrase. Because of the intelligent segment selection algorithm, the impact of the limited database on the results was higher than anticipated. This caused the nonsense language to sound less natural.

Also some of the listeners have feedback that, hearing a recognisable word as a segment in the carrier utterance caused them to focus on the potential meanings of the word and impacted the way they recognised the emotional affects. Even though having some words as segments in the database was a conscious decision to make the nonsense language more attractive for the children, this side affect was not considered during the initial design.



Chapter 5

5. Future Works

Even though for some synthesized emotional speech output samples very high success rates were achieved, there are improvement areas which could be worked further before the implementation of the corresponding application.

The future works suggested to improve the quality of the synthesized emotional speech output can be categorized in two areas; the improvements in both neutral and emotional databases and the improvements in the system algorithm.

In terms of the databases, the number of samples in each database can be significantly increased. This would have a very positive impact in the results achieved. Also having the neutral samples really neutral and emotional samples reflecting the desired emotion at a high success rate would increase the efficiency of the system developed.

In the software side, an intelligent segment selection technique has already been implemented in the system algorithm. But this technique needs to be advanced further to reach a natural sounding nonsense language. The cut points of the neutral sample segments used during this study were manually pinpointed through an interactive segmentation tool. A new algorithm to do the segmentation automatically can be included in the system to increase the efficiency.

Appendix A - INITIAL DATA FOR THE DATABASES

1. Using the Same Sentences

NEUTRAL:

“I will tell all that to my mom.”
 “Here are your badges and conference packages.”
 “The telephone did not ring too often today.”
 “Mom said the story was a lie.”
 “You have asked that question so many times.”

ANGER: *increased intensity, high pitch with dynamic changes, higher speech rate, strong stress/emphasize*

“I will tell all that to my mom.”
 “Here are your badges and conference packages.”
 “The telephone did not ring too often today.”
 “Mom said the story was a lie.”
 “You have asked that question so many times.”

HAPPINESS/JOY: *slightly increased intensity, increase in pitch, highest speech rate*

“I will tell all that to my mom.”
 “Here are your badges and conference packages.”
 “The telephone did not ring too often today.”
 “Mom said the story was a lie.”
 “You have asked that question so many times.”

SADNESS: *decreased intensity, as same pitch as neutral speech with no dynamic changes, slowest speech rate*

“I will tell all that to my mom.”
 “Here are your badges and conference packages.”
 “The telephone did not ring too often today.”
 “Mom said the story was a lie.”
 “You have asked that question so many times.”

FEAR: *low intensity, slightly higher pitch than neutral with no dynamic changes, slightly faster speech rate with pauses between words.*

“I will tell all that to my mom.”
 “Here are your badges and conference packages.”
 “The telephone did not ring too often today.”
 “Mom said the story was a lie.”
 “You have asked that question so many times.”

2. Using Different Sentences

NEUTRAL TEXT:

“This is the story of Albert Le Blanc. When Albert arrives in Mr. Jolly’s toy shop, the other toys think he is the saddest-looking bear they have ever seen. But then, Jack-in-a-box has an idea that might just put a smile on Albert’s face.

This is Albert Le Blanc. Even from the back he looks sad. His head hangs low. His shoulders are hunched. His arms flop loosely by his side.

From the front, Albert Le Blanc looks very sad indeed. He has the saddest eyes you ever saw. Which is strange, because Albert Le Blanc... But wait. Let me tell this story from the very beginning...

When Albert Le Blanc first appeared, sitting all by himself in Mr. Jolly’s toy shop, the other toys could only stare. He did look so sad.

- ‘Poor love’, said Sally the hippo. ‘We must try and cheer him up.’
- ‘You could do your funny dancing’, said Toby the cat. ‘That would make him laugh.’
- ‘My dancing is not funny’, said Sally. ‘It is beautiful and artistic.’

The other toys looked at each other and tried not to smile.

- ‘I know a joke’, said a little mouse called Pickle. ‘But I can’t remember the funny bit at the end.’

Everyone agreed that this could make the joke a lot less funny. It might even make it not funny at all. Pickle flopped down and stared at the floor.

It was then that Jack-in-a-box (who at that moment was not in his box) had an idea.

- ‘Why don’t we all try very hard to think of something happy and funny? Something to cheer up a very sad bear. Then we could put all our things together and make a show!’

It was a good idea. Everybody thought so. At first the toys sat quietly thinking. What could they say? What could they do? Something happy... Something funny... Maurice, the steam engine, let off a little steam as he tried to think. Lizzie, the humming top, hummed to herself as she thought.

.....”

ANGER: *increased intensity, high pitch with dynamic changes, higher speech rate, strong stress/emphasize*

“Silence, please! I keep trying to tell you! I am not unhappy! I am not sad at all. It is just the way I am made. I just happen to have a sad look on my face.”

"I don't believe this! That's the third time I came running for nothing! I think this big job has gone to that little boy's head!"

HAPPINESS/JOY: *slightly increased intensity, increase in pitch, highest speech rate*

“I can’t believe that! That is a present for me! I will play many games with him, go to school with him and even share my lunch with him. Oh Jimmy, that is great!”

“When we got to the Grand Canyon, I was speechless! I can't even explain how cool and beautiful it was! You would have to see it for yourself”

SADNESS: *decreased intensity, as same pitch as neutral speech with no dynamic changes, slowest speech rate*

“Oh! Silly me, I am sorry. I must have broken your heart by mistake! ”

“I'm really sorry I hit you when I was mad. That was wrong. I won't do it anymore, but please forgive me and give a small smile.”

FEAR: *low intensity, slightly higher pitch than neutral with no dynamic changes, slightly faster speech rate with pauses between words.*

“Listen! I heard a noise outside! The sound of a breath! Who is that? They may already know he was here! Once they find out he was here, believe me, they will kill you, me and everyone in this room.”

“It is getting dark! And we still don’t know where we are and how to go back. I see everywhere two looking eyes to us in the darkness! We must find somewhere to hide immediately!”

Appendix B - UTTERANCE EVALUATION

This is to evaluate the recorded utterances which are spoken by the same speaker, according to some evaluation criterions. The results will be used in the selection process of the research on emotion synthesis.

There is only one male speaker but two different coloured voice of him. The first one is his normal speaking tone and the other one is cartoon-like coloured tone.

The evaluation could be done according to four different criterions. The evaluator was able to listen the recorded speech by double clicking to the utterance at the beginning of each row and evaluate it for the following criterions:

- 1- To evaluate the colour of the voice: **“Does the utterance sound *HUMAN-LIKE* or *CARTOON-LIKE*?”**
- 2- To evaluate the emotion in the utterance: **“Which emotion do you feel in the utterance?”**
- 3- To evaluate the desired emotion in the utterance (if the desired emotion can be understood or not): **“To which emotion do you think the utterance is closest?”**
- 4- To evaluate the quality of the portraying: **“Does the utterance sound *REAL* or *IMITATED (=FAKED)*?”**

Thanks in advance for your time...

EVALUATOR PROFILE	
AGE:	
GENDER:	
OCCUPATION:	
MARITAL STATUS:	
DO YOU HAVE A CHILD?	

EVALUATION FORM TEMPLATE

UTT ERA NCE	Does the utterance sound human-like or cartoon-like?		Which emotion do you feel in the utterance? EMOTIO N FELT	To which emotion do you think the utterance is closest?						Does the utterance sound real or imitated (=faked)?	
	HU MA N- LIK E	CAR TOO N- LIKE		NEU TRA L	AN GE R	JO Y	SA DN ES S	FE AR	NO NE	REA L	IMIT ATED (=FA KED)
1 											
2 											
3 											
...											
...											

COMMENTS

Appendix C - FINAL DATABASES

1. Neutral Database

cr1: "I will tell all that"
cr2: "to my mom"
cr3: "the telephone did"
cr4: "not ring often today"
cr5: "you have asked"
cr6: "the question so many times"
cr7: "this is the story of anty"

2. Emotional Database

Anger

ang1: "You have asked that question so many times"
ang2: "Silence please"
ang3: "I keep trying to tell you"
ang4: "I am not unhappy"
ang5: "I am not sad at all"
ang6: "It is just the way I am made"
ang7: "It just happen to have a sad look on my face"
ang8: "I do not believe this"
ang9: "That is the third time I come running for nothing"
ang10: "I think this big job has done"

Joy

joy1: "Here are your badges and conference packages"
joy2: "When we got to the Grand Canyon I was speechless"
joy3: "I can't even how good and beautiful it was"
joy4: "You would have to see it for yourself"

Sadness

sad1: "You have asked that question so many times"

Fear

fea1: "I heard a noise outside"
fea2: "Who is that?"
fea3: "They may already know he was here"
fea4: "Believe me; he will kill you, me and everyone in this room"
fea5: "It is getting dark"
fea6: "And we still do not know where we are and how to go back"
fea7: "I see everywhere two looking eyes"
fea8: "We must find somewhere to hide immediately"

References

- [01] Daniel C. Dennett, "**Consciousness in Human and Robot Minds**", London, 1994
- [02] Werner Verhelst and Henk Brouckxon, "**Rejection Phenomena in Inter-Signal Voice Transplantations.**" In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, October 19-22, 2003.
- [03] P. Soens, W. Verhelst, "**Split Time Warping For Improved Automatic Time Synchronisation of Speech**", In Proc. of the first annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2005), Antwerp, Belgium, April 19-20, 2005.
- [04] W. Verhelst, D. Van Compernelle and P. Wambacq, "**A Unified View On Synchronized Overlap-Add Methods for Prosodic Modification of Speech**," in Proc. of ICSLP 2000 , Beijing, pp. II.63-II.66, October 2000.
- [05] Sami Lemmetty, "**Review of Speech Synthesis Technology**", Espoo, March 30, 1999.
- [06] Dennis Klatt, "**Review of text-to-speech conversion for English**" J. Acous. Soc. Amer. 82, 737-793, 1987
- [07] Ekman P, Friesen W, Ellsworth P. "**Emotion in the Human Face: Guidelines for Research and an Integration of Findings**", New York, 1972
- [08] Selma Yilmazyildiz, Wesley Mattheyses, Yorgos Patsis, and Werner Verhelst, "**Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication**", Brussels, 2006
- [09] Suthathip Chuenwattanapranithi, Yi Xu, Bundit Thipakorn, and Songrit Maneewongvatana, "**The Roles of Pitch Contours in Differentiating Anger and Joy in Speech**", in Transactions On Engineering, Computing And Technology V11, February 2006
- [10] Iain R. Murray and John L. Arnott, "**Synthesizing Emotions In Speech: Is It Time To Get Excited?**", Dundee, 1996.
- [11] Marc Schröder , "**Emotional Speech Synthesis: A Review**", Saarbrücken

- [12] Sieb Nooteboom, “**The Prosody Of Speech: Melody And Rhythm**”, Netherlands, 2001
- [13] E. Moulines, W. Verhelst, “**Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech**”, 1995
- [14] Wesley Mattheyses, Werner Verhelst and Piet Verhoeve, “**Robust Pitch Marking For Prosodic Modification Of Speech Using TD-PSOLA**”, Belgium, 2006
- [15] Werner D.E. VERHELST, “**A System for Prosodic Transplantation with Research Applications**”, Belgium, 1991
- [16] Werner VERHELST, “**Automatic Post-Synchronization Of Speech Utterances**”, Belgium, 2005
- [17] Verhelst, W., Borger, M.: “**Intra-Speaker Transplantation of Speech Characteristics. An Application of Waveform Vocoding Techniques and DTW**”. Proceedings of Eurospeech’91, Genova (1991) 1319–1322.
- [18] Mattheyses, W.: “**Vlaamstalige tekst-naar-spraak systemen met PSOLA (Flemish text-to-speech systems with PSOLA, in Dutch)**”. Master thesis, Vrije Universiteit Brussel (2006)
- [19] CHAPTER 3 M. Bulut, S. S. Narayanan, A. K. Syrdal, “**Expressive Speech Synthesis Using a Concatenative Synthesizer**”, in Proceedings of ICSLP, Denver, CO, 2002.
- [20] E. Douglas-Cowie, R. Cowie, M. Schröder, “**A New Emotion Database: Considerations, Sources and Scope**”, ISCA Workshop on Speech and Emotion, Northern Ireland, 2000.
- [21] Yilmazyildiz, “**Mevcut Iliskisel Veri Modellerini XML (Extensible Markup Language)'e Aktarmayi Saglayan XTABLES Teknolojisinin Incelenmesi, Avantajlarinin ve Potansiyel Kullanım Alanlarinin Belirlenmesi (Review of XTABLES Technologies that Creates XML Views of Existing Relational Data, in Turkish)**”, BSc. Thesis, Uludag University, Bursa, Turkey, 2004.

- [22] Mattheyses, W., Verhelst, W., Verhoeve, P.: **“Robust Pitch Marking for Prosodic Modification of Speech Using TD-PSOLA.”** Proceedings of the IEEE Benelux/DSP Valley Signal Processing Symposium, SPS-DARTS (2006) 43–46.
- [23] Verhelst, W.: **“On the Quality of Speech Produced by Impulse Driven Linear Systems”.** Proceedings of the International Conference on Acoustics, Speech and Signal Processing - ICASSP-91 (1991) 501–504.