# Emotional Speech Synthesis

*Gregor O. Hofer*

Master of Science

School of Informatics

University of Edinburgh

2004

# Abstract

The goal of this MSc project was to build unit selection voice that could portray emotions in various intensities. A suitable definition of emotion was developed along with a descriptive framework that supported the work carried out. Two speakers were recorded portraying happy and angry speaking styles, additionally a neutral database was also recorded. One voice was built for each speaker that included all the speech from that speaker. A target cost function was implemented that choses units from the database according to emotion mark-up in the database. The Dictionary of Affect [30] supported the emotional target cost function by providing an emotion rating for words in the target utterance. If a word was particularly emotional, units from that emotion were favoured. In addition intensity could be varied which resulted in a bias to select more emotional units. A perceptual evaluation was carried out and subjects were able to recognise reliably, emotions with varying amounts of emotional units present in the target utterance.

# Acknowledgements

I would like to thank my supervisors, Robert Clark, Korin Richmond, and Simon King for their valuable comments and support throughout this project. In addition I would like to thank all my collegues form the 2003/2004 MSc course for their companionship, in particular Andreas Vlachos and Tamara Polajnar for their help with programming. Finally I would like to thank my parents and my partner Daniela for their support and patience during this not always easy year.

# Declaration

I declare that this thesis was composed by myself, 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.

(*Gregor O. Hofer*)

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Motivation and Overview

This thesis concerns the process of building an emotional speech synthesiser using the Festival speech synthesis system developed at the University of Edinburgh. The goal was to build a unit selection synthesis system that can portray emotions with different levels of intensity. To achieve this, the system was based on theoretic frameworks developed by Psychologists to describe emotions.

The overall goal of the speech synthesis research community is to create natural sounding synthetic speech. To increase naturalness, researchers have been interested in synthesising emotional speech for a long time. One way synthesised speech benefits from emotions is by delivering certain content in the right emotion (e.g. good news are delivered in a happy voice), therefore making the speech and the content more believable. Emotions can make the interaction with the computer more natural because the system reacts in ways that the user expects. Emotional speech synthesis is a step in this direction.

The implementation of emotions seems straightforward at first but a closer look reveals many difficulties in studying and implementing emotions. The difficulties start with the definition of emotions. We will see in chapters two and three that there is no single definition of an emotion and that people have developed many concepts of what can be called an emotion. Also it is not at all clear what emotional speech sounds like,

meaning people can recognise emotional speech but they cannot describe it. Despite all this there have been various attempts in creating emotional speech synthesis.

Previous emotional speech synthesisers have focused mostly on doing signal processing to make synthesised utterances sound emotional. The work carried out during this project is based on unit selection synthesis which in general does not do signal processing. One of the inspirations for this thesis came from a paper written by Alan Black [4] on how to implement emotions in unit selection voice. He mentioned a technique called *blending* that was adapted for this project. This technique allows for mixing units from different speech databases. The technique was expanded to use psychological knowledge of emotions to produce a synthesised utterance.

The first part of this thesis is concerned with finding a definition of an emotion relevant for this project and identifying problems and mistakes previous researchers have encountered when building an emotional speech synthesiser. The second part of the thesis describes the process of building an emotional speech synthesiser; starting with the databases collection and ending with the modification of the synthesis algorithm. The differences between building a standard synthesis system and an emotional one are described and improvements are suggested. The last part of the thesis summarises the evaluation of the system and points out directions for further research.

## 1.2 Unit selection synthesis

The second part of this thesis describes the building and evaluation of an emotional unit selection voice. Although unit selection and the Festival speech synthesis system are further described in section 5.1, this section gives a short introduction to unit selection synthesis.

Unit selection synthesis generates speech by concatenating segments of speech or units found in a speech database. Usually the database consists of a few hours of speech recorded from a single speaker. The waveforms of the recordings are transcribed at the phone level with starting and end times of every phone. However units can have varying lengths, ranging from a single phone to a whole phrase. The unit selection algorithm selects the appropriate units from the database by minimising two costs. The

cost of how well a unit fits or the target cost, and the cost of how well two units join together or the join cost. The target cost is computed based on differences of features. A feature might be the position in a word. So, if a unit is to be used in a certain word, the unit's position, in the word it comes from, is compared to its position in the word it will be used in. The join cost is computed by calculating the spectral mismatch of two units. Thus, the larger the speech database is, the more units an algorithm can chose from, which makes it easier to find suitable units for a given sentence.

The quality of unit selection synthesis depends heavily on the quality of the speaker and on the coverage of the database. If no suitable units are found, the speech can sound very bad because there is almost no signal processing done to join two units. If suitable units are found, the synthesised speech can sound almost like the speaker. Ultimately the quality of the speaker determines how good the speech synthesis sounds, and this is even more true if unit selection is used for synthesising emotional speech.

# Chapter 2

# Background

It is not straightforward to define an emotion. There are many different perspectives on what we call an emotion. This chapter highlights some of the different meanings and traditions in emotion research while following closely the overview of emotion theories given by Marc Schrder in his PhD thesis [28]. Further, it described how emotion relates to speech and what the sources of emotional speech for research can be. When I refer to an emotion in this chapter, a full blown emotion is usually meant unless other wise specified. A full blown emotion incorporates all or nearly all of the aspects mentioned in the componential view.

## 2.1 Componential View

Researchers agree that emotions are not as often thought of, just a subjective experience or feeling. An emotion seems to be made up of several components:

- evaluation or appraisal of antecedent event, the meaning of the stimulus for the individual [20]
- physiological change, e.g. pupil dilation or blushing
- action tendencies, flight or fight patterns [16]
- subjective feeling [23]

- expressive behaviour such as non-verbal expressions including facial expression [15] and vocal expression [2]

By looking at this probably incomplete list of components it becomes clear that a given research project can only focus on one or two aspects that make up an emotion. Depending on the view one takes on emotion different components are highlighted in a study. The next section gives a short description of the different views.

## 2.2 Perspectives on Emotion

There are four basic traditions in emotion research in Psychology [10]. Each theory focusing on different components and making different assumptions on what is important for describing an emotion.

### 2.2.1 The Darwinian perspective

Charles Darwin in his book *The Expression of Emotion in Man and Animals* laid the groundwork for much of modern psychology and also for emotion research. He describes emotions as reaction patterns that were shaped by evolution. This implies that emotions are common in all human beings and also that some emotions might be shared with other animals.

The concept of basic emotions was also developed by Darwin. The function of the emotions may be a biological activation to make an animal more responsive to certain situations, including the tendency to perform certain actions. Another function might be a signal to an external observer, such as threat and therefore influencing the observers behaviour.

The universality of facial expression as found by Ekman [15] was an important finding in support of the Darwinian view. It makes clear that emotions have a biological basis and are therefore evolutionarily shaped. He demonstrated at least six emotions (happiness, sadness, anger, fear, surprise, and disgust) that were expressed in the face and recognised in the same way in many cultures.

### 2.2.2 The Jamesian perspective

For William James the body is essential for an emotion. Bodily changes follow some stimulus automatically and the emotion arises through the perception of these changes. Therefore without the perception of the body there are no emotions.

The facial feedback hypothesis follows the Jamesian perspective. It states that the facial expression of a person has an effect on the subjective emotional experience. For example if a person has a facial muscle configuration corresponding to a happy face the person reports feeling happier (e.g. smiling makes one happy).

### 2.2.3 The cognitive perspective

Cognitive emotion theories relate emotions to appraisal, which is the automatic evaluation of stimuli by low level cognitive processes. It determines how important a given stimulus is for the individual and increases the chances of an appropriate response.

Scherer's component process model [27] makes physiological predictions relevant to speech from such appraisal processes. The model details the appraisal process as a series of stimulus evaluation checks (SEC) in a certain temporal order: novelty check, intrinsic pleasantness check, goal/need significance check, coping potential check, and norm/self compatibility check. Each SEC is associated with an appropriate response in the various components of emotions (see Componential View). An emotion is denoted in the component process model as a configuration of SEC outcomes.

### 2.2.4 The social constructivist perspective

In the social constructivist perspective, emotions are seen as socially constructed patterns that are learned and culturally shared [10] [1]. Emotions have a social purpose that regulates the interaction between people. The expressions of emotions and the emotions themselves are described as culturally constructed. Although the biological basis of emotions is recognised, the socially constructed mechanisms are given more weight.

### 2.2.5 Discussion

Superficially these four perspective on emotions seem rather contradictory. Taking a closer look one can see that they are just different perspectives capturing different aspects of an emotion. The Darwinian tradition looks at the evolutionary context of emotions, the Jamesian tradition looks at the body in terms of emotions, the cognitive perspective looks at psychological phenomena concerning emotions, and the social constructivist perspective looks at the social function of emotion in wider context [10]. Taking this view on emotion theories has important implication on the descriptive system employed to describe an emotion.

## 2.3 Descriptive systems of Emotions

There are a number of descriptive frameworks of which only two are described here because of their relevance for this MSc dissertation. This section is only supposed to give a quick overview over the two main description systems used by most researchers.

### 2.3.1 Emotion categories

A common way of describing emotions is by assigning labels to them like emotion-denoting words. There are a number of researchers that have compiled lists of emotional words [30]. Of course some of these terms are more central to a certain emotion than others. Also different emotion theories have different methods for selecting such basic emotion words.

In a Darwinian sense the basic emotions have been evolutionarily shaped and therefore can be universally found in all humans. There is a Jamesian extension that expects to find specific patterns for the basic emotions in the central nervous system. The number of basic emotions is usually small. Ekman identified just six basic emotions by looking at the universality of facial expressions.

To describe different forms of basic emotions like hot and cold anger, finer grained emotion categories are used which the basic emotions are inclusive of. Scherer [27] suggests that an emotion is more general than another if its appraisal components form

a subset of the other emotion. So just ”anger” can be subdivided into ”hot anger” and ”cold anger” depending on the outcomes of particular SEC’s not specified for ”just anger”.

### 2.3.2 Circumplex Model of Affect

Instead of independent emotion categories, several researchers have described an affective space [23]. Russell has developed a circular ordering of emotion categories that makes it straightforward to classify an emotion as close or distant from another one. By having subjects rate similarity of different emotion words and converting the ratings into angles, a circular ordering emerged [22].

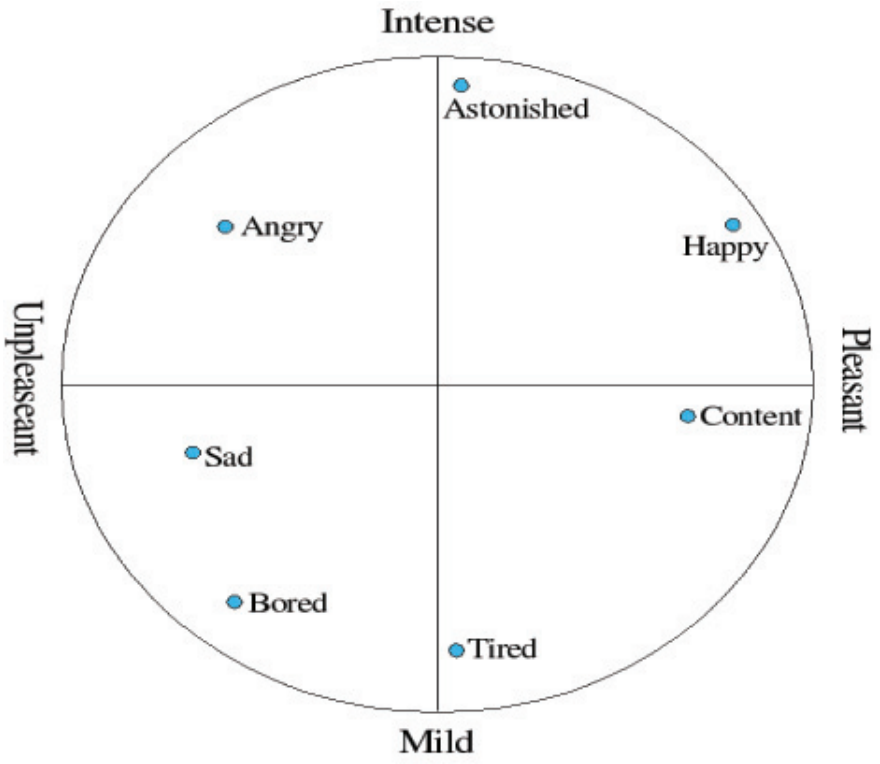


Figure 2.1: **Circumplex Model of Affect as described by Russell (1980)**

In addition to the circular ordering, Russell found evidence of two dimensions describing affect [23]. He called the two dimensions ”valence” and ”arousal”. These

terms correspond to a positive/negative dimension and an activity dimension respectively. Scherer found further evidence supporting two dimensions but proposed a different interpretation relevant to the component process model [26].

Some researchers suggested that using emotion dimensions gives an impoverished description of emotions [21]. Depending on the application, the use of emotion dimensions is often sufficient, sometimes even necessary. Especially the ability to measure similarity becomes very useful in computing applications where distance measures are used. The dictionary of affect that is used in the practical part of this MSc dissertation makes use of emotion dimensions.

## 2.4 Speaker vs. Listener centred

The expression and perception of emotion can be described via the Brunswikian lens model that Scherer adapted for research on vocal correlates of personality [25]. The version presented here has been further adapted by Marc Schrder [28].

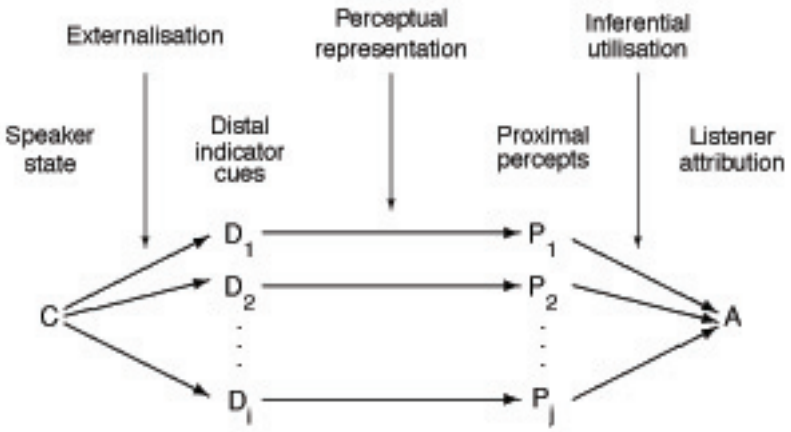


Figure 2.2: **Brunswickian lens model adapted by Schrder(2003) from Scherer (1984)**

Figure 2.2 reads like this. The speaker state C1 is expressed through the ”Distal indicator cues” which in the case of speech and emotion corresponds to the various acoustic parameters belonging to a certain emotion. The cues are perceived by the

listener as ”proximal percepts” which are the pitch and other voice parameters. The Attribution A is the perceived speaker emotion. This model gives visual support to different research orientations.

Studies can be speaker-centred or listener-centred. A speaker-centred study is interested in making a link between the speakers emotion and his vocal parameters. Scherer calls this an ”encoding study”. A listener centred study is interested in the perception of the listener, which Scherer further differentiates into inference studies and decoding studies. In inference studies, the distal cues are modified, while in decoding studies the speaker state changes. When acoustic parameters of speech are measured, the study is speaker-centred while perceptual tests of emotional speech are listener-centred.

## 2.5 Emotion in Speech

Vocal expression has been recognised as one of the primary carriers of affective signals for centuries. Darwin [12] in his pioneering monograph on the expression of emotions in man and animals underlined also this importance of the voice as an affective channel. In recent times studies have been undertaken to find the specific vocal patterns for certain emotions and further, how accurately listeners can infer emotions from the voice. Notably, Scherer [2] has done important work in his studies on acoustic profiles. But many other studies have been undertaken that examine the relationship of vocal expression and emotion. To examine the vocal correlates of emotions, one has to analyse a speech database. The source of the content of such a database has been widely debated [13].

### 2.5.1 Sources of Emotional Speech

To obtain authentic emotional speech data is one the biggest challenges in speech and emotion research. The goal is to have a closely controlled corpus with spontaneous speech. Because one cannot have spontaneous speech that is closely controlled, researchers have devised a number of strategies to obtain somewhat natural and spontaneous emotional speech data.

The most frequently used and oldest method is to have an actor portray certain emotions. This method has the advantage of control over verbal, phonetic, and prosodic speech content. Because only the emotion is varied in the actors' portrayal, direct comparisons of voice quality and prosody between different affective states are feasible. Another advantage is that it is easier to obtain expressions of full blown and extreme emotions. The problem with an actor-based approach is the ecological validity of the obtained data. Actors might portray only stereotypical expressions of emotions and their portrayal may differ from spontaneous expression in natural circumstances. Banse and Scherer challenge these criticisms on two grounds. Actors actually feel the emotion they are portraying and that natural emotion expression is also ”staged” because of the control of oneself required in different social contexts[2].

Nick Campbell [11] proposed a method similar to the actor approach. Instead of having the speaker read the same sentence in different emotions he had emotionally rich carrier sentences read which seemed to evoke genuine emotions. This idea was was further developed in the creation of the Belfast Structured Emotion Database [13] where different carrier paragraphs were written for the different emotions. The big disadvantage of this approach is that the text material will not be identical for different emotions, which makes a comparison very difficult.

The elicitation of authentic emotions in participants in a laboratory has been tried by number of researchers [17]. There are a few techniques that are termed mood induction procedures which can be used in different settings. Few studies in the field of speech and emotion have used these kinds of methods. Johnstone & Scherer [20] have used a computer game in a speaker-centred study. In the case of a computer game the subjects were asked to give a verbal protocol of their current emotional state while playing. This allowed the experimenter to vary variables through the computer game.

Recorded speech from spontaneous human interaction is the most natural but also uncontrolled. There are a few studies that are concerned with this kind of data. The Belfast Naturalistic Emotion Database is a collection of recordings from interviews and TV programs. Marc Schrder [28] analysed this corpus in his PhD thesis to find acoustic correlates for emotion dimensions, valence and arousal.

As described there are a large variety of methods for obtaining emotionally coloured

data. Different techniques are suitable for certain investigations. This MSc dissertation will focus on portrayal of emotions by actors. For the application of speech synthesis, carefully controlled voice parameters are needed which can be obtained relatively simply by having actors read carrier sentences in different emotions.

### 2.5.2 Acoustic Correlates of Emotions

During evolution speech was added to a ”primitive analog vocal signalling system” [28]. This means that the study of speech parameters expressing emotions is very complex. Acoustic parameters vary in their function of linguistic information carriers and non-verbal information carriers. Therefore it is not clear which parameters should be measured. Parameters like voice quality are important carriers of emotion in speech but are very difficult to measure. Therefore many studies have focused on measuring different aspects of F0 and intensity [2].

Scherer conducted an extensive experiment where actors portrayed 14 different emotions, varying in intensity [2]. It was found that vocal parameters indicated the degree of arousal but also quality aspects or valence. Other studies also found a strong correlation between vocal parameters and arousal, but there is very little evidence for correlation between valence and parameters.

Marc Schrder conducted an extensive analysis of the Belfast Naturalistic Emotion Database and relating the results to three emotion dimensions activation (arousal), evaluation (valence) and power [28]. He used the ASSESS system [9] for acoustic analysis of natural speech. The corpus was not homogenous as it consisted of recorded TV and radio clips and other recordings spoken by different speakers in different conditions. Despite this, his results also suggest a strong correlation between activation and vocal parameters. There were only weak correlations between valence and any voice parameters.

When people listen to speech they can clearly identify the emotion [2]. It is not clear why a correlation between valence and acoustic parameters has not been found yet. The information is present in the signal because people can identify the emotion just from speech. However, the problem seems to be that it is not clear what to measure. The variables used in studies so far do not seem to capture the essential properties of

valence. This problem of not being able to predict qualitative aspects of an emotional speech signal carries over to emotional speech synthesis.

# Chapter 3

# Emotional Speech Synthesis

This chapter explains the motivation of emotional speech synthesis. It also gives a short summary of what has been done before to implement emotions in speech synthesisers. It motivates the practical work undertaken for this project and gives reasons for certain design decisions.

## 3.1 Emotional speech synthesis

Natural sounding speech is the long term goal of research in speech synthesis. Speech synthesis systems have existed for quite some time now but are only applied in limited domain settings or in niche markets like screen readers for the blind. Wide-scale acceptance of full speech synthesis systems can probably only be reached with higher quality voices.

The measures *intelligibility* and *naturalness* are usually employed to describe the quality of synthetic speech. Major improvement in both these aspects have been taken place since the onset of the early rule based synthesisers. Modern unit selection voices are sometimes almost indistinguishable from natural speech, especially in limited domain settings. Still, compared to computer graphics which have been widely accepted by the movie industry, speech synthesis has a long way to go until it can replace the voice of an actor.

Telephone dialogue systems is a technology that depends heavily on speech synthe-

sis systems. In larger dialogue systems the text of a conversation might not be known beforehand and therefore requiring speech synthesis. In such systems the most important criterion is intelligibility but the believability of such a system depends heavily on naturalness. It also might be better to speak certain items in the appropriate emotional tone rather than in a flat tone. If good news are delivered a happy voice seems more appropriate than a flat voice.

One way to make speech synthesis more natural is to implement emotions into the voice. There is an argument that this is not necessarily important to have full emotional voices because speech synthesisers need to express not full blown emotions but things like certainty, friendliness, sense of urgency, or empathy when bad news are delivered. In this case full blown emotions might not be of much help. But if emotions can be modelled in varying degrees ranging from weak emotional states to full blown emotions, the expression of emotion will be able aid in the type of expressions mentioned above.

When modelling emotions in speech synthesis two main approaches have been identified. One can model a few emotional states as closely as possible. Recent work in unit selection has used this approach with three full blown emotions [19]. In this research a separate database was recorded for happy, angry, and sad tone of voice. Although this type of method almost guarantees good results, it is impossible to generalise the voice to other types of emotional expression, because the emotion is not explicitly modelled but just reproduced from the database. The other approach would be to build a general speech synthesis system that allows control over various voice parameters and therefore should be able to model almost any type of expression. As became clear from the literature on vocal correlates of emotion (see chapter 1) the exact acoustic parameters of an emotion are not known. Therefore such a system will be very difficult to build. The alternative approach taken in this MSc thesis is a mixture between the two methods mentioned above. The system built here is based on a unit selection voice, which does not explicitly model the prosodic realisation of an emotion, but can also model varying degrees of an emotion.

The goal in any emotional speech synthesiser should not be to model full blown emotions but to model emotions in varying degrees so to be able to express emotional

content more gradually. In a telephone conversation between a sales person and a customer full blown emotions are rarely used but a mild emotional colouring is quite frequent. Emotion dimensions seem to be the better choice than emotion categories to model varying degrees of intensity. With such an underlying framework, it seems to be easier to model a large variety of emotions in varying degrees.

## 3.2 Speech synthesis and emotion theory

The practical part of this MSc project explores the use of a dimensional representation of emotions in a speech synthesis system that can portray varying degrees of emotion. This section describes how the system relates to the theoretical background laid out in chapter 1.

To link the system to the different research traditions is not straightforward. There is is a minor link to the Darwinian perspective because I assume that the expression of emotion is universal. The emotions portrayed by the speech synthesiser are expected to be recognisable in every culture and therefore the emotional part of the system does not have to be adapted to work for Japanese for example. The system is not embodied and therefore the Jamesian tradition is not relevant to this research, however, it might become relevant if the system is used in conjunction with a talking head. The cognitive perspective is related to the system because of the emotion concepts that are evoked in the listener's mind. The social constructivist perspective is important because the aim of this project is to portray emotions and degrees of emotions relevant to everyday communication.

Emotional speech synthesis is always listener-centred because the aim is to evoke the intended perception in the listener's mind. Through the Brunswickian lens model (see Figure 2.2) this can be seen as synthesising the distal indicator cues so the proximal percepts evoke the intended attribution in the listener. The distal indicator cues are represented by a dimensional framework that is implemented in the speech synthesiser through the Dictionary of Affect [30] (see section 2.5 on Dictionary of Afffect).

## 3.3 Previous emotional synthesisers

Different synthesis techniques allow control over voice parameters in varying degrees. In most previous systems, only three to nine full blown emotions were modelled. This section includes a brief discussion of previous rule based systems and diphone synthesis systems. Unit selection systems are discussed in more detail because this technique is used for the system described in this MSc dissertation.

### 3.3.1 Rule based synthesis

Rule based synthesis, also known as formant synthesis, creates speech through rules of acoustic correlates of speech sounds. Although the resulting speech sounds quite unnatural and metallic it has the advantage that many voice parameters can be varied freely. When modelling emotions this becomes very interesting.

Cahn's Affect Editor [7] used DEC-talk as the underlying synthesis system. The acoustic parameter setting for each emotion were derived from the literature and implemented in the system. Burkhardt in his PhD thesis [5] employed a format synthesiser but took quite a different approach for generating the voice parameters. He used perception tests to find the best parameter settings for different emotions. Both systems were able to produce emotional speech but the output was crippled by the unnatural quality of the synthesised speech.

### 3.3.2 Diphone synthesis

Diphone synthesis uses recorded speech that is concatenated. A diphone stretches from the middle of one phone to the middle of the next. The diphone recording is usually made in a monotonous pitch and an F0 contour is generated through signal processing at synthesis time. This technique allows only limited control over voice parameters. F0 and duration can be controlled but control over voice quality seems to be impossible.

It is probably not enough to just vary F0 and duration to express emotion[28]. Nevertheless Copy synthesis [6] has been used to model emotion in diphone synthesis. F0 and duration are measured for each phone in an utterance from the portrayal of a given emotion by an actor and used for synthesis of the utterance from diphones.

The result is a synthesised utterance with the same duration and F0 as the original but the voice quality is based on the diphones. The naturalness of this technique is much higher than with formant synthesis but the usability of this technique for a large variety of emotions is questionable. It will be also very hard to model the intensity of an emotion with this kind of technique.

## 3.4 Emotional unit selection synthesis

Unit selection Synthesis is the most natural sounding speech synthesis technique today. It uses a cost function to select segments that vary in length from a large database of speech. The selected segments or units are concatenated to form the speech signal. When usually building such systems, the speech in the database should be consistent because segments from different parts of the database are expected to join seamlessly. In addition to the speakers distance from the microphone and loudness, the speaker is also asked to maintain a certain speaking style for all recordings. The output of the speech synthesiser will be largely consistent with the speaking style selected for the database. Signal processing after the concatenation of units is usually not conducted because of artefacts. This is also one of the biggest drawbacks of unit selection: most unit selection systems do not have an implementation that allows for modification of any acoustic parameters, therefore making it non-trivial to use in emotional speech synthesis.

Alan Black [4] identified two main techniques of building emotional voices for unit selection using different speaking styles. Separate voices can be built for each style or emotion and the synthesiser can switch between them. This technique can work well if there are well defined borders between the voice types and the voice is not too large. Limited domain synthesis ,where each style is for a particular domain, might be a good application for this. For example, one domain is to tell good news where the other domain tells bad news. Naturally, creating a large number databases for different emotions will take a long time and is not trivial. It might be sufficient and more efficient to have a general database and just a few important units recorded in certain styles.

The second method is called *blending*. With this technique all databases are combined into one voice, meaning that the voice can chose units from each database. The appropriate units are selected by a criterion that is based on the words or phrases being synthesised. For example, a command word would be more likely to be synthesised from command phrases in the database while other units might come from a more general part in the database [4]. Something similar to blending is the approach taken in this MSc thesis and will be described in more detail in Chapter 3.

Akemi Iida for his PhD has implemented an unit selection system that can portray three emotions: angry, happy and sad [19]. He recorded a separate database for each emotion from non-actors reading emotionally tainted stories. To synthesise an utterance in a given emotion, only units from the corresponding database were used. He found very high emotion recognition rates (60-80 percent).

Eide et al. [14] have implemented a system that uses a variant of the blending technique described above. The system did not portray emotion per se, but was an expressive speech synthesiser that could speak in three different styles: delivering good news (more upbeat), delivering bad news (more subdued) and asking a question. The researchers trained a prosody engine for each speaking style on speech from a professional actor and built a general unit selection voice from neutral speech from the same actor. When synthesising an utterance, a small number of units from the expressive speech were also included in the database. In addition, a cost matrix was constructed by hand that specified the cost of choosing a unit from a different expression than the one being synthesised. This matrix was employed when calculating the target cost. Although the matrix was not empirically verified and not motivated by previous research, very high recognition rates for the different speaking styles were found (70-85 percent).

Most work in emotional speech synthesis has been on creating full blown emotions or expressions with the notable exception of Marc Schrder [28] who did not use unit selection because of its limitations in modifying acoustic parameters but used MBROLA instead.

To implement a unit selection voice that can portray varying degrees of emotion is not trivial given the limitations of unit selection. The two options are to record

databases for each emotion in each intensity or to create a blending technique that is able vary the amount of emotional units selected for a synthesised utterance. This technique should of course be theoretically motivated, meaning the type and number of units should correspond to what is expected by the listener.

## 3.5 Dictionary of affect

This section describes in detail the Dictionary of Affect [30] because of its use as an outside knowledge source for the algorithm developed for the target cost function of the speech synthesiser. It motivates the type of units used for varying degrees of emotion.

The Dictionary of Affect was compiled by Cynthia Whissell et al. and it consists of 8742 words that are rated on two dimensions which corresponds to Russell's valence and arousal. The words come from other research, or are common English words with affective connotations [30]. Each word has an associated score along both dimensions. Subjects had to rate a number of words along these two dimensions. Not all subjects were presented with the same word list but each word was rated by at least 4 subjects. The mean for the arousal or activation dimension is 1.67 with a standard deviation of 0.36 and a range of 1 to 3, where 1 means not aroused and 3 means very aroused. The mean for the evaluation or valence dimension is 1.85 with a standard deviation of 0.36 and a range of 1 to 3, where 1 means unpleasant and 3 means pleasant.

Any verbal material can be scored in terms of the dimensions used by the Dictionary of Affect. Evidence for its reliability comes from a test-retest of a random sub sample. Evidence for validity comes from comparrissons of word scores with earlier studies [23] and from subjects that were presented with an imaginary situation. They were asked to give a 10 sentence description of their feelings in response to the situation. The sentences were correlated with the number of words included in the dictionary. The activation dimension has a lower reliability than the evaluation dimension. The dictionary works best when whole passages are scored where the mean score along each dimension is reported as opposed to single words. Also the reliability increases as the distance of words from the centre of the evaluation-activation space increases [30]. A sample of the dictionary is shown in Table 3.1. The version of the dictionary

presented also has a rating for imagery which gives a rating to a word of how easy it is to form a mental picture of that word.

Table 3.1 suggests that most words are very close to the centre of the rating scale for evaluation. It seems sensible that ”a” is in the centre but ”absent” is associated with a negative evaluation. The activation dimension is not as straightforward to interpret but was also found to be less reliable than the other dimensions. Certain words still seem to be more active than others; ”accelerated” is more active than ”abnormal”. As can be seen from the activation mean, most words are actually less active than the centre. All in all, the Dictionary of Affect has been validated many times and seems like a good choice for an knowledge source on the emotional connotation of words.

| word | evaluation | activation | imagery |
|---|---|---|---|
| a | 2 | 1.3846 | 1 |
| abandon | 1 | 2.375 | 2.4 |
| abandoned | 1.1429 | 2.1 | 3 |
| abandonment | 1 | 2 | 1.4 |
| abated | 1.6667 | 1.3333 | 1.2 |
| abilities | 2.5 | 2.1111 | 2.2 |
| ability | 2.5714 | 2.5 | 2.4 |
| able | 2.2 | 1.625 | 2 |
| abnormal | 1 | 2 | 2.4 |
| aboard | 1.8 | 1.875 | 2.8 |
| abolition | 1.5 | 2.1818 | 1.6 |
| abortion | 1 | 2.7273 | 2.6 |
| about | 1.7143 | 1.3 | 1.4 |
| above | 2.2 | 1.25 | 2.4 |
| abroad | 2.6 | 1.75 | 2.2 |
| abrupt | 1.2857 | 2.3 | 2.4 |
| abruptly | 1.1429 | 2.2 | 2.2 |
| abscess | 1.125 | 1.5455 | 2 |
| absence | 1.5 | 1.5556 | 2.6 |
| absent | 1 | 1.3 | 2 |
| absolute | 1.6667 | 1.4444 | 1.6 |
| absolutely | 1.6 | 1.5 | 1.8 |
| absorb | 1.8 | 1.75 | 2.2 |
| absorbed | 1.4 | 1.625 | 2 |
| absorption | 1.7778 | 1.6667 | 1.6 |
| abstract | 1.6667 | 1.4444 | 1.8 |
| abstraction | 1.4286 | 1.4 | 1.6 |
| absurd | 1 | 1.5 | 2.2 |
| abundance | 2.6667 | 1.5556 | 3 |
| abuse | 1.4286 | 2.5 | 3 |
| abusers | 1.25 | 2.7273 | 2.6 |
| abusing | 1.25 | 2.8182 | 2.6 |
| abusive | 1.6667 | 2.6667 | 2.4 |

Table 3.1: **First 40 entries in the Dictionary of Affect compiled by Cynthia Whissell.**

# Chapter 4

# Building an Emotional Unit Selection Voice

This chapter describes the process of building an emotional unit selection voice using festvox [3], which is a collection of scripts and tools that automate large parts of the voice building process. The initial approach taken was to build a separate voice for each emotion. This means that for each emotion a new database was recorded. In the end, a single voice with a combination of three databases was built for every speaker.

## 4.1 Speaker selection

Usually a speaker for a unit selection voice has to be able to speak in the same tone of voice constantly for long periods of time. In the case of building an emotional unit selection voice, the speaker should be able to portray an emotion for long periods of time. Both attributes are important for recording a emotional database. To determine the quality of a speaker, auditions were held. The following paragraphs describe the aspects that were looked at in determining which speaker to chose.

During the audition, the speakers had to read 10 sentences of about equal length that were printed on one single sided page. They were asked to read the sentences three times; once in a neutral voice, once in an angry voice, and once in a happy voice. They always read the neutral part first, then they could decide if they wanted to read

either the angry part or the happy part next. Recordings were made of the auditions in the studio where the actual database collection would take place. This was to see if the speaker felt comfortable in a studio setting.

The most important factor in determining a speaker's quality was his or her reading abilities. A speaker would need to read for a few hours, preferably without mistakes. It is very tiring on the speaker if he had to repeat sentences because he made a mistake. It was also important that the researcher liked the voice. He would have to listen to that voice for extensive periods of time during the voice building process. Another important factor was the quality of the voice. Since many aspects of the voice building process were automated, a ”cleaner voice” would get better results. A breathy voice would not work as well with the tools provided.

The previous attributes of a speaker apply to all unit selection voices, but in the case of emotional speech synthesis, it was also important that the speaker was able to portray emotions convince-ably. It is not clear if one should use actors to portray emotions or if amateurs are convincing enough. Both approaches were tried in this MSc thesis.

Ideally a number of people would have rated the recordings of the auditions according to the emotions portrayed to see how good the speakers were. But in this case the recordings were only listened to by the researchers to decide which speaker was the best.

## 4.2 Which emotions?

As can be seen from the quick review in Chapter 1 of the different traditions in emotion research, it is not trivial to define an emotion. Therefore it was also not trivial to communicate to the speaker which emotion he or she should portray. Because there is no way of empirically evaluating what has been said during the recording sessions, only limited cues were provided as to how the speaker should say a given sentence.

It was decided to use emotion categories for collecting the database instead of using a two dimensional approach. The reason for this was that most people are more comfortable with using labels to describe emotions. When describing emotions in

a dimensional framework, most speakers were not able to produce the correct emotion. This resulted in speakers being told to say a section in either a happy, angry, or neutral voice. The speakers were usually not provided with more detail. Only if the researcher could not detect any emotional colouring or opposite emotional colouring than intended, was the speaker told to repeat the sentence in the correct emotion. A categorical label like ”happy” is not exact but it seemed to convey enough information for the actor to produce the desired emotion.

## 4.3 Database collection

This section describes what a database for a unit selection voice should entail and how it is recorded. Two speakers were selected; one male speaker without acting training and a female speaker with acting training. Three databases in each of the emotions were recorded from each speaker.

### 4.3.1 Diphone coverage

During unit selection synthesis, variable sized units selected from a speech database are concatenated. The goal was to have a database that covers most prosodic and phonetic variation but at the same time is as small as possible. The script that the speaker would have to read during recording was what determined the actual coverage. Therefore it was of advantage to have a script that was written to have an optimal coverage.

Yoko Saikachi [24] in her MSc dissertation devised an algorithm to optimise the coverage of a speech database. She optimised according to the diphone coverage of the sentences in the database. The first 400 sentences of the script produced by the algorithm were used which gave enough coverage to have a fully functional voice.

The script was made up from newspaper sentences, which in my opinion, was not the best source. Many sentences in the script had grammatical errors. Also the script contained many headlines which are very hard to read without excessive intonation. The connotation of many of the sentences was far from neutral and therefore not really suitable. The script contained too many sentences about war and other human catastrophes. It also contained many foreign words and proper names for which the

pronunciation was very unclear. I was also asked many times by both speakers how to say a certain word. Some of the sentences in the script were very long and therefore very hard to read. Although I did not change the script, I told the speaker that they could omit sentences if they were too difficult. I also marked utterances that had foreign words other words that the speakers had problems pronouncing.

### 4.3.2 Recording procedure

All recordings were done in a soundproof recording booth. The speech was recorded directly into the computer and also onto DAT for backup purposes.

The speaker was asked to sit always the same distance from the microphone in the same position. When reading the script, the speaker would hear a beep if the sentence he just read was correct and no beep if the sentence was incorrect. The beep is a 7000 Hz tone for half a second and it served two purposes; it let the speaker know if he had to repeat the sentence and it told the computer were to cut the waveform.

The recording was done during the same time of the day for both speakers. The female speaker did three sessions during one week in the morning. The male speaker did three recording sessions during a week but in the afternoon. One recording session consisted usually of 400 sentences spoken in one emotion. Each sentence is one utterance and the session waveform had to be cut up into separate utterance waveforms.

One recording session for the female speaker lasted only about 1 hour 30 minutes to 2 hours where as the recording session for the male speaker lasted for about 3 to 4 hours.

### 4.3.3 Professional Actor vs. Regular Speaker

During the voice building process, it became very clear that trained speakers are more suitable for collecting an emotional speech database. The actress was more comfortable in the studio setting and therefore made less reading mistakes. This is reflected in the time difference for the recording sessions. The non-professional speaker needed about twice as much time for 400 sentences than the female speaker due to reading errors.

Conveying the intended emotion to the non-professional speaker was very difficult. The approach taken was to play him emotional speech to let him imitate the voice. Some sessions sounded more convincing than others. In genera,l he was very good in portraying the angry voice and not as good at portraying the happy voice. The speaking rate of the angry voice was very fast which made it harder to build the voice. The speaker told me that at points, he really felt a certain emotion which was reflected in the quality of the performance. Also, it was very hard for him to read sentences in a certain emotion if the connotation of the sentence was suggesting a different emotion. Since the script was made out of newspaper sentences, there were a few sentences about war. The speaker made noticeably more mistakes when he had to read these sentences in a happy voice.

The actress was able to speak with the same speaking rate for extended periods of time while making about 1 or 2 mistakes every 50 sentences. It was also straightforward to convey the intended emotion to her. She was told to read a session in a certain voice at a certain speaking rate, which was enough information for her to be able to convey the emotion convince-ably. The actress did not have any problems with the content of the sentences. In general, her angry voice was better than her happy voice.

There are several advantages in using an actor to portray emotions. The recording sessions with the actress were a lot quicker and her session waveforms sounded a lot more ”professional”. But it is unclear if her portrayal of emotions was natural. The recordings were clearly understandable as being in a certain emotion but they were not necessarily natural. Most of the recordings of the regular speaker did not sound very emotional, but when they did, it sounded natural. It seemed he really felt that emotion where as the actress’ portrayal did not convey the same feeling.

## 4.4 Labelling of the data

To synthesise speech from the database an utterance structure had to be created for each of the utterances. The utterance structure would consists of labels for segment, syllables, words, phrases, F0, intonation events and emotion. Because it is impractical to hand label all the data, a mostly automatic procedure was used to label the data.

### 4.4.1 Forced Alignment

The first step in labelling the data was to align the speech waveforms with the correct transcription. Automatic alignment can be seen as a simplified speech recognition task where the transcription of the speech data is known. A Hidden Markov Model (HMM) with a known sequence of phoneme models was generated. The Viterbi algorithm was used for finding the optimal phonetic alignment. Although the task of aligning a transcription with speech data is a somewhat different task from actual speech recognition, which has as its goal, word accuracy, and not segmental alignment, the HMM approach achieved high precision of alignment.

The alignment was carried out using the Hidden Markov Model Toolkit developed at Cambridge University. Depending on the speaker's accent, a different phone set was used. For the male speaker an Received Pronunciation (RP) phone set was used and for the female speaker the Edinburgh phone set was used. The speakers accents were determined by having them read sentences which were later judged to be of a certain accent. Parts of the labelling was checked manually to see if the system was successful and was corrected if necessary. In general, the accuracy of the labelling was satisfactory for all six voices.

After the alignment process, the waveforms had to be normalised so they had the same power ratios. In order to find the powerfactors needed for normalisation, the correct labelling was needed. It was suggested that after normalisation, the waveform should be realigned. This was done for all the voices.

### 4.4.2 Pitchmarking

Pitchmarking was required to generate the LPC coefficients needed for measuring the perceptual mismatch between units. Festvox has an implementation of an algorithm that extracts pitchmarks directly from the waveform. Finding the correct pitchmarking for a waveform was not straightforward and involved a lot of trial and error. The algorithm that performs the pitchmarking takes four parameters that determine the spacing of the pitchperiods:

- cutoff frequency for low-pass filter

- cutoff frequency for high-pass filter
- minimum pitch periods in seconds
- maximum pitch periods in seconds

To determine the pitchmarks, the speech was first low pass filtered to remove higher frequencies. Then it was high pass filtered to remove lower frequencies. Post processing was done to remove pitchmarks that were too close together. The minimum and maximum pitch periods determined which pitchmarks were removed. Also unvoiced regions in the speech signal were filled with pitchmarks. The optimal parameters for each voice were found by trial and error. The festvox manual suggested that by moving the pitchmarks to next highest peak in the waveform, the accuracy could be increased even more. This step was performed for all voices.

The pitchmarking was in general not as good for the male speaker as for the female speaker. It was also observed that the pitchmarking was more problematic in the angry voices for both speakers. When building the voice, festival reported the number of ”bad pitchmarks” meaning regions that do not have pitchmarks. This number was generally much higher for the male speaker. There were about 600 to 700 ”bad pitchmarks” for the male voices where as for the female voices, only 100 to 150 were reported. It is not clear why there is such a large difference between the speakers but one reason might be that the female speaker was a trained speaker with acting training where as the male speaker was an amateur. By just listening to the voices, the female speaker was more constant in speaking rate and pitch than the male speaker. Another explanation would be that the algorithm handles female data better because females tend to speak in a higher pitched voice. The last statement is contradictory to what has previously been found and therefore it seems that the algorithms developed to find pitchmarks might have to be rewritten for emotional speech.

### 4.4.3 Generate utterances

Once the the speech data was aligned and the pitchmarks were extracted, the utterance structure could be built for each of the voices. During this process, a specific lexicon

was used for each speaker that had the phonetic transcriptions for each word that was in the script. The phonetic transcription can vary according to the speaker's accent. As stated earlier the female speaker had an Edinburgh accent and the male speaker an RP accent. The differences between these two accents are not discussed here. There is enough difference between the pronunciation of certain words that a different lexicon had to be used for each of the speakers.

Festival also uses the pitchmarks to build the utterance structure and marks the segments "bad" that don't have any pitchmarks. Segment duration information and F0 contours are also added to the utterance. The information is stored in an data structure that was specifically developed to store this kind of linguistic information. The utterance data structure is explained in detail in the next chapter. Once an utterance structure is built for a voice, festival has access to the data and can use it to synthesise text into speech.

At the end there were three databases of utterances from each speaker. Festival allows for voice definitions to include more than one database. So for each speaker one voice definition was written that included three databases.

# Chapter 5

# Emotional Festival

The Festival speech synthesiser was developed by Alan Black and Paul Taylor at Edinburgh University. This chapter describes the technical details of how Festival can be employed to function as a emotional speech synthesiser. It gives a general overview of the synthesis procedure. This follows a detailed description of the underlying data structure of the synthesiser and a description of how the target cost is computed meaning how a unit is chosen from the database. The last part of this chapter is devoted to describing the changes that had to be made to the standard system for emotional speech synthesis.

## 5.1 Synthesis

The current version of Festival uses the multisyn unit selection algorithm developed by Korin Richmond, Robert Clark, and Simon King at the University of Edinburgh. The algorithm works by predicting a target utterance structure. Then, suitable candidates from the database are proposed for each target unit. The best candidate sequence is found by minimising target and join cost. The unit size used is diphones, larger sized units are not explicitly used, but the selection is performed by search where the join cost for adjacent units in the database is zero which effectively favours longer units.

The target utterance structure is predicted from text. Unit selection has the advantage that a lot of information that is necessary for other synthesis techniques does not

have to predicted. Properties like segment durations and prosody can be safely omitted because this information is inherently present in the units that are selected at synthesis time. Therefore only phrasing and pronunciation are predicted. The resulting target utterance is a sequence of phones with an appropriate linguistic structure attached.

The target phone sequence is converted into target unit sequence. A list of candidates for each of the units is constructed from the database. The list contains all diphones of the target type in the database. In the case the inventory does not contain a specific diphone a backing-off mechanism is used. The backing-off rules substitute missing diphones with existing diphones. For example, a full vowel is substituted for a missing reduced vowel.

After the list is complete and the missing diphones are substituted, a standard Viterbi search is used to find the diphone sequence that best matches the target sequence. The search reduces the sum of all target costs and join costs.

## 5.2 Utterance Structure

Any text to speech system needs to store many different types of linguistic information. The data-structure that stores this information is supposed to allow for a straightforward implementation of speech synthesis algorithms, therefore having straightforward access to its contents.

Festival uses a data structure called an utterance structure that consists of relations. Taylor et al. [29] define a relation as "...a data structure that is used to organise linguistic items into linguistic structures such as trees and lists. A relation is a set of named links connecting a set of nodes." Depending on how the data is stored, the direction of the link changes. Lists are linear relations where each node has a previous and a next link. The nodes in trees on the other hand have up and down links. Nodes don't carry any information by themselves but have, apart form links to other nodes, a link to an item which contains the information.

The information in items is represented as features, which are a list of key-value pairs. An item can have an arbitrary amount of features including zero. A feature is usually a string or a number but can also be a complex object or a feature function. A

feature function reduces redundancy in the utterance structure. Instead of storing each value in a feature, a feature function computes its values from the information stored in other functions. This helps greatly when timing information is needed because the end and start times of items are often the same, and therefore only need to be explicitly stored once. It also makes changes in the data easier to implement because the feature function calculates its value new every time it is accessed.

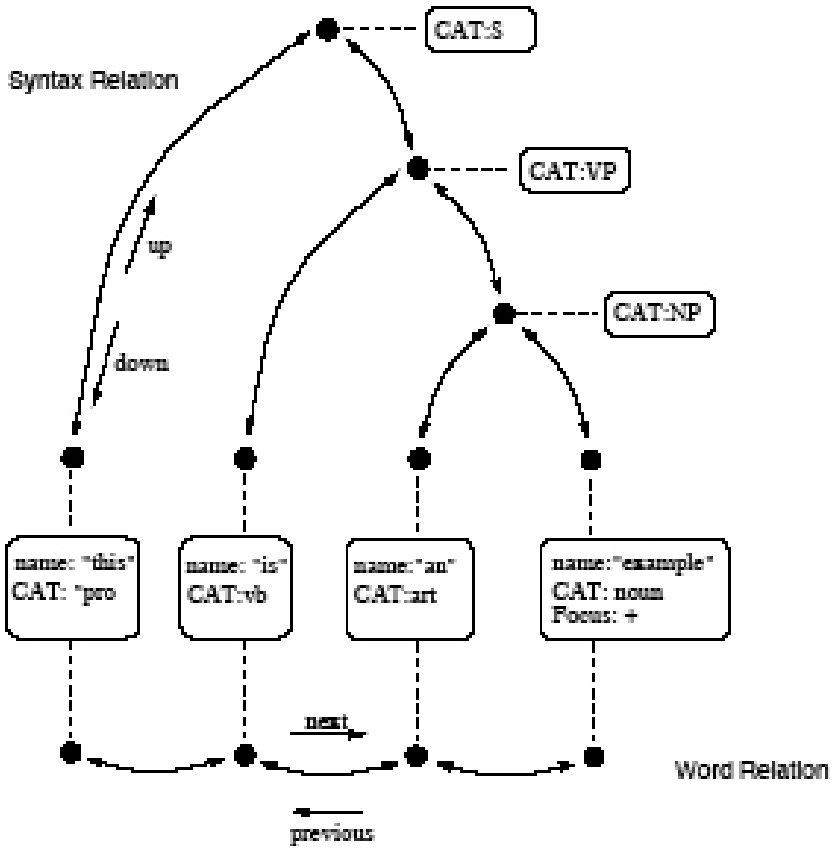


Figure 5.1: **From Taylor et al. (1998). The example of an utterance shows the word relation and the syntax relation. The syntax relation is a tree and the word relation is a list. The links are connecting the black dots. The items which contain the information are shown in the rounded boxes.**

Figure 5.1 shows an example of an utterance structure. An utterance is made up of a set of items each containing various features. Relations are structures that consist

of nodes which are linked to an item. Different nodes can link to the same item. In Figure 4.1, the word relation has four nodes represented by the black dots. Each node has a link to an item represented by the curved box with various features, stating the name and word category. The same items are linked to by the syntax relation which is a tree structure. This relation has some more nodes with their own items denoting the syntactic category.

Both the utterances in the database and the to be synthesised utterance are stored in utterance structures. This allows for a straightforward comparison between the features of the utterances in the database and the target utterance.

## 5.3 Target Cost

During synthesis time, the challenge is to find the units in the database that best match the target utterance. Two measures are used to find the best units. The join cost gives an estimate of how well two units join together and the target cost describes how well a candidate unit matches the target unit. The following formula denotes the target cost as proposed by Hunt & Black [18]:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} {w_j}^t {C_j}^t(t_i, u_i) \tag{5.1}$$

The target cost is a weighted sum of functions that check if features in the target utterance $t_i$ match features in the candidate utterance $u_i$ and is computed for each unit in the candidate list. Each function returns a penalty if a feature does not match or if a penalty feature is set in the candidate. The standard target cost proposed by CSTR involves the following computations of differences:

- stress cost: is target stress equal to candidate stress?
- position in syllable cost
- position in word cost
- position in phrase cost

- POS cost: does the candidate have the same POS tag as the target?
- left context cost
- right context cost
- bad duration cost: if the candidate unit has a bad duration feature a penalty is added
- bad f0 cost: if the candidate unit has a bad f0 feature a penalty is added

Both the *bad f0* and *bad duration* function are weighted the highest, probably because a bad unit can reduce the quality of the synthesis substantially. The *stress cost* is also set relatively high followed by the *position in phrase cost* because they can really make a difference in intonation. The context costs and the other position cost are weighted lower. The ratio of the weights of the target cost functions changed when the emotion target cost function was added. The ratio of the weights of the target cost functions, before and after the emotion target cost has been added can be seen in Table 5.1.

| **target cost function** | **ratio before emotions** | **ratio after emotions** |
|---:|:---:|:---:|
| stress | 0.13 | 0.09 |
| position in syllable | 0.07 | 0.05 |
| position in word | 0.07 | 0.05 |
| position in phrase | 0.09 | 0.07 |
| POS | 0.08 | 0.06 |
| left context | 0.05 | 0.04 |
| right context | 0.04 | 0.03 |
| bad duration | 0.13 | 0.09 |
| bad f0 | 0.33 | 0.23 |
| emotion | 0 | 0.29 |

Table 5.1: **Ratio of weights of the different target costs before and after an emotion target cost function was added**

## 5.4 Emotion Markup

The standard Festival mode is to just do text to speech meaning that everything that is inputted will be synthesised. In order to communicate with festival and tell it to synthesise something in a certain way, a special text markup is used. Festival supports various markup modes. The most widely used ones are OGI markup, Sable markup, and APML.

OGI and Sable markup are part of the text-mode which supports input from a text file with tags. The Text mode in Festival supports any implementation of a text mode. This means that text modes can be implemented not only for OGI or Sable, but also for other purposes. It lets you define what should be said and how it should be said. For example there could be an HTML implementation that reads different parts of a web-page in different voices with different emphasis, and lets you know the structure of a table or tells you that there is link to another web-page. OGI and Sable are already implemented in the text mode and therefore the obvious choice for markup. The difference of OGI markup and Sable markup is the tags supported. OGI tags are straightforward to use but do not support complex tags, where you can assign a value to a tag. Also OGI files do not use a dtd file, therefore there is no external structure imposed on the tags. Any OGI tag must be enclosed in "<" and ">" and needs to be specifically implemented in the OGI file. This means that $< 1 >$ is not read as a number with the value 1 but as a tag "1". The OGI markup implementation was changed so it supported <neutral>, <angry>,<happy> and switched voices accordingly.

Sable on the other hand is XML based and therefore has a more complex structure than OGI. It uses a dtd file for evaluating the structure of the input. The number of tags supported by Sable is already substantial. It supports complex tags where a tag can be assigned a value. So it would parse $< 1 >$ as a value and not as a "1". For emotion markup an EMOTION tag was added that could take three values: neutral, angry, and happy.

The problem with text mode is that it produces more than one target utterance. It splits the input into a number of utterance structures. This happens at an intermediate level where the interpreter has no access to the utterances. Festival returns only a waveform and not the corresponding target utterances when text mode is used. This

makes it very complex to analyse the synthesis process since the intermediate steps are hidden from the user.

The APML markup language , which was originally developed to drive conversational agents, is not a text mode but was separately implemented in C++ to support APML tags. The implementation supports complex tags and also uses a dtd file for structure. An Emotion tag was added that can take the values: neutral, angry, and happy. Festival returns both an utterance structure and waveform when it parses APML input which makes debugging and evaluating the system more straightforward.

In the final version of the system, only APML markup was used because it gave more control to the user than the text mode.

## 5.5 Utterance structures with Emotional information

For Festival to take emotion information into account in the target cost, the utterance structures of both the utterances in the databases and the target utterances had to be changed to carry emotional information. This section describes in detail how this was accomplished for the target utterance and the database.

### 5.5.1 Emotion Feature

The speech database was also marked according to the emotion a particular utterance was spoken in. An emotion feature that can have the values neutral, angry or happy was added the items in the word relation. Figure 4.2 gives a graphical perspective of the modified word relation.

### 5.5.2 Emotion Relation

When Festival parses input with emotion tags, it stores the tags in the target utterance. It was first suggested to just add a an emotion feature to one of the relations in the utterance structure but as it turned out this was not feasible. The information flow between the different parts of the system is very complex, one of reasons being that

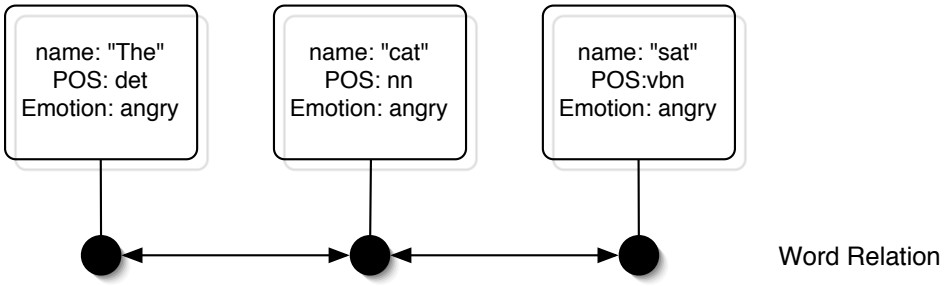


Figure 5.2: **Candidate Utterance: The emotion feature is part of the items in the word relation.**

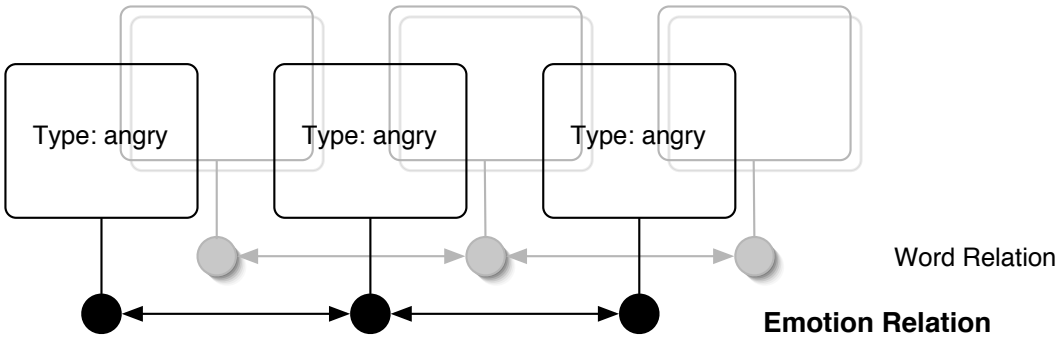


Figure 5.3: **Target Utterance: A separate Emotion Relation is added to the utterance.**

the implementation is done both in C++ and Scheme. This makes is not trivial to implement new information in the utterance structure.

The APML markup implementation was changed so it would add a new Emotion relation connected to an item with a Type feature to the target utterance. This relation can be seen as a separate layer since it does not share any items with the other relations. Figure 4.3 gives a graphical perspective of the modified utterance structure.

Of course it was not intended to add a separate relation just for one feature but this was a way of circumventing the technical complications that other methods would have caused. It was a straightforward modification that could be done to the target utterance so that the target cost functions have access to an emotion feature.

## 5.6 Emotion Target Cost Function

Instead of having a separate voice for each emotion, a target cost function was added to the target cost to find the most suitable units for synthesising a specific emotion. A new target cost class was derived from the standard CSTR target cost. It included the new emotional target cost.

Instead of just comparing the emotion feature of the target utterance to the candidates, an outside source of knowledge is applied to find the most suitable units. It was hypothesised that only certain words are really important to be synthesised from an emotional database where as others can be synthesised from the neutral database. Whissell's Dictionary of Affect [30]is consulted during the selection algorithm to determine which words should be synthesised in a given emotion. If a word was marked in the dictionary as belonging to a certain emotion, the target cost will favour units that belong to the database of that emotion. This means that if the target utterance is marked up as being ”happy”, words that are associated with happy have a very high chance of being synthesised from the happy database. In general, neutral units are favoured. This means that if a word is not in the dictionary, or belongs to a different emotion category, the chances of it being synthesised from an emotional database are very small.

Since the dictionary of Affect uses dimensional ratings for its words but the system uses categorical emotions for markup a suitable translation scheme between the two representations was developed. Each word in the dictionary has a rating for activation and evaluation. These ratings are translated into a categorical emotion information. A ”happy” word would have an evaluation of higher than 2 where as an ”angry” word would have an evaluation of lower than 2. These translations were based on the Russell's circumplex model [23] (see Figure 2.1) where evaluation is called valence and activation is called intensity. The origin of this model corresponds to 2 in Whissell's model. A word that has an evaluation of less than 2 might be considered a ”negative” word and respectively a word that has en evaluation of more than 2 might be considered a ”positive” word. Happy, in general, is seen as a positive emotion and angry, in general, is seen as a negative emotion. The value of the evaluation dimension was associated with either happy or angry based on it being positive or negative. The ac-

tivation dimension was also considered in the translation. The position of happy and angry in the cirumplex model suggests that happy and angry have a higher than origin activation. In general, happy and angry are seen as active emotions. Therefore, it was decided to favour words that have a high activation in the target cost.

The exact algorithm that computes the emotion target cost is described in Algorithm 1. The algorithm, in general, favours units that are neutral, but if the target word is happy, it will favour candidate units that are happy and vice versa for angry. Depending on the target word and on the candidate unit, the function returns different penalties. In Algorithm 1 uses three different kind of penalties; small, medium, and large. The best values for these penalties are not specified yet but a few sets have been evaluated.

In general, the amount of emotional units used in the synthesised utterance depends not only on the weights and penalties of the emotional target cost but also on the words used. If a particular sentence consists of many words that have a happy connotation, the resulting synthesised utterance will be more likely to include happy units because the happy units are favoured for the ”happy” words. The challenge was to find a suitable set of penalties that allows for varying intensity with mixing units from the different databases.

## 5.7 Description of Voices

Three sets of penalties were implemented which resulted in voices that used varying amounts of emotional units. This section describes the differences between these voices and gives an informal description of the voices.

The main difference between the voices was the percentage of units from different databases were included in a synthesised utterance. With the first set of penalties utterances consisted usually only of units from one database. Of course with this database the quality of the synthesis was the best because there were very few joins between units from different databases. Also the the emotions sounded the strongest with this set of penalties.

The next set of penalties was aimed at keeping the balance of units about 50/50

```
cEmo ⇐ emotion of candidate unit
tEmo ⇐ target emotion
tWord ⇐ unit is in this word in the target
ee ⇐ evaluation of target word
aa ⇐ activation of target word
if tWord in dictionary AND tEmo = cEmo then
  if (tEmo = angry AND ee < 2) OR (tEmo = happy AND ee > 2) then
    if aa > 2 then
      return no penalty
    else
      return small penalty
    end if
  else
    return medium penalty
  end if
else if tEmo = neutral AND cEmo = neutral then
  return no penalty
else if cEmo = neutral then
  return small penalty
else if tEmo = neutral AND cEmo ≠ neutral then
  return big penalty
else
  return medium penalty
end if
```

**Algorithm 1:** The Emotional Target Cost Function

between the neutral database and an emotion database. If the target word was in the Dictionary of Affect and had an emotional connotation that corresponded somewhat to the target emotion, chances were high that i would be synthesised from emotional units. With this the synthesis quality was generally worse than with the previous set but there were a few instances when it improved the synthesis quality. Sometimes, units that were missing in the neutral database or were corrupted, were present in the emotion databases and this improved the quality of the synthesis. The emotions did not sound as strong with this set which was intended.

The final set of penalties was designed to use more neutral units by using the minimum amount of emotional units possible. Emotional units were only used if the the target word had very emotional attributes in the Dictionary of Affect. The synthesis quality was very similar to the previous set, but it sometimes was very difficult to perceive an emotion. The emotion effect was very weak with this set of penalties.

It was attempted to modulate the intensity of emotions by varying the amount of emotional units used in an utterance. A web-page[1] with examples of synthesised utterances was created and a formal evaluation was conducted to asses the accuracy of the synthesised emotional speech.

[1] http://homepages.inf.ed.ac.uk/s0343879

# Chapter 6

# Evaluation

The accuracy of the developed emotional synthesis was assessed in a perceptual test. It was decided to do a perceptual study because there is no clear evidence (see Chapter 2) of which acoustic parameters are important for a given emotion. Only the voice of the actress was evaluated, but it is planned to evaluate the male voice in the future.

## 6.1 Method

To evaluate the system three versions of the emotional unit selection algorithm were implemented with different values for the penalties returned by the emotion target cost function. One version was set so it used only units from the right emotion database (full emotion). The next version was set so it favoured units from the right emotion database, but not exclusively (half emotion). The last version was set so it rarely favoured units from the emotion database but to use more neutral units instead (some emotion).

Four carrier sentences were synthesised with the three implementations. A neutral version of each sentence was also included for control. The exact number of each type of unit and the number of joins for each sentence can be seen in the Tables 6.1-6.4. The problem was that with the limited coverage of the databases, the content of the carrier sentences had to stay very close to the script. There was an effort to keep the number of joins constant for one sentence because too much variation in the number of joins

could introduce a lurking variable in the evaluation. The total number of synthesised sentences was 28.

It was very difficult to generate a satisfactory set of carrier sentences because the target cost cannot be forced to chose certain units and to achieve a constant number of joins. A large number of sentences was created by mixing and matching different parts of the original script. Each sentence had to be synthesised with the various voices to get a valid estimate of the number of joins, and of how good the synthesis was. The goal of the evaluation was not to assess synthesis quality and therefore only examples were chosen that sounded good across all the conditions. From these sentences the four carrier sentences were chosen.

The actual experiment took place on a computer where subjects had to rate the synthesised sentences. The rating was done using a continuos scale represented by a slider bar that was labelled angry on the left side, happy on the right side and neutral in the centre. The program recorded the input on a scale from 0 to 100 where 50 meant neutral, 0 meant angry, and 100 meant happy. It is not correct to say that happy and angry are opposite of each other but for the purpose of this experiment it should not had any effect on the outcome. This type of rating scale had the advantage that it is forced choice with a continuos response. A continuos response was needed because the strength of a given emotion was part of the assessment. Each subject had to listen to three blocks of 28 sentences over headphones. Each block was randomised in itself. There was no way of repeating a just heard sentence. The instructions were given on paper and there was no time limit on the experiment.

## 6.2 Results

The experiment had 13 participants from which 4 were female and 9 were male. There were 6 native English speakers and 7 non-native speakers. All of them were post-graduate university students between the age of 22 and 35.

The descriptive statistics of the results are shown in Table 6.5 and in Figure 6.1. A one-way ANOVA (Analysis of Variance) was carried out to measure difference between the ratings for each condition. A Fisher's LSD (Least Significant Difference)

| emotion | joins | neutral | angry | happy |
|---|---|---|---|---|
| neutral | 12 | 58 | 0 | 0 |
| full angry | 15 | 0 | 58 | 0 |
| half angry | 14 | 27 | 31 | 0 |
| some angry | 12 | 50 | 8 | 0 |
| full happy | 11 | 0 | 0 | 58 |
| half happy | 11 | 30 | 0 | 28 |
| some happy | 12 | 39 | 0 | 19 |

Table 6.1: Sentence one:“Most people will now build large housing complexes in Dijon and not in Worcester.”

| emotion | joins | neutral | angry | happy |
|---|---|---|---|---|
| neutral | 21 | 48 | 0 | 0 |
| full angry | 18 | 0 | 48 | 0 |
| half angry | 21 | 23 | 25 | 0 |
| some angry | 22 | 34 | 15 | 0 |
| full happy | 19 | 0 | 0 | 48 |
| half happy | 19 | 21 | 0 | 29 |
| some happy | 19 | 25 | 0 | 25 |

Table 6.2: Sentence two:“My former tutors at Edinburgh were taking this action to clean the oil.”

| emotion | joins | neutral | angry | happy |
|---|---|---|---|---|
| neutral | 10 | 58 | 0 | 1 |
| full angry | 10 | 0 | 59 | 0 |
| half angry | 13 | 26 | 31 | 2 |
| some angry | 11 | 45 | 12 | 2 |
| full happy | 11 | 0 | 0 | 59 |
| half happy | 10 | 31 | 0 | 28 |
| some happy | 9 | 51 | 0 | 8 |

Table 6.3: Sentence three:“His wife Zoe dressed in Prada and Gucci while studying Zoology at University.”

| emotion | joins | neutral | angry | happy |
|---|---|---|---|---|
| neutral | 19 | 32 | 0 | 1 |
| full angry | 18 | 0 | 33 | 0 |
| half angry | 18 | 17 | 15 | 1 |
| some angry | 18 | 21 | 11 | 1 |
| full happy | 16 | 0 | 0 | 33 |
| half happy | 18 | 14 | 0 | 19 |
| some happy | 19 | 23 | 0 | 10 |

Table 6.4: Sentence four:“Tomorrow I will get up and dance tango with you.”

comparison was carried out to measure individual differences between the mean ratings of the emotions. The results are summarised in table 6.6.

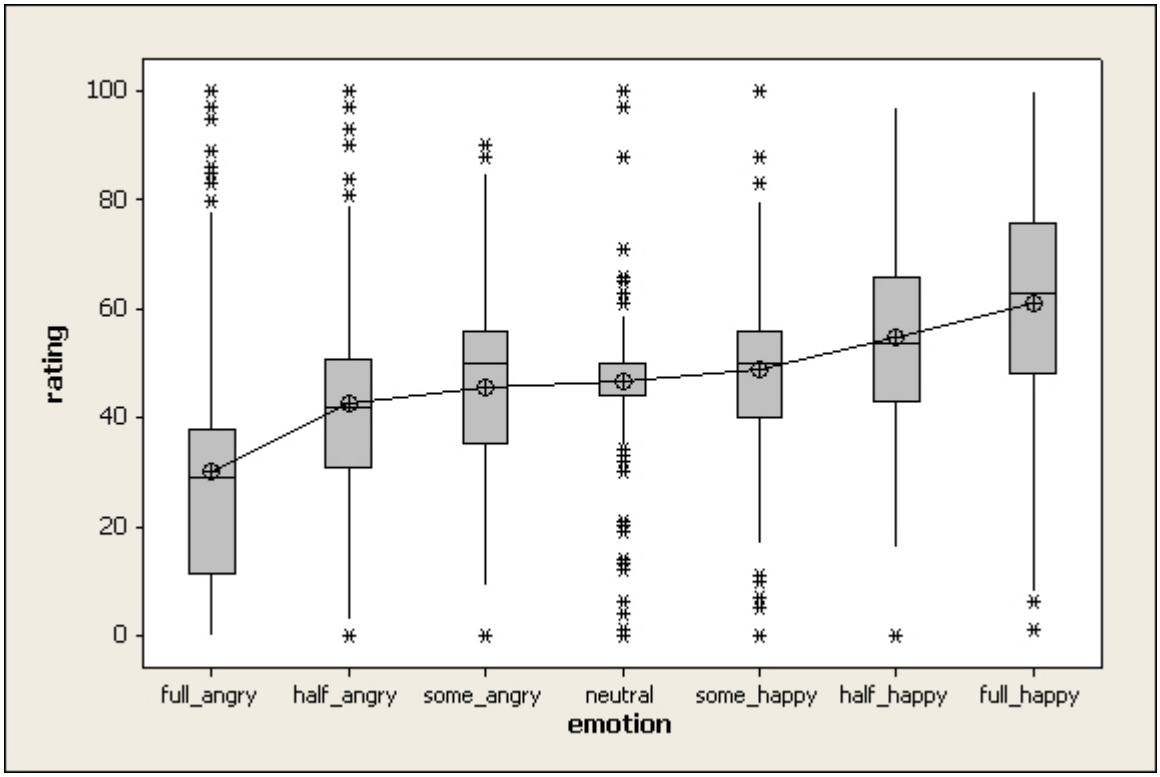


Figure 6.1: The means and variance of the ratings according to the emotion. Outliers are marked with *

A significant difference between the means was found (F=39.79, p=0.00). The Fisher comparison found significant difference between the neutral condition and both of the extreme emotions. A significant difference was also found between the half-happy condition and the neutral condition. The half angry condition was marginally not significantly different from the neutral condition. There were also significant differences between the extreme emotion conditions and the other conditions.

A two-way ANOVA was carried out to estimate the interaction between sentence type and emotion with the interaction being significant(F=10.39, p=0.00). A one-way ANOVA with a Fisher's LSD was carried out for rating and sentence, and a significant difference was found between sentence four ($\mu_4 = 41.54$) and the other three sentences

| emotion | rating mean | standard deviation |
|---|---|---|
| neutral | 46.6 | 15.00 |
| full angry | 30.27 | 24.14 |
| half angry | 42.79 | 20.45 |
| some angry | 45.45 | 16.62 |
| full happy | 61.01 | 22.03 |
| half happy | 54.78 | 17.32 |
| some happy | 48.78 | 16.58 |

Table 6.5: Descriptive Statistics of the Evaluation

| emotion | lower | center | upper | |
|---|---|---|---|---|
| full angry | -20.58 | -16.33 | -12.08 | * |
| half angry | -8.05 | -3.80 | 0.45 | |
| some angry | -5.40 | -1.15 | 3.10 | |
| full happy | 10.17 | 14.42 | 18.67 | * |
| half happy | 3.94 | 8.19 | 12.44 | * |
| some happy | -2.06 | 2.19 | 6.44 | |

Table 6.6: Results of Fischer's LSD for neutral vs. emotions. If the confidence interval does not include 0 then there is a significant difference between two means. Significant differences are marked with *

($\mu_1 = 50.99, \mu_2 = 46.92, \mu_3 = 48.95$). No other significant difference was found for the remaining sentences. Figure 6.2 gives the descriptive statistics for the ratings according to sentence.

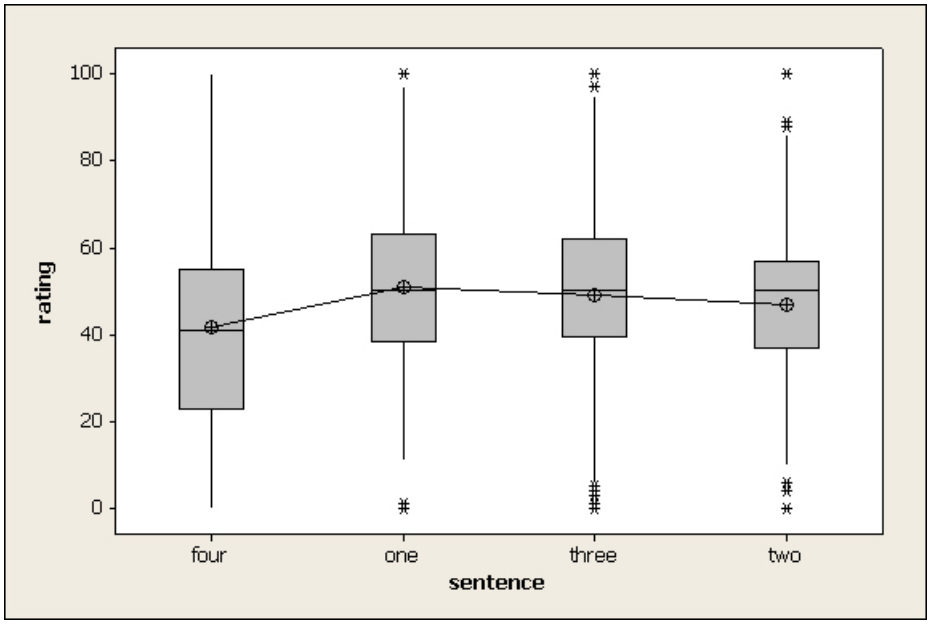


Figure 6.2: The means and variance of the ratings according to the sentence. Outliers are marked with *

To check for consistency between the three blocks, a one-way ANOVA was carried out. A marginally significant difference between the first and the third block was found. The descriptive statistics are shown in Figure 6.3.

## 6.3 Discussion

In general, the hypothesis was supported that more units from a given emotion database resulted in higher the ratings for that given emotion. It is clear from the results that the extreme emotions were recognised correctly. Their mean ratings were higher in the correct directions. But their variance was very high which means that they were not clearly recognised. The neutral condition on the other hand had very low variance.

The conditions that only used about half emotion units were recognised correctly

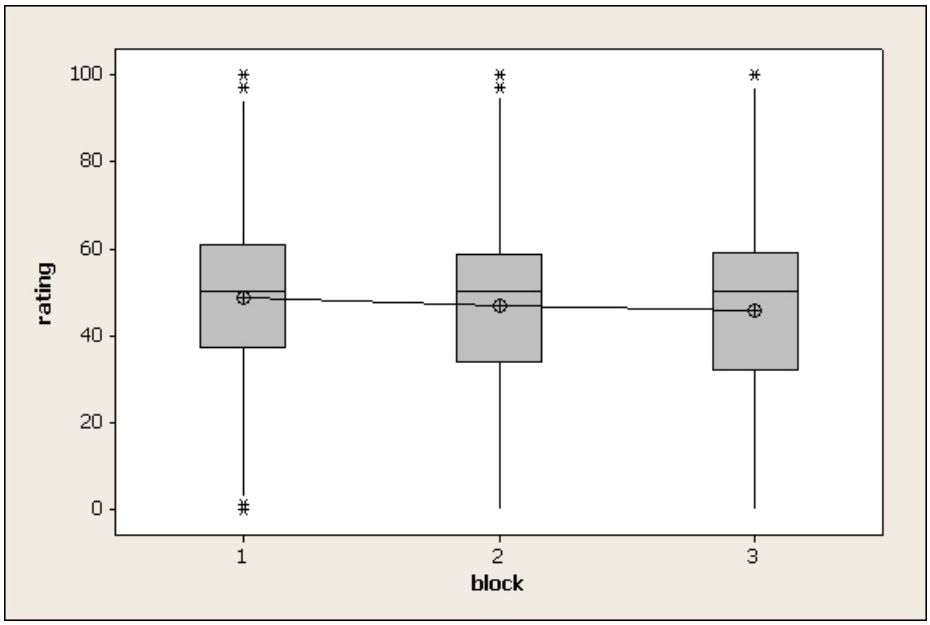


Figure 6.3: The means and variance of the ratings according to blocks. Outliers are marked with *

as well, but did not show such a clear effect. They also had a lower variance than the other extreme conditions which might be explained by their relative proximity to the neutral condition. The two conditions that used only a few emotional units were very close to the neutral condition with no significant difference. Although this was to be expected there is also no clear distinction between the some_happy and some_angry condition. Many outliers were reported for the neutral condition. One explanation for the larger number of outliers might be that subjects were uncertain about that condition in the context of emotion. This could mean that right after hearing a happy example a neutral example sounded more angry and vice versa. In some circumstances, subjects might have been primed from previous examples.

What was surprising was that sentence four was judged significantly more angry than the other sentences, where as the overall ratings of the other sentences were very close to the centre. The only difference apart from the content, which is relatively neutral is the length of the sentence. Sentence four is shorter than the other sentences.

There was not enough data to see if there was a difference in ratings for the na-

tive and the non-native speakers. The literature suggests that there should not be a difference but more participants are needed to confirm that.

Another interesting result was the difference between the three blocks. It looks like subject rated the sentences angrier as time progressed. The difference between blocks 2 and 3 are negligible which suggests that the subjects got more familiar with the task after the first block. This can also explain the difference between blocks 1 and 3.

# Chapter 7

# Discussion and Conclusion

In this project the main goal was to implement a functional unit selection speech synthesiser that produces recognisable emotions. A further target was the implementation of varying degrees of emotions by mixing units from different databases in the final synthesised utterance. To achieve these goals, a number of voices had to be built with varying emotional tone. An actor was recorded reading a script in different emotions. The target cost in the Festival speech synthesis program was changed in order to accompany emotional speech. The Dictionary of Affect was employed to guide the unit selection process by providing knowledge to the emotional target cost function.

The project was aimed at staying as close as possible to the theoretical framework for emotions proposed in psychology. Without a proper theoretical grounding, it would have been difficult to implement the different parts of the emotional speech synthesiser. There was a theoretical motivation for designing the target cost function so that it incorporated the Dictionary of Affect.

The final system was evaluated by a formal perceptual test. Participants had to rate utterances spoken in different emotions along a continuos scale. The hypothesis was confirmed that emotions are perceived as more intense when more emotional units are included in the utterance. It has been shown that by varying the amount of emotional units the intensity of a perceived emotion can be varied. Although the Dictionary of Affect motivated the choice of units it is not clear if its inclusion had an effect on the perceived intensity of an emotion. To test this theoretical foundation of the system, random selection of emotional units should have also been evaluated. There was also

an interaction between the sentence and the emotion, which means that better care should have been taken on ensuring the neutrality of a sentence. The general quality of the emotions synthesised was very good since all tested emotions showed an effect in the right direction with the right intensity. Also, several participants of the rating study mentioned that they were surprised how good the quality of the emotional synthesis was.

Although the goals set at the beginning of this project were achieved, Festival proved to be not the most straightforward platform to work with. The documentation for the voice building process and for the general system was very limited. The text-mode in Festival which was supposed to be used for this project turned out to be unusable. It only returns a waveform and no utterance structure, and the mark-up that is parsed is not automatically accessible from the target cost. It was easier to modify the APML implementation than to work with the text-mode.

Apart from fulfilling the goals set at the beginning, several other interesting possibilities for extensions have opened up during the course of this project. First of all, the database collection process could be modified so it caters better to the elicitation of emotions. All databases collected during this project had very similar coverage. It would be much more efficient to just have a full coverage for the neutral database. For the emotion databases a new script would have to be designed in accordance with the Dictionary of Affect to ensure only the units are recorded that are needed to cover a certain emotion. The distribution of units for the utterances synthesised during this project can give some idea on what needs to be recorded for a given emotion. The emotion mark-up in the database could be enhanced by introducing more fine grained labels or by switching to a dimensional representation. This way units from the database can be better matched with the intended intensity of the utterance.

The synthesiser so far can only portray two different emotions, happy and angry. Over time more emotions could be added but it is not clear if this is necessary. It might be enough to have a synthesiser that can portray positive and negative affect in various degrees. The ability to portray a wide range of emotions might not be as important as the ability to portray them in various intensities. It is the idea of ”how good” and ”how bad” that is the most important to communicate.

A way of enhancing the emotional experience created by the synthesiser would be to fill the pauses of the speech signal with other vocal cues. It was found that disgust is mainly expressed through noises that the speaker makes between the words and not through prosodic variation of the speech signal. The current festival implementation has a pauses module that could be used for adding non-linguistic content to the synthesised signal.

Usually a synthesis system is not a standalone product, but is integrated in a larger system. The quality of the mark-up becomes very important in a wider context. The APML language has been used before to mark-up text for a talking head. But instead of mark-up, the prediction of emotion from text and other mulitmodal cues can be attempted. This would lead to a more natural expression because the content of an utterance would fit to the emotional tone in the speech. In a conversational system the agent's linguistic expression is a clear indicator for it's emotional state. For that expression to be believable it needs to be accompanied by appropriate non-verbal cues, like prosody. These expression will seem unnatural or false if they do not match what is expressed linguistically. The synthesis system should therefore be able to predict emotional information from linguistic input. The system implemented in this MSc project does not predict emotional information explicitly but is able to vary the intensity of an emotion according to the words used in the utterance. This means that the current system varies the emotional intensity not only according to mark-up but also to how strong the emotional connotation of a given sentence is.

The difference between professional actors and regular speakers portraying emotions needs to be explored. The general intuition is that actors are more suited for such a task, but as pointed out earlier, it is not clear if the emotions that they portray sound natural. As with all unit selection synthesis, the quality of the synthesis ultimately depends on the speaker and even more so with emotion synthesis. How natural a synthesised emotion sounds depends on how natural the speaker portrays that emotion. Therefore if natural emotional speech could be used for the synthesis process, the quality of the portrayed emotion would improve.

It has been argued before that unit selection is a dissatisfactory method from a research perspective because it is not true model of speech but rather plays back pre-

viously recorded speech in an intelligent way. Systems that try to model the speech apparatus might be more interesting for research purposes but as long as there is no clear understanding of the acoustic parameters involved in emotional speech, unit selection synthesis remains the best option for synthesising emotional speech.

# Bibliography

[1] Averill, J. R. (1975). A semantic atlas of emotional concepts. JSAS Catalog of Selected Documents in Psychology, 5:330. Ms. No. 421.

[2] Banse, R., & Scherer, K. R. (1996). Acoustic Profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), 614-636

[3] Black, A. W., & Kevin, K. (2000). Building voices in the Festival speech synthesis system. http://festvox.org.

[4] Black, A. W. (2003). Unit Selection and Emotional Speech, Eurospeech 2003, Geneva, Switzerland.

[5] Burkhardt, F. (1999). Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren. PhD thesis, TU Berlin.

[6] Bulut, M., Narayanan, S. S., & Syrdal, A. K. (2002). Expressive Speech Synthesis using a Concatenative Synthesizer. ICSLP 2002.

[7] Cahn, J. E. (1989). Generating expression in synthesized speech. Master's thesis MIT Media Lab.

[8] Clark, R. A. J., Richmond, K., & King, S. (2004) Festival 2 - Build Your Own General Purpose Unit Selection Speech Synthesiser. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA.

[9] Cowie, R., Sawey, M., & Douglas-Cowie, E. (1995). A new speech analysis system: ASSESS. In Proceedings of the 13th Internationl Conference of Phonetic Sciences, volume 3, pages 278-281, Stockholm, Sweden.

[10] Cornelius, R.R. (2000). Theoretical approaches to emotion. In Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland.

[11] Campbell, N. (2000). Databases of emotional speech. In Proceedings of the ISCA Workshop on Speech and Emotion, pages 34-38, Northern Ireland.

[12] Darwin, C. (1872). The expression of emotion in man and animals. London: John Murray.

[13] Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication Special Issue Speech and Emotion, 40(1-2):33-60.

[14] Eide, E., Aaron, A. Ahem (2004). A corpus-based approach to Expressive Speech Synthesis. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA.

[15] Ekman, P. (1993). Facial Expression and Emotion. American Psychologist, 48(4):384-392.

[16] Frijda, N. H. (1986). The Emotions. Cambridge University Press, Cambridge, UK.

[17] Gerrards-Hesse, A., Spies, K., & Hesse, F. W. (1994). Experimental Induction of emotional states and their effectiveness: A review. British Journal of Psychology, 85:55-78.

[18] Hunt, A., & Black, A. W. (1996) Unit selection in a concatenative speech synthesis system using a large speech database. In ICASSP-96, volume1, pages 373-376, Atlanta, Georgia

[19] IIda, A. (2002). Corpus based speech synthesis with emotion. PhD thesis. University of Keio, Japan.

[20] Johnstone, T., Banse, R., & Scherer, K. R. (1995). Acoustic Profiles from Prototypical Vocal Expressions of Emotion. Proceedings of the XIIIth International Congress of Phonetic Sciences, 4, 2-5.

[21] Lazarus, R.S. (1991) Emotion and Adaptation. Oxford University Press, New York.

[22] Plutchik, R. (1994). The Psychology and Biology of Emotion. HarperCollins College Publishers, New York.

[23] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology. 39:1161-1178.

[24] Saikachi, Y. (2003). Building a Unit Selection Voice for Festival. MSc thesis. University of Edinburgh.

[25] Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. European Journal of Social Psychology. 8:467-487.

[26] Scherer, K. R. (1984) Emotion as a multicomponent process: A model and some cross-cultural data. Review of Personality and Social Psychology. 5:37-63.

[27] Schere, K. R (1986) Vocal affect expression: A review and a model for future research. Psychological Bulletin. 99:143-165.

[28] Schrder, M. (2003). Speech and Emotion Research. An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis. UniversitŁt des Saarlandes. Saarbrcken.

[29] Taylor,P. A., Black, A. W., & Caley, R. (1998) The architecture of the festival speech synthesis system. In The Third ESCA Workshop in Speech Synthesis, pages 147-151, Jenolan Caves, Australia.

[30] Whissell, C. M. (1989). The dictionary of affect and language. In Plutchik, R. and Kellerman, H., editors, Emotion: Theory, Research, and Experience. volume 4: The measurement of emotions, pages 113-131. Academic Press, New-York.