Emotional speech synthesis for a Radio DJ: corpus design and expression modeling

Martí Umbert i Morist

Master Thesis MTG - UPF / 2010

Supervisors:

Jordi Bonada
Jordi Janer
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

UNIVERSITAT
POMPEU FABRA

Acknowledgments

During this Master course, it has been huge the amount of people that has contributed to grow my personal interest for music and the technologies that deal with it. Without any of you it would not have been possible to learn what we have been taught during this year.

Firstly, I sincerely thank Dr. Xavier Serra for giving me the opportunity to join the Music Technology Group, letting me participate in the group activities such as the Sound and Music Computing conference and for broadening my future perspectives.

I would like to thank Dr. Jordi Janer and Dr. Jordi Bonada for supervising my master thesis, as well as Oscar Mayor and Merlijn Blaaw for providing me the necessary tools and technical support.

Furthermore, I thank Dr. Emilia Gómez, Dr. Hendrik Purwins and the rest of MTG researchers for their help during this year. My thanks are also to the classmates who have shared this journey to the Master's degree. Thanks to Pratyush, Marco, Zuriñe, Andreas and all those who have shared these imaginary ”parties” in room 55.312.

Finally, I would like to thank my friends and family for their enless support.

Abstract

This master thesis concerns the design of a corpus for speech synthesis as well as the modeling of different emotions in the context of a Radio DJ speaker.

In the context of the radio DJ speaker we designed a corpus that represents what radio DJs use to present songs being played in a radio show. A professional speaker has been recorded uttering a set of these sentences in different levels of arousal and speed. By labeling the phonemes of the recorded phonemes, control parameters have been extracted from these sentences in order to transform or synthesize them in other emotion and speech rate conditions, and thus change the control parameters accordingly or the synthesized keywords such as a band or a song name.

More precisely, the aim of this project is to model how different acoustic parameters behave according to a given emotion. The model considers syllable energy, duration and pitch which will be used to transform (or even synthesize) a recorded sentence into another with a different emotion. These results are objectively compared to the training data as well as subjectively evaluated in terms of emotion activation and speech rate.

Contents

Abstract **vi**

List of Figures **xii**

List of Tables **xiii**

1 Introduction **1**

1.1 Overview 1

1.2 Radio DJ Project 2

1.3 Goals 2

2 State of the Art **5**

2.1 Speech and singing voice synthesis 5

2.2 Emotions in speech 6

2.2.1 The Circumplex Model of Affect 7

2.3 Synthesis control 8

2.3.1 Acoustic Correlates for speech synthesis 9

2.4 Related Work 10

3 Design, Recording and Labeling of the Radio DJ Corpus **11**

3.1 Design 11

3.1.1 Overview 11

3.1.2 Getting radio shows 11

3.1.3 Analyzing radio shows 12

3.1.4 Different criteria to generate corpora 14

3.1.5 The generated corpus 15

3.2 Recording 15

3.2.1 Final recorded set 15

3.2.2 Recording Setup 16

3.3 Labeling 16

3.3.1 Emotion 16

3.3.2 Phonetic transcription 18

3.3.3 Phoneme segmentation 18

3.3.4 Syllable duration 18

3.3.5 Syllable pitch and energy 19

4 Methodology for modeling emotion 21
4.1 Simplifying the circumplex model of affect 21
4.2 Emotion modeling workflow 21
4.2.1 Training the model 22
4.2.2 Generating predictions 23
4.2.3 Postprocessing the predictions 24
4.2.4 Pitch shifting and energy modification 25
4.2.5 Time domain transformations 25
4.2.6 Objective tests 25

5 Evaluation and Results 31
5.1 Evaluation Methodology 31
5.1.1 Perceptual test 31
5.2 Results 33
5.2.1 Perceptual test 33
5.2.2 Correlation between perceptual and objective results 34

6 Future work and conclusions 41
6.1 Future Work 41
6.1.1 Research work 41
6.1.2 Possible applications 42
6.2 Conclusions 42

References 44

A Example: Song Label structure 49

B Example: Clustered data 53

C Recorded sentences 55
C.1 Sentences placed before the song 55
C.2 Sentences placed after the song 55
C.3 Sentences placed before and after the song 56
C.4 Hello and goodbye sentences for a show 57

D Labels of the sentences 59
D.1 Labels for sentences placed before the song 59
D.2 Labels for sentences placed after the song 59
D.3 Labels for sentences placed before and after the song 60
D.4 Labels for hello and goodbye sentences for a show 60

E Objective test results 61
E.1 Using an anchor sentence close to the target 61
E.2 Using the same neutral-calm anchor sentence for all transformations 62

List of Figures

1.1 Thesis and internship main blocks 3
1.2 Thesis workflow 3

2.1 Voice organ blocks 6
2.2 Unit dimensions 7
2.3 Circumplex model of affect 8

3.1 Radio DJ corpus generation workflow 12
3.2 Speaker and microphones setup for the recording sessions 17
3.3 Phoneme labeling process sample 18
3.4 SPPTools: used to extract pitch and energy 19

4.1 Circumplex model of affect activation dimension 22
4.2 Simplified circumplex model of affect 23
4.3 Workflow to generate predicted sentences 23
4.4 Predictions placed in both the train and unseen regions of the control space 24
4.5 Syllable duration prediction for sentence 2: fast-excited emotion with 0.18 sec/syll as speech rate, close anchor 27
4.6 Syllable energy prediction for sentence 2: fast-excited emotion with 0.18 sec/syll as speech rate, close anchor 28
4.7 Syllable pitch prediction for sentence 2: calm-slow emotion with 0.30 sec/syll as speech rate, close anchor 29
4.8 Syllable duration prediction for sentence 1: calm emotion with 0.25 sec/syll as speech rate, calm-fixed anchor 29
4.9 Syllable energy prediction for sentence 2: calm-slow emotion with 0.30 sec/syll as speech rate, calm-fixed anchor 30
4.10 Syllable pitch prediction for sentence 1: excited emotion with 0.19 sec/syll as speech rate, calm-fixed anchor 30

5.1 Perceptual evaluation: Answer sheet 32
5.2 Perceptual evaluation: sentence 1 ratings of real recordings for calm and energetic samples 33
5.3 Perceptual evaluation: sentence 1 ratings of real recordings for fast and slow samples 33
5.4 Perceptual evaluation: sentence 2 ratings of real recordings for calm and energetic samples 34
5.5 Perceptual evaluation: sentence 2 ratings of real recordings for fast and slow samples 34

5.6 Perceptual evaluation: Ratings for sample ID 1 35
5.7 Perceptual evaluation: Ratings for sample ID 2 35
5.8 Perceptual evaluation: Ratings for sample ID 3 36
5.9 Perceptual evaluation: Ratings for sample ID 4 36
5.10 Perceptual evaluation: Ratings for sample ID 5 36
5.11 Perceptual evaluation: Ratings for sample ID 6 37
5.12 Perceptual evaluation: Ratings for sample ID 7 37
5.13 Perceptual evaluation: sentence 1 error distance 38
5.14 Perceptual evaluation: sentence 2 error distance 38
5.15 Speech rate correlation between rated and real values 39
5.16 Activation vs energy correlation 39

List of Tables

3.1 Duration of complete shows and labeled data 14
3.2 Final recorded corpus information 16
3.3 Recording setup description 16
3.4 Vectors used to label the emotion 17
3.5 Sentence phonetic transcription example 18

4.1 Control parameters of the evaluated sentences 26
4.2 Summary of the mean values for all features in sentence 1 27
4.3 Summary of the mean values for all features in sentence 2 28

5.1 Perceptual evaluation: error modules from the mean rating to ground-truth 35

D.1 Labels and occurrences for sentences placed before the song 59
D.2 Labels and occurrences for sentences placed after the song 59
D.3 Labels and occurrences for sentences placed before and after the song 60
D.4 Labels and occurrences for hello and goodbye sentences 60

E.1 Objective evaluation with close anchor sentence: syllable duration feature for sentence 1 61
E.2 Objective evaluation with close anchor sentence: syllable duration feature for sentence 2 61
E.3 Objective evaluation with close anchor sentence: Energy feature for sentence 1 . . 62
E.4 Objective evaluation with close anchor sentence: Energy feature for sentence 2 . . 62
E.5 Objective evaluation with close anchor sentence: Pitch feature for sentence 1 . . . 62
E.6 Objective evaluation with close anchor sentence: Pitch feature for sentence 2 . . . 63
E.7 Objective evaluation with neutral anchor sentence: syllable duration feature for sentence 1 63
E.8 Objective evaluation with neutral anchor sentence: syllable duration feature for sentence 2 63
E.9 Objective evaluation with neutral anchor sentence: Energy feature for sentence 1 . 63
E.10 Objective evaluation with neutral anchor sentence: Energy feature for sentence 2 . 64
E.11 Objective evaluation with neutral anchor sentence: Pitch feature for sentence 1 . . 64
E.12 Objective evaluation with neutral anchor sentence: Pitch feature for sentence 2 . . 64

Chapter 1

Introduction

1.1 Overview

Controlling emotions is a current trend in speech synthesis research. It is important in terms of having a natural sounding synthesized voice and it can help in many applications concerning human-computer interaction.

In this master thesis, the application that contextualizes the current project is a virtual Radio DJ speaker. In this case there is no direct interaction with the listener but, nevertheless, when broadcasting a message it is important to make the receiver of the message feel the intended emotion that wraps the message. Since the communication is uni-directional, the transmitted emotion has more significance.

In this case, the messages contain sentences that present a song in a virtual radio show. There are several keywords that radio DJs commonly repeat in order to introduce a song, such as the song name, the band, some member name, the album the song belongs to, the city and date of the next concert, etc. We identified the most frequent keywords have been covered in a set of sentences that were recorded; these are the corpora the virtual radio DJ will be able to use.

Concerning emotions, in this project we deal with the concept of arousal (calm, neutral or energetic) and added also speech rate in order to have a higher degree of manipulating the synthesized emotion. The valence dimension of emotions was not relevant to us since the target speaker is a radio DJ, which in general reflect a rather neutral/positive mood and not a negative one.

A main objective in this project is to study how emotions are mapped into the parameters that are used to control the text-to-speech synthesis engine (TTS). From this study models can be build in order to be able to make the virtual radio DJ express in intermediate ways not given in the database recordings used for training.

This master thesis is organized as follows. In chapter 2 the *State of the Art* presents the main concepts, from speech synthesis techniques, its control parameters, emotions and some related previous works. The design process of the Radio DJ corpora is described in chapter 3. Here it is also included how it has been recorded and labeled afterwords in terms of the control parameters.

Then, chapter 4 addresses which strategies are adopted to model emotions as well as an objective evaluation. Then, the evaluation methodology and the perceptual test results are presented in chapter 5. Chapter 6 presents aspects that can be taken into account as a future work as well as the conclusions of this master thesis.

1.2 Radio DJ Project

As mentioned in section 1.1, this master thesis related to the development of the Radio DJ project. The aim is to generate a sequence of songs presented by a virtual Radio DJ. The sentences used to introduce the songs should be based on the available information for each particular song, which are the corresponding keywords (song name, band, etc). To get this set of keywords it has been necessary to identify which are the frequent sentence structures and relevant words used to introduce the songs played in radio shows, which will afterwards be recorded.

Then, a built-in voice synthesizer could be used as the engine to synthesize them according to a playlist of a virtual radio show. Obviously, those keywords representing the song name, the band, the concert date, etc, will have changed with respect the original sentences selected for the corpora in order to be flexible to present any list of songs.

1.3 Goals

With respect to the global Radio DJ project context, the goals of this master thesis are:

- Creation of a corpus-specific voice synthesizer.
- Recording and annotation of a corpus.

And more precisely, concerning the master thesis, the main goals can be summarized as:

- Generate an expressive virtual radio DJ in terms of emotions. This project, will focus on the transformation of a reference or anchor sentence instead of using a speech synthesizer.
- Study how to model emotions in terms of syllable duration, mean energy and pitch.
- How to build a continuous emotion control space.

The proof of concept presented in this introductory chapter can be summarized in the following figures. Figure 1.1 shows the main blocks explained in these first two sections and Figure 1.2 focuses on the relationship between the inputs of the system, the control parameters and final output. In this project, the developed idea of transforming an anchor sentence could be extended to controlling a speech synthesizer.

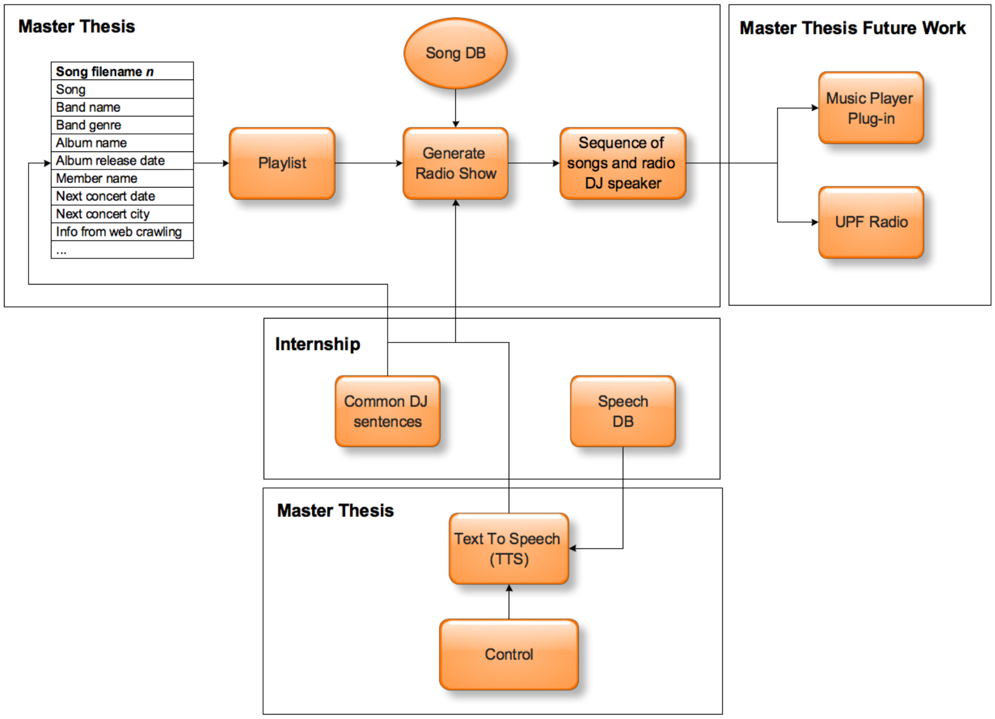


Figure 1.1: Thesis and internship main blocks

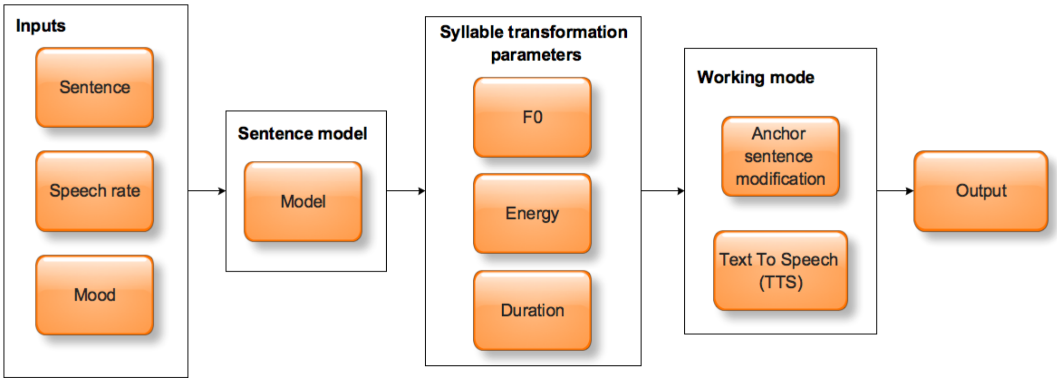


Figure 1.2: Thesis workflow

Chapter 2

State of the Art

2.1 Speech and singing voice synthesis

Different synthesis strategies are typically used in speech and singing voice synthesis depending on the approach according to [Rod02] and [BS07]. These approaches are perception or production of the speech or singing voice. Thus, there are the ones that use a waveform synthesizer and those that use modification and concatenation of recorded units of natural singing voice (concatenative synthesis). The former, is again split into synthesizers as physical models of the voice production mechanism and synthesizers that model the voice signal.

Physical model synthesis models take advantage of the fact that its control parameters are related to the ones used by a speaker or singer. Although the mapping may be intuitive, the drawback is that there are lots of parameters. Figure 2.1 is a representation taken from [BS07] that shows the voice organ and the different parts that take place in the voice production. Each of these parts is modeled and concatenated to get the output. Next section mentions which would the parameters that model this system.

In the second case, the best units from a database are chosen according to the phonemes to utterate and the parameters from the rules. The selected units are then changed to get the expected speech or singing properties. In the context of singing voice, this would be the case of the Vocaloid singing synthesizer [KO07].

In general, the best systems come from a combination of different approaches. In this case, it is important the labeling of the selected units (phonemes, diphones, etc) according to those parameters that later on will need to be modified. In the case of [BS07], these are labeled according to four dimensions: the type of diphone, its pitch, tempo and loudness. These are shown in Figure 2.2.

These systems also tend to introduce post-processing steps in order to improve performances. As in [BS07], which uses a sample-based synthesizer transforming and concatenating samples, transforming techniques are important. Then, there is the need for audio-processing techniques adapted to the particular characteristics of the singing voice.

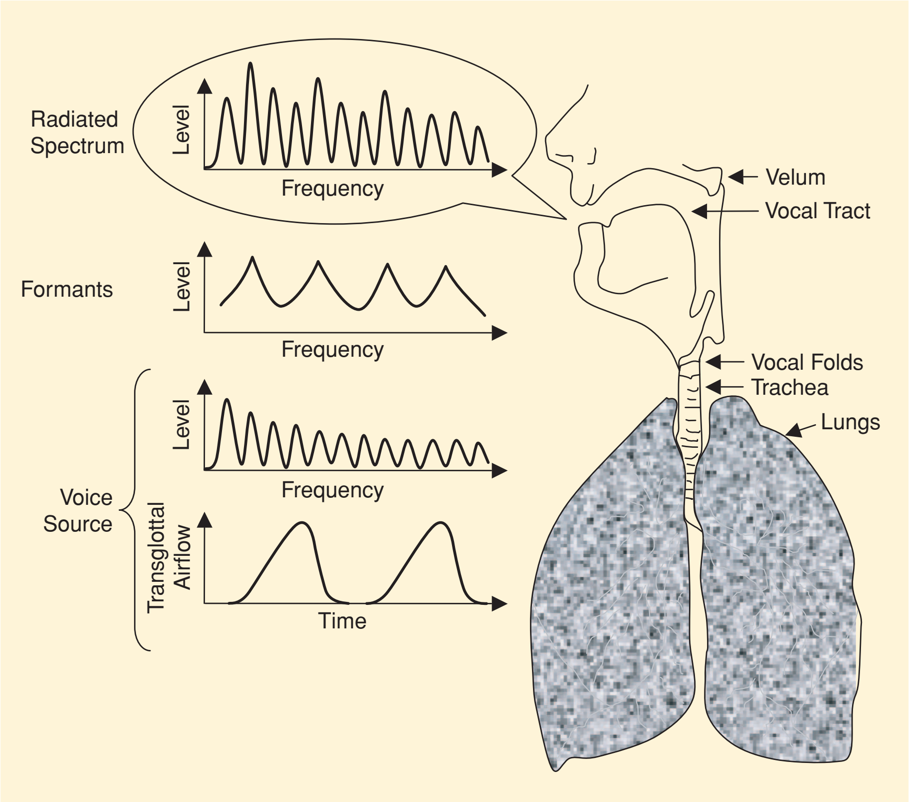


Figure 2.1: Voice organ blocks

2.2 Emotions in speech

The ability of showing emotions in speech and in a musical performance gives them an extra value. It is through a right mapping of emotions into control parameters that this may be achieved. It is necessary, then, to study which are the emotions that can conveyed and also how different expressivity resources are organized and related to low-level features.

In speech, an emotional space was set in [Rus80] and [PRP05], where emotions are placed in a two-dimensional space (circumplex model of affect) relating them to two main processes: arousal (or activation or alertness) and valence (how positive or negative the emotion is). In speech and music, it has been studied for instance in [Sch95]. This model of affect is also introduced in section 2.2.1.

In [MPR10], in this case working with instruments, different emotional intentions are analyzed both acoustical and perceptual commonalities of instrumental songs are studied. Machine learning techniques are applied to observe how expressive intentions are organized. PCA technique is applied to get a visual 2D representation.

Still, with respect to emotion, growl and rough voice have been studied in [LB04]. This vocal disorders are also used by performers as an expressive resource. In this work, the studied roughness is the one due to inter-period variations of the pitch (jitter) and the period amplitude (shimmer). Concerning growl, singers use it as an expressive accent. Their algorithm generates the new sub-harmonics that appear due to this disorder.

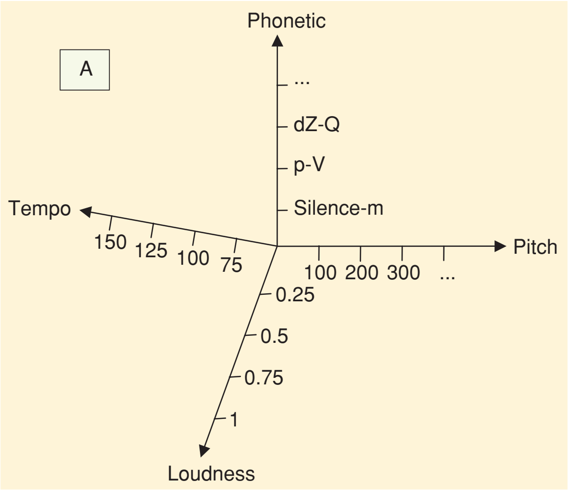


Figure 2.2: Unit dimensions

Concerning music performances, a state of the art review is done in [WG04] concerning computational modeling of expressive performance with respect to parameters like tempo, timing, dynamics and articulation. These models catch commonalities and differences between performances and performers. Focusing on singing voice performance, [BLK03] presents a systems that mimics a real recording after analyzing its parameters.

Finally, an approach for modeling and controlling the emotion can be found in [CDPD$^+$04]. The authors apply morphing techniques to change expressive intentions continuously working both at high (symbolic) and low (features) levels.

2.2.1 The Circumplex Model of Affect

In [Rus80], the authors studied which was the emotional scope of the human voice. They came out with different representations of to the *circumplex model of affect*. Its representation in the direct circular scaling coordinates form for 28 affect words is shown in Figure 2.3. It is characterized by a 2-dimensional representation in terms of arousal (or activation or alertness) and valence (how positive or negative the emotion is).

In voice and music, it has been studied for instance in [Sch95]. This work studies how listeners identify speaker's emotion as well as discusses the similarities of emotion expression in speech and music.

The cicumplex model has also been studied in [PRP05]. In this case it was extended to a 3-dimensional space. Also, in [IT06] the extended version of the emotion space to three dimensions of affect (energy arousal, tension arousal and valence) was used to compare acoustic parameters in both music and speech. The authors studied the degree of overlap between affective qualities in music and speech by directly comparing intensity, pitch and tempo. They conclude that there is a general mechanism that links acoustic features to emotions. However, some differences were found in the behavior of features with respect to dimensions and emotions, which could be taken into account. Nonetheless, these differences show that different attentional strategies may be used for speech and music.

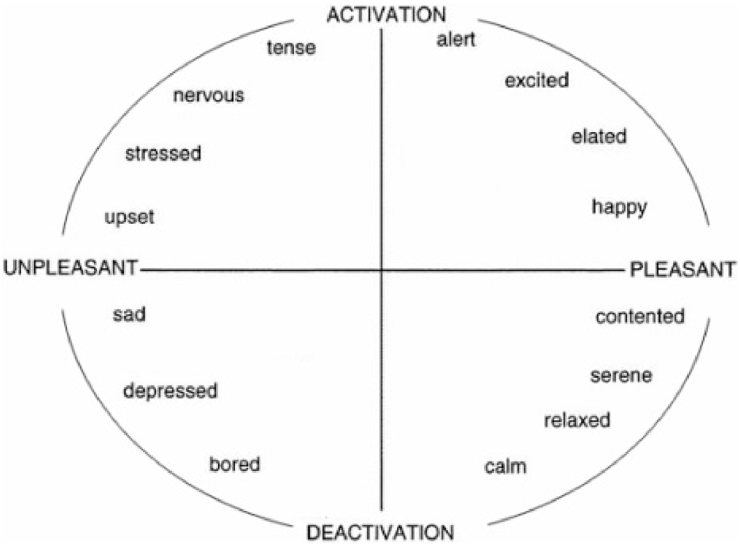


Figure 2.3: Circumplex model of affect

2.3 Synthesis control

One of the main targets of an effective control of the speech or singing voice synthesis is to produce a natural and expressive voice. Current techniques are still far from achieving such results since it is needed to model the speaker or the singer.

In the case of singing voice, for instance, it involves multidisciplinary concepts from music theory, cognition and motor control problems as stated in [BS07]. Two main inputs have to be considered [Rod02]: score and lyrics. The score controls a variety of parameters of the synthesizer. On the other hand, the lyrics need to be converted into a phonetic sequence in order get the phonemes to sing.

From the production point of view of the physical models, parameters were described for example in [Coo91]. As summarized in [BS07], the parameters to control in this case are jaw opening, tongue sharp, sub-glottal air pressure, tensions of the vocal folds, etc. The high amount of parameters make this kind of systems difficult to map to a desired output. Models for each of the vocal tract parts are needed as reviewed in [KB09].

For instance, in this context, recent work on the control of the expressiveness of the synthesized singing voice has been done in [dA09]. In this work, resulting synthesis engine, the flow parameters are changed in order to access to the perceptual dimensions of a voice quality control space.

From the perception perspective in the spectral models, this mapping is made more easily. On the other hand, the parameter space is not as intuitive as in the physical models, as could be glottal pulse, spectral shape, formant frequencies and bandwidths, etc. Some others can be taken from the score, as the mean pitch, energy dynamics and musical articulation. The physical articulation (vibrato, pitch tracking, loudness, timbre) of the vocal tract is calculated with rules if the guiding performance is not given.

Finally, in [GL08], artificial neural networks (ANN) are used to model expressivity. Particularly, these are used to model one of the important emotional factors, the vibrato.

2.3.1 Acoustic Correlates for speech synthesis

In the case of speech synthesis, the acoustic correlates to model emotions that are used focus on intonation, tempo and intensity as stated in [SCDc$^+$01]. These can be summarized as follows:

For intonation:

- the values for F0 mean and range,
- accent structure: number of F0 maxima / minima per time unit, duration, magnitude and steepness of F0 rises and falls.

For tempo:

- duration of pauses,
- articulation tempo.

For intensity:

- the values for intensity mean and range,
- a measure of dynamics (the difference between mean intensity for intensity maxima and overall mean intensity).

This last reference also provides an interesting tool, FEELTRACE, to perceptually annotate emotional content. It allows to place the tone of an audio file in a 2-D representation in terms of activation-evaluation space, continuously over time.

The design features and their relationship with different affective states have also been described in [SB04]. These design features are intensity, duration and synchronization as response characteristics and event focus as antecedents, and appraisal elicitation and rapidity of change and behavior impact as consequences in terms of stability and impact on behavior choices.

The effect of prosodic features such as pitch level and range, articulation rate and loudness have been studied in [TSS$^+$06]. In this case the aim is to model personality features more than emotion. Still concerning prosody, robust F0 modifications with respect to the emotions perception have also been studied in [BN08]. The authors perform modifications on F0 mean, range and stylization (pitch contour representation with linear segments).

Finally, other strategies reach emotional speech by converting from neutral speech, for example as in [TKL06]. In this work, different models are tested, such as linear model modification (LMM), Gaussian mixture model (GMM) and a classification and regression tree (CART). The LMM modifies the F0 contours, syllabic durations and intensities. On the other hand, the GMM and CART models map prosody distributions between neutral and emotional speech. The CART model also includes linguistic information.

2.4 Related Work

Some previous works share common aspects with the current master thesis that could be applied to the idea of predicting control features of the speech or singing voice synthesizer with just a few input parameters.

In [CLB$^+$00], the voice features (like timbre, pitch, articulations and vibrato) of a user are morphed with the ones from a prerecorded singer. It is applied in real-time in a karaoke context in order to impersonate, at different possible degrees, the voice of the user and make it more or less similar to the target singer.

Other approaches that use another performance to control the synthesis as in the performance driven strategies, for example in [JBB06]. In this case, a singing performance directly controls the synthesized expression.

In [NG], their system tries to mimic a reference user voice by automatically predicting its parameters (in this case F0, energy and onset time and duration of notes) for singing synthesis with the help of song lyrics. The problems this engine tries to overcome are that the parameters' configuration is not easy to achieve, therefore being time consuming, and to be robust to changing conditions like the database.

Concerning the modeling of the control parameters according to inputs of the system, previous work can be found for example in [SUA05] and [SGUA07]. Both deal with modeling F0 in order to generate pitch contours mainly using second order exponential damping and oscillation models.

Linking the F0 contour and the music style there is the work of [KOK$^+$09], matching Gaussian mixture models to the extracted F0 fluctuations. The inverse path of classification is used in order to generate pitch contours representative of the desired singing style and melody score.

Having in mind the concept of the control of synthesis and how expression is mapped, in other fields different to speech or singing voice, there is the work of [Mae09]. In this case, it is worth mentioning that the control parameters are the bowing contours and are used to get a natural violin sound, extracted from an annotated input score. Two sound synthesis approaches (physical modeling synthesis and sample-based synthesis) were taken into consideration.

Chapter 3

Design, Recording and Labeling of the Radio DJ Corpus

3.1 Design

This section summarizes the work done concerning the task of preparing the Radio DJ prototype sentences of the corpus database.

3.1.1 Overview

A first task in the project is to create a corpus of typical sentences Radio DJs use in order to introduce the songs in their shows. In order to narrow the language variability of the speaker and the way songs are introduced, we have focused on pop/rock radio stations.

Several real radio shows from different sources have been downloaded. The segments where the DJ gives relevant information have been extracted, transcribed and labeled. For the labeling, a syntax of tags has been created in order to identify keywords with generic more content (the song, the group, the position of the sentence with respect to the song, etc).

Labels have different frequency according to the number of occurrences in the whole group of transcribed shows. Also, sentences can be clustered concerning the information (labels) they provide. These clusters can be ordered in terms of relevance: clusters are weighted as the mean of their label frequencies. By recursively taking labels from more to less important clusters, more sentences/labels/clusters will be covered. The corpus can be generated according to different stop criteria: number of sentences/labels/clusters in the corpus and percentage of weight clusters covered.

All these steps are shown in Figure 3.1 and described in the following sections.

3.1.2 Getting radio shows

Getting radio shows available on the net has not been an easy task, both in the sense of finding them but also with respect to their content.

The most useful tools have been:

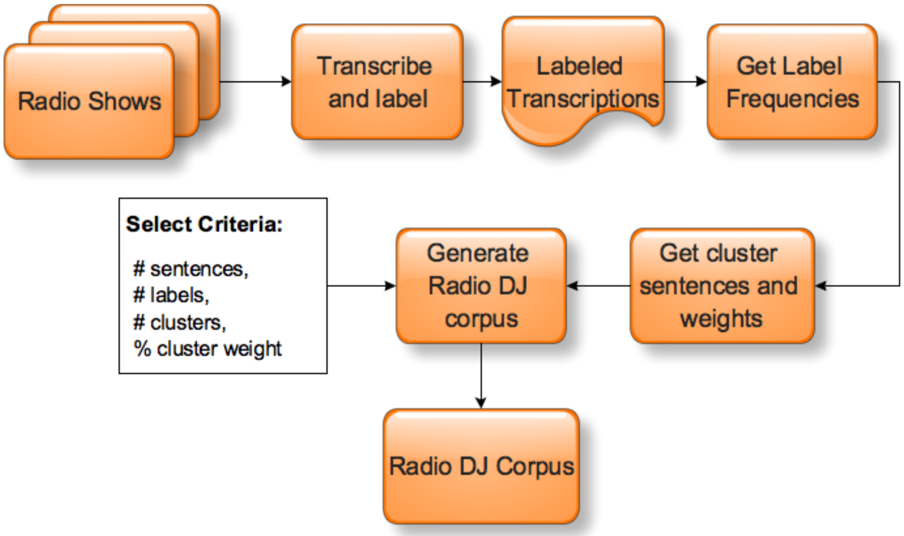


Figure 3.1: Radio DJ corpus generation workflow

- iTunes (there are some free Pop/Rock radio shows available),
- some web pages that collect radio shows with links to their podcasts [1]

Some issues have been observed while selecting the radio shows that are worth taking into account and we'll have to check how they affect the future work:

- Some radio shows are based on a dialog between a couple of speakers. In this case, it becomes more difficult to get sentences, since they usually overlap when talking, or even the same sentence is said the first half by one speaker and the second half by the second one.
- Also, the speech signal may not be completely clear if there's music on the background while the DJ is talking.
- It's recommendable that the shows are of the same kind of music. The vocabulary/type of sentences used in Pop/Rock content shows is different from the one used in House music (more content based on remixes, featuring), or cover programs (covers vs original version), etc. In our case, it's better to focuse on Pop/Rock music shows.

3.1.3 Analyzing radio shows

The relevant information given about each song, or the show, the speaker, etc has been labeled in order to get an idea of which kind of information radio DJs usually provide.

Labels

The labeling process has been a trial an error process, specially in the begining. The outcome has been a specific syntax for the labels that wrap the tagged keyword.

[1] http://www.podcastalley.com/podcast_genres.php?pod_genre_id=3

It has been considered that the most convenient way of tagging information is working with 2 levels of information: the object it tags and the concept tagged. For instance, if the word being tagged is the name of a song (*Rebel Hero*), the label used follows this format:

- $song.name(Rebel Hero)

Once all sentences have been labeled, the number of times each label appears and its frequency with respect to the global number of labels can be computed. This information is shown in Appendix A.

The information shown in this appendix are the different labels grouped according to the first level; that is to say, all labels regarding song information are presented together, the same for the group, and so on. At each level, the number of occurences (*n_elem*) and overall frequency (*freq*) are shown, as well as the examples found for each case with the number of occurences. As an example, some of the most common elements are:

- $group.name, 118 times
- $song.name, 103 times
- $album.name, 33 times

The following lines show how labeled transcriptions look like:

```
$sentence.before(pos) And now to $group.name(Eluvium).
$sentence.before(pos) Here it is, they're the $group.name(Repeat Offenders).
$sentence.before(pos)  It's $song.name(Lovers).
```

Time duration of labeled data

Most of the time, there is music content rather than speech content. Table 3.1 shows these differences in time concerning the whole shows duration and the labeled data duration as well as some of the introduced points such as accent and speech quality commented in section 3.1.2.

Clustering labeled sentences

The different labels the appear in each sentence have been used in order to cluster sentences: those sentences with the same set of labels belong to the same cluster. Each cluster has been weighted according to the mean value of the labels' frequencies that belong to it. Ordering clusters by this weight gives us an idea of what the final corpus should cover.

Appendix B shows the cluster structure used for processing.

Some observations

A part from the infromation provided about the song, group, etc, sentences vary depending on if these are said before the song is played or after. To take this information into account it has been included a label concerning the position of the sentence:

- $sentence.before(pos): before the song,

Radio show	Duration	Labeled data	Accent	Speech quality
662_Coverville	1h:24m:00s	0m:45.39s	American	Good
01Chopin...NewsomMore	0h:35m:01s	2m:40.49	American	Good
01FelaKuti...GibsonMore	0h:15m:36s	0m:57.25s	American	Good
01MagneticFields...JazzBand	0h:27m:21s	0m:53.11s	American	Good
RadioOrphansPodcast235	0h:31m:51s	1m:34.52s	American	2 speakers
RadioOrphansPodcast241	0h:27m:24s	2m:17.19s	American	2 speakers
RadioOrphansPodcast245	0h:33m:15s	0m:05.54s	American	2 speakers
RadioOrphansPodcast247	0h:30m:18s	0m:23.67s	American	2 speakers
RadioOrphansPodcast248	1h:04m:40s	0m:14.56s	American	2 speakers
ReleaseYourselfPodcast102	1h:01m:27s	2m:28.36s	American	Music overlapped
TRIntroFreeIntro...08Mar10	0h:45m:02s	3m:00.86s	British	Some music overlap
TRIntroFreeIntro...22Feb10	0h:44m:50s	4m:34.82s	British	Some music overlap
XYRocks_011	0h:29m:46s	0m:21.96s	American	Good
XYRocks_082	0h:58m:20s	0m:29.79s	American	Good
XYRocks_083	0h:52m:44s	0m:23.72s	American	Good
XYRocks_084	0h:54m:34s	0m:29.17s	American	Good
ALL	**18h:22m:20s**	**21m:40.46s**		

Table 3.1: Duration of complete shows and labeled data

- $sentence.after(pos): after the song,
- $sentence.sequence(pos): the current sentence is related to the previous labeled sentence. This label is used to help find those pieces of information about a song that is given both before and after the song.

These labels have been used in order to differenciate between clusters, explained in section 3.1.3, but are not taken into account to weight clusters, since it has been considered that 2 sentences carrying the same information can be equally considered no matter what their position is with respect to the song in order to have a balanced corpus in terms of where the sentence is placed. Of course, it could have been done the other way round (taking these weights into account) since in general sentences are places after the song in the shows that have been analyzed (123 occurrences vs 82).

At the same time of building the cluster data structure, sentences are grouped into sequences, information provided by the label *$sentence.sequence(pos)*. This is to generate a corpus that reflects how sentences where linked in real radio shows. The final corpus is generated recursively by taking sentences by groups of sequences that are covered from high to low cluster weight until reaching the stop criteria defined in next section.

3.1.4 Different criteria to generate corpora

The generation of the nearly final corpus is by processing clusters from high to low weight, and ordering its labels from high to low frequency. Then, label by label, cluster by cluster, different information is gathered:

- The approximate number of words in the output corpus,

- The cumulated number of sentences covered up to this point,
- The cumulated number of labels covered,
- The cumulated covered clusters (in number, percentage, percentage in terms of weight and amount of words).

One thing that has to be taken into account, is that by following this procedure, once a label is added to the corpora, clusters other than the current one can be added to the final corpus, since clusters share labels, and after some loops a certain low weight cluster may be covered by the higher weighted clusters already covered.

These are the possible criteria in order to decide when to stop generating the output corpus. To generate the corpora for this project, it was finally done by setting the number of words as the main criteria since it is easier to relate it to the amount of time needed to record it. It is typically considered that in 1 hour can be pronounced around 9000 words. Since our corpus should be repeated 4 times, it has to be considered that reading it once can take around 15 minutes. Taking also account the pauses and that it usually takes time to explain the instructions and repetitions it has been set to be a corpus for 10 minutes, which means around 1500 words.

3.1.5 The generated corpus

As explained in section 3.1.4, the corpus has been generated with the criteria of 1500 words. The resulting corpus has been split according to where is said with respect to the presented song: the information is just said before the song, after the song, or a little bit of both. Appendixes C and D show the final corpus labeled sentences that could be recorded and the summary of the covered labels.

3.2 Recording

This sections describes what was finally carried out during the recording sessions. Since it was done with a professional speaker and therefore limited in time, just a part of the corpus was recorded.

3.2.1 Final recorded set

The recorded set of sentences can be found in appendix C. The corpus has been grouped according to where sentences were placed in the real show, that is to say, whether the information was given only before or after a song or a little bit of both. Also, welcome and goodbye sentences were recorded as well as a set of sentences with additional such as the hour, day, month, etc.

Description

Part of the final recorded set was uttered in all emotions (calm, slow, fast and excited), and part only in slow and fast or neutral emotions. Those sentences recorded in all moods are the ones used to model the emotion space. Those sentence that were recorded in just some emotions limit the control space and might just be used as control parameters themselves for a synthesizer.

Table 3.5 shows the amount of sentences recorded according their position with respect to the song, the emotions in which these were recorded and the amount of difference labels for that set. Additionaly to this table, around 70 more short sentences or words where recorded taking account the hour of the day, the day of the week, the month and ordinal numbers.

Sentence position	# of sentences	moods	# different labels
Before	21	slow, fast	20
After	16	slow, fast	15
Before & after	33	slow, fast, calm, energetic	20
Hello/Goodbye	12	neutral	12
Total	82	-	30

Table 3.2: Final recorded corpus information

3.2.2 Recording Setup

The recording session was done with Merlijn Blaaw and Jordi Janer during the Vocaloid recording sessions for its new database. Table 3.3 summarizes the characteristics of the recording setup and Figure 3.2 was taken during the recordings.

Feature	Description
Harware	2x Neumann U87 microphones: 1x cardioid (on-axis), 1x omni-directional (off-axis)
	2x mic stand, shock mount
	Anti-pop filter for cardioid mic
	Headphones for talk-back
	Yamaha 03d mixing console
Software	Steinberg Nuendo
Script	Printed and placed on top of the microphones

Table 3.3: Recording setup description

3.3 Labeling

This section describes the process followed in order to label the data that has been used to model emotion.

3.3.1 Emotion

In section 2.2 the concept of model of affect and has been introduced. This model has been simplified in order to get the control space in terms of activation and speech rate, as explained later

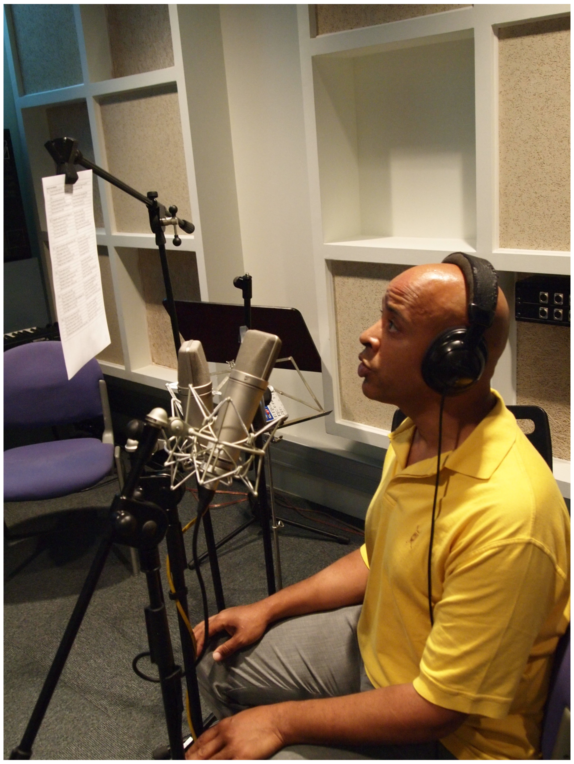


Figure 3.2: Speaker and microphones setup for the recording sessions

in this text in section 4.1. This section describes how the different moods or emotions have been labeled. Table 3.4 shows how emotions have been identified as input vectors with different activated coordinates. The idea here is to that for the training examples, all emotions are set to 0 except the one corresponding to each utterance. To generate new elements in the control space, any other vector could be used so that it represents a combination of the training examples. This representation has been chosen because of emotions has been chosen in order to be able to express unseen training data by activating more than one coordinate. It could also have been considered to use just 3 activation coordinates and thus relate it directly to the activation dimension by setting, for example, -1 for calm, 0 to neutral fast and slow and 1 to energetic (or excited).

Mood	**Coord. 1**	**Coord. 2**	**Coord. 3**	**Coord. 4**
calm	0.0	0.0	0.0	*1.0*
slow	0.0	0.0	*1.0*	0.0
fast	0.0	*1.0*	0.0	0.0
energetic	*1.0*	0.0	0.0	0.0

Table 3.4: Vectors used to label the emotion

3.3.2 Phonetic transcription

The phonetic transcription has been obtained by means of the in-house transcription tool provided by Merlijn Blaauw which transcribes text into a sequence of phonemes using the symbols of the SAMPA dictionary, with accents and word and syllable boundaries.

Example:

Sentence	And that was We the Kings with This is out of town
Phonemes	[Sil] [@ n d] [D @ t] [w @ z] [w i] [D @] [" k I N z] [" w I T] [" D I s] [I z] [" aU 4] [@ v] [" t aU n] [Sil]

Table 3.5: Sentence phonetic transcription example

3.3.3 Phoneme segmentation

This section describes the manual process of phoneme segmentation. This process is necessary because the working unit are syllables, made of phonemes. The syllable features (duration, energy and pitch) are extracted after the syllable time boundaries have been set. To this purpose, Sonic Visualizer software has been used. As an example of the process, Figure 3.3 shows the result of labeling phonemes for a sentence with the help of the spectrogram.

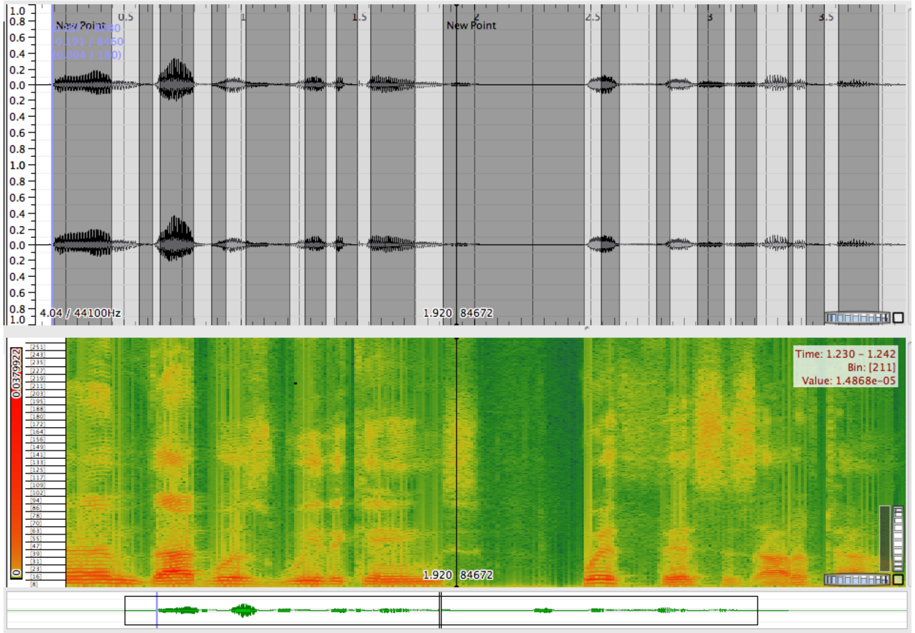


Figure 3.3: Phoneme labeling process sample

3.3.4 Syllable duration

Using the information from the phonetic transcription (section 3.3.2) and the phonetic segmentation (section 3.3.3), the information concerning the syllables duration has been extracted.

Speech rate

The mean value of syllable duration has been used to define the speech rate concept for each sentence.

3.3.5 Syllable pitch and energy

Pitch and energy have been extracted using the in-house SPPTools software. It allows to analyze a speech file in order to extract these two features frame by frame. Their mean energy and pitch have been computed using the corresponding frames per syllable. Figure 3.4 corresponds to the waveform, pitch and energy values computed by this software during the feature extraction process.

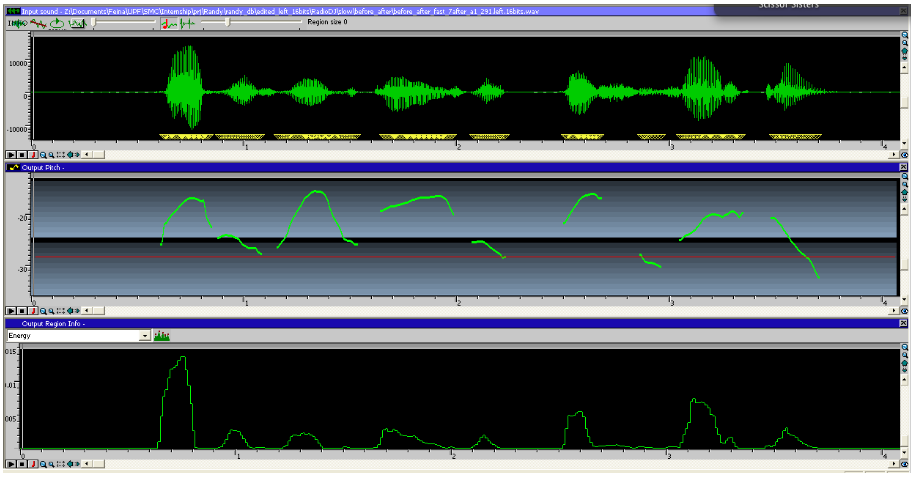

Input sound - Z:\Documents\Feina\UPF\SMC\Internship\prj\Randy\randy_db\edited_left_16bits\RadioDJ\slow\before_after\before_after_fast_7after_a1_291.left.16bits.wav
INFO
Region size 0
10000
0
-10000
1
2
3
4
Output Pitch -
-20
-30
Output Region Info -
Energy
015
.01
005

Figure 3.4: SPPTools: used to extract pitch and energy

Chapter 4

Methodology for modeling emotion

In order to model the emotion expressed in a sentence, different strategies where considered at the beginning project. An original sentence of the recorded database could be used in order to extract the control parameters of the synthesized sentence. This control, and therefore the synthesis, could be either performed at the keywords level (such as the band name, the song title, etc) or for the entire sentence. In the first case the control sentence and the synthesized keywords would be concatenated in order to generate the resulting sentence.

Another strategy, the one finally used in this project, has been to use the different utterances of the same sentence to build a model which predicts several control features used by the synthesizer for the entire sentence.

4.1 Simplifying the circumplex model of affect

For the purposes of the project, the circumplex model of affect referenced in 2.2.1 can be simplified. The activation dimension is mainly the one that discriminates between *calm* and *excited*. It is the one that makes the difference for the Radio DJ project in terms of these two emotions, which are highlighted in Figure 4.1. The Pleasant-Unpleasant dimension is not taken into account since it doesn't change in both emotion levels and, besides, in the context of a Radio DJ application, it would not be a common situation, for example, a *sad* speaker. This speaker is supposed to be in a formal and rather *pleasant* mood.

Thus, by taking the activation dimension and adding speech rate as an extra dimension to be modeled, the model of affect that is taken into account in this project is shown in Figure 4.2. The depicted dots represent the training data in this simplified 2D emotion control space, where the intention is to cover different utterances of the sentence in terms of activation and speech rate. This simplification of the circumplex model of affect is the control space of the emotion model.

4.2 Emotion modeling workflow

As introduced in the beginning of this chapter, a model per sentence can be built by using all utterances of the same sentence to predict several control parameters of the synthesized sentence. In section 3.3.1, it has been explained that these utterances have the particularity that have been

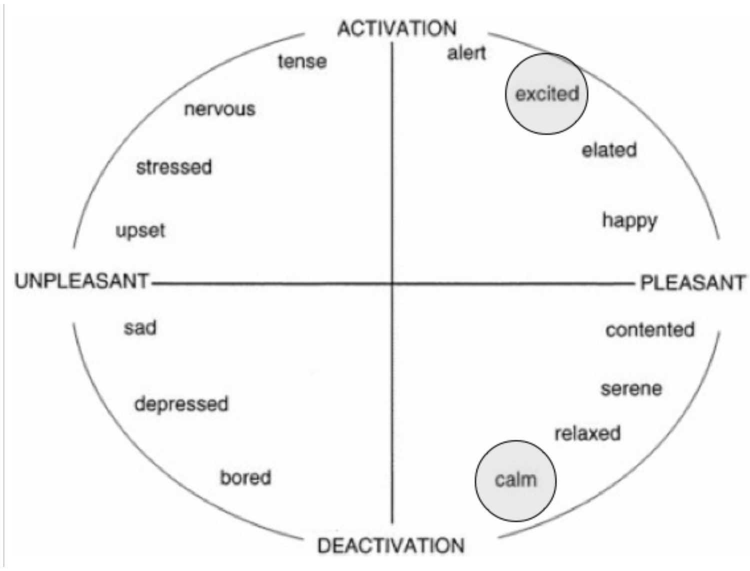


Figure 4.1: Circumplex model of affect activation dimension

recorded with different emotion conditions, both in terms of speech rate and activation dimension of the circumplex model of affect.

4.2.1 Training the model

Concerning the emotion modeling after the phoneme labeling process, three features are extracted per syllable in order to model the control space, the most commonly used amongst the ones introduced in the state of the art. In this project, the speech rate is directly related to the syllable duration and computed the mean of all syllable durations in a sentence. On the other hand, energy and pitch are related to the speaker emotion, in our case the activation dimension. These 3 features are used to train the model output according to an input vector representing the activation and the mean syllable speech rate. The main workflow of the modeling is shown in Figure 4.3 and described in the current and following sections.

In order to perform the train and test (generation of predictions) steps, the labeled data described in section 3.3 has been used. In the literature algorithms such as GMMs, CART, LMM are used to model emotion with respect to acoustic correlates, as introduced in section 2.3. On the other hand, artificial neural networks (ANN) have also been used in singing voice synthesis to model emotion by means of the vibrato expressive factors. The idea in this project has been to evaluate how can ANNs model emotional speech by predicting the control parameters.

Neural networks with back-propagation have been used from the neural networks toolbox in *MATLAB* (more precisely, the *newff* function). The target is to model the relationship between the input feature vector (speech rate and activation) with respect to the output feature vector (syllable energy, pitch and duration). From the different configurations of hidden layers that were tested at the beginning, this parameter was set to 10. Its results are the ones that have been brought to evaluation. For coding purposes, the number of bands needs to be configured according to the amount of nodes in the output. This is sentence dependent and corresponds to 3 (that is to say, features) times

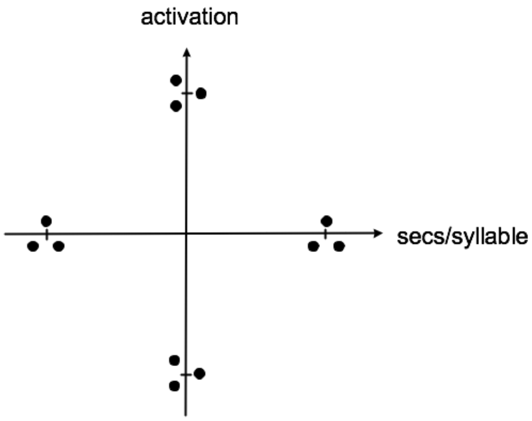


Figure 4.2: Simplified circumplex model of affect

the number of syllables. For instance, a sentence with 12 syllables configures the corresponding ANN to 36 nodes in the output. The backpropagation network training function has been set to the default one (*trainlm*), which updates weight and bias values according to Levenberg-Marquardt optimization. It is the fastest backpropagation algorithm in the toolbox, and is recommended as a first-choice supervised algorithm. The transfer function for the hidden layers has been set to the typical hyperbolic tangent sigmoid (*tansig*) and the output transfer function to a linear function (*purelin*).

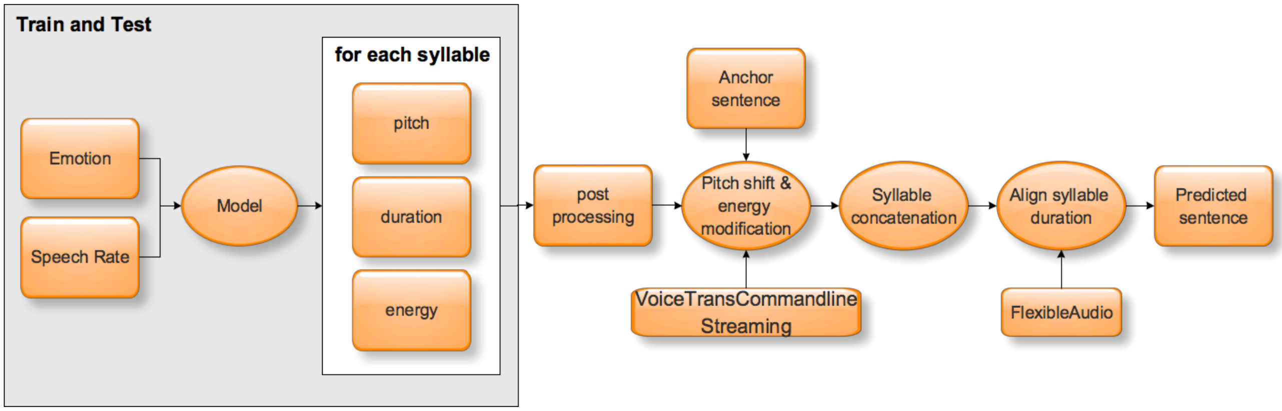


Figure 4.3: Workflow to generate predicted sentences

4.2.2 Generating predictions

Once the model has been trained, it has been tested by generating predictions of new output feature vectors giving new activation and speech rate conditions. Predictions have been generated with input values both from a similar range as the training values and also from new values of the 2D control space, in order to test if, a part from generating similar conditions to the training set, the space models correctly the unseen data in the training process. In figure 4.4 both regions have been highlighted in black (for the training region) and grey (for the unseen data) dots.

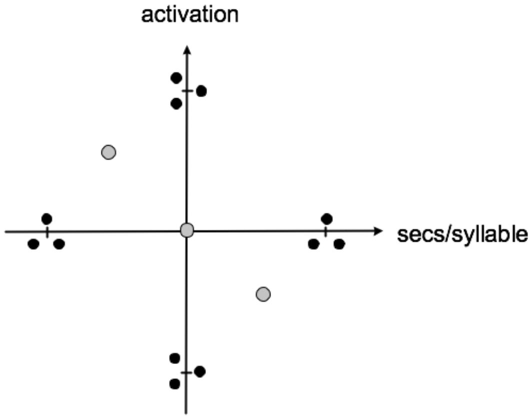


Figure 4.4: Predictions placed in both the train and unseen regions of the control space

4.2.3 Postprocessing the predictions

Predictions have been postprocessed since in many cases values fall outside a physically reasonable area. Sometimes these values are far too high or low values or even negatives, and don't make sense for syllable duration, energy or pitch. Also, energy and pitch values have been smoothed in order to have more continuous changes amongst syllables.

Checking right boundaries

In some cases, the output predictions of the model have been changed so that these are positive and between the right low and high boundaries. The criteria to set these boundaries has been by testing the quality of the output sound generated by software that performs the gain and pitch shifting. Thus, a pitch shifting factor lower than 0.25 or higher than 2.0 is not recommended. With respect to energy, its conversion factor boundaries have been set to 0.5 and 1.2.

Concerning the syllable duration, negative values have been replaced by the mean of the remaining positive values.

Ensuring smoothness

In order to ensure more continuous changes in energy and pitch than the real predicted values, these have been smoothed by taking the mean of the previous and next predicted values per feature.

Anchor sentence selection

The postprocessed predictions are applied to an anchor sentence from the real recording sentences of the database set. This sentence will be first modified in terms of pitch and energy with a specific software and afterwards in terms of syllable duration.

The selection of this anchor sentence for this project is manual, and can be done in two different ways. It can either be a fixed neutral sentence from our database or the closest sentence to the target prediction so that the modification produces less distortion as possible.

4.2.4 Pitch shifting and energy modification

In order to perform the pitch shifting and energy modification, the software described in [Bon08] has been used, which, as an example, was used in [MBJ09].

This software has been used to obtain another version of the anchor sentence with the target pitch and energy. Since each syllable has its own target values, the algorithm generates as many transformations of the anchor sentence as syllables. These are processed in the next step.

4.2.5 Time domain transformations

Once the previous step has generated as many pitch-shifted versions of the anchor sentence as syllables, the right transformed syllables in each one have to be processed in order to get the final transformed sentence. This last process involves syllable concatenation and time stretching.

Syllable concatenation

From each transformed version of the anchor sentence, only a particular syllable is necessary for our purpose. Thus, the target syllables are concatenated in order to generate a first version of the target sentence.

Time stretching to align syllable duration

The remaining step is to modify the syllables' durations in order to match the predictions. To this purpose, an in-house algorithm has been used. *FlexibleAudio* is described in [Bon02]. This step enlarges or shortens the time duration of a syllable according to the predictions of the model.

4.2.6 Objective tests

The training process of the model and the predictions have been objectively evaluated with 2 sentences and the results are shown from Tables E.1 to E.12. The evaluation computed the ratio between the predicted values per syllable and the mean value of the corresponding feature of the training sentences for that emotion class. Also, the mean of these values are presented sentence-wise as well as syllable-wise. The lower-right value is the overall mean. The closer the ratios are to 1, the closer is the prediction to the mean value of the corresponding feature for that sentence.

Dataset

Both the perceptual and the objective evaluation have the same generated sentences in common. This set of generated sentences have the control parameters showed in Table 4.1 as input of the trained model for the predictions. The vector columns is the coded information showed in Table 3.4. Coding sentences with emotions from the training set is straightforward. Unseen data has been coded by considering it as a point with contributions of different moods in the 2D control space (see 4.4).

	Activation		***Speech Rate***
Sample ID	**Name**	**Vector**	**Mean seconds/syllable**
1	energetic	1.0,0.0,0.0,0.0	0.19
2	fast	0.0,0.1,0.0,0.0	0.20
3	slow	0.0,0.0,0.1,0.0	0.29
4	calm	0.0,0.0,0.0,0.1	0.25
5	energetic-fast	0.5,0.5,0.0,0.0	0.18
6	slow-calm	0.0,0.0,0.5,0.5	0.30
7	all	0.25,0.25,0.25,0.25	0.22

Table 4.1: Control parameters of the evaluated sentences

The output predicted values of the model where applied to 2 sentences, one corresponding to a sentence that would be placed before the song it is presenting, and the other one after the song. Also, two anchor sentences have been used for each set of predictions, corresponding to the 2 strategies explained in section 4.2.3. Thus, one anchor sentence is common for all predictions and belongs to a rather neutral and calm utterance from that sentence in the training set. The other anchor sentence depends on the activation, so it changes from prediction to prediction.

Using an anchor sentence close to the target

The objective test ratios results using a different anchor sentence which is close to the target for each transformation are presented in Tables E.1 to E.6 in Appendix E. The mean values per sentence and syllable are also computed in order to have and overall idea of the model performance.

Figures 4.5 to 4.7 show some examples of a sentence prediction values and the training values for that class. The corresponding legends show which are the training examples, the predictions of the model and the corresponding corrected values in terms of boundaries and smoothness. In general, for these selected examples, the predicted values fall inside the training feature values per syllable. This, of course, doesn't mean that the predicted values doesn't also fall inside other emotions' range values.

Using the same neutral-calm anchor sentence for all transformations

The objective test ratios results using a neutral-calm anchor sentence for the transformation are presented in Tables E.7 to E.12 in Appendix E. The mean values per sentence and syllable are also computed in order to have and overall idea of the model performance.

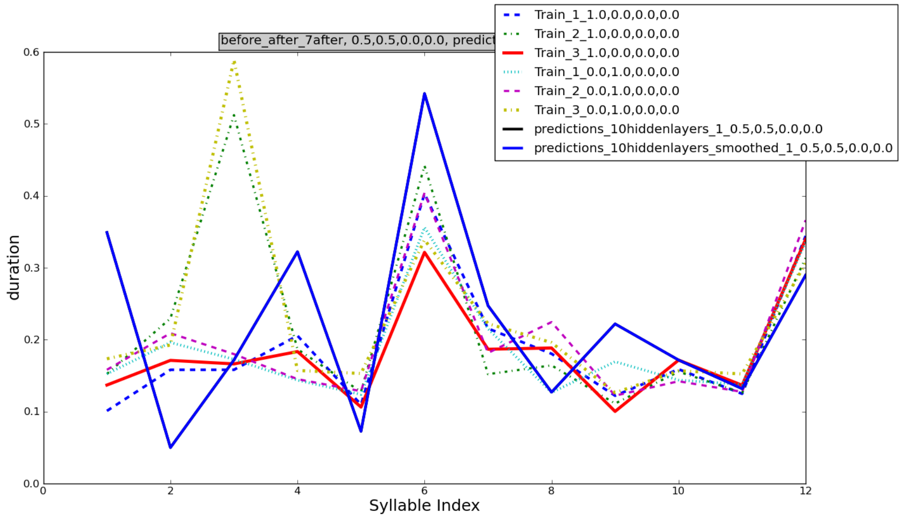


Figure 4.5: Syllable duration prediction for sentence 2: fast-excited emotion with 0.18 sec/syll as speech rate, close anchor

Figures 4.8 to 4.10 to show some examples of a sentence prediction values and the training values for that class. The corresponding legends show which are the training examples, the predictions of the model and the corresponding corrected values in terms of boundaries and smoothness.

Objective test results summary

Tables 4.2 to 4.3 summarize the test tables shown in appendix E. Just the sentence mean values are shown in this case. From the overall means (last row) it can be observed that in general the predictions seem to be too far from the train class mean, except for duration and pitch in sentence 2.

	Close anchor			***Neutral anchor***		
Sample ID	**Duration**	**Energy**	**Pitch**	**Duration**	**Energy**	**Pitch**
1	1.51	1.32	1.03	1.51	0.15	1.04
2	1.69	1.58	1.51	1.69	0.22	1.50
3	1.27	2.13	1.63	1.27	0.58	1.66
4	0.96	0.61	1.39	0.96	0.61	1.39
5	1.35	1.24	1.45	1.35	0.16	1.42
6	1.25	2.94	1.86	1.25	0.68	1.87
7	1.25	0.97	1.55	1.25	0.25	1.56
Mean	1.33	1.54	1.49	1.33	0.38	1.49

Table 4.2: Summary of the mean values for all features in sentence 1

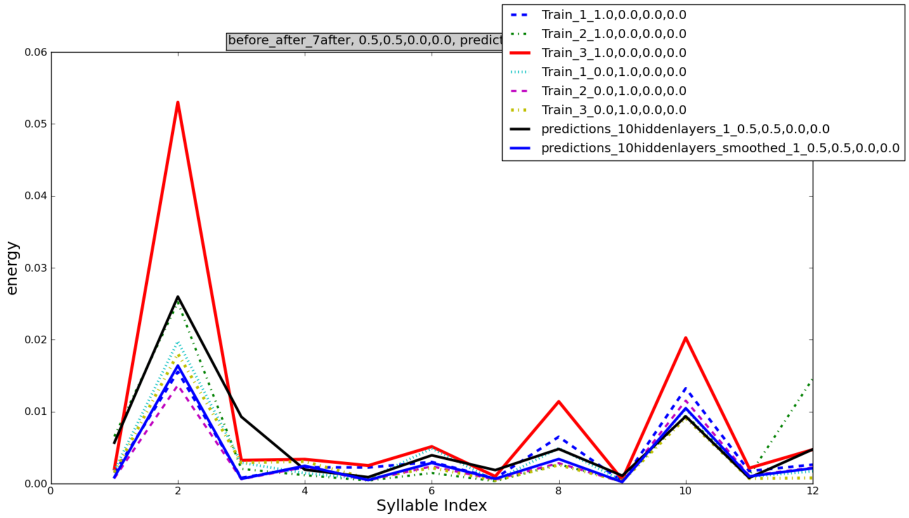


Figure 4.6: Syllable energy prediction for sentence 2: fast-excited emotion with 0.18 sec/syll as speech rate, close anchor

	Close anchor			*Neutral anchor*		
Sample ID	**Duration**	**Energy**	**Pitch**	**Duration**	**Energy**	**Pitch**
1	1.09	1.12	1.24	1.09	0.25	1.04
2	1.10	0.76	1.10	1.10	0.40	0.92
3	0.85	0.43	0.99	0.85	0.93	1.05
4	0.81	0.50	1.25	0.81	0.50	1.25
5	1.15	0.69	1.13	1.15	0.30	1.00
6	0.75	0.60	1.07	0.75	0.60	1.07
7	0.82	0.13	1.12	0.82	0.31	1.10
Mean	0.94	0.60	1.13	0.94	0.47	1.06

Table 4.3: Summary of the mean values for all features in sentence 2

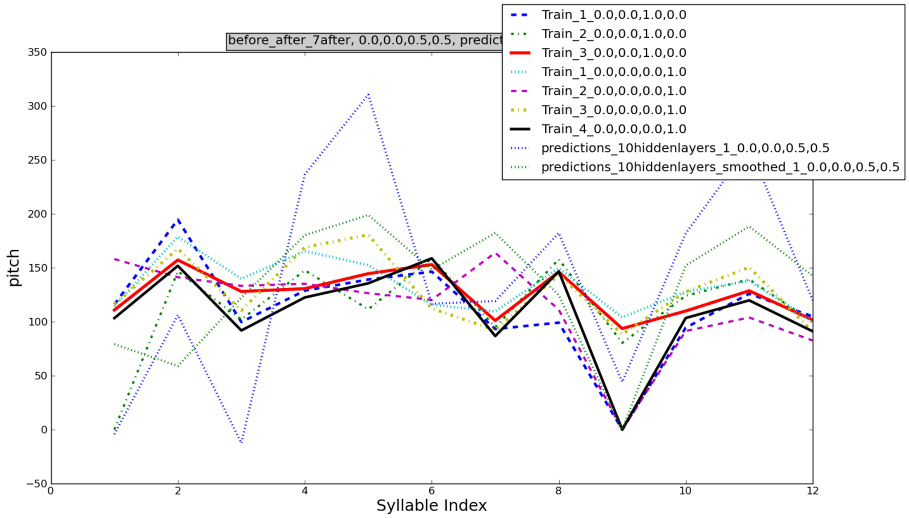


Figure 4.7: Syllable pitch prediction for sentence 2: calm-slow emotion with 0.30 sec/syll as speech rate, close anchor

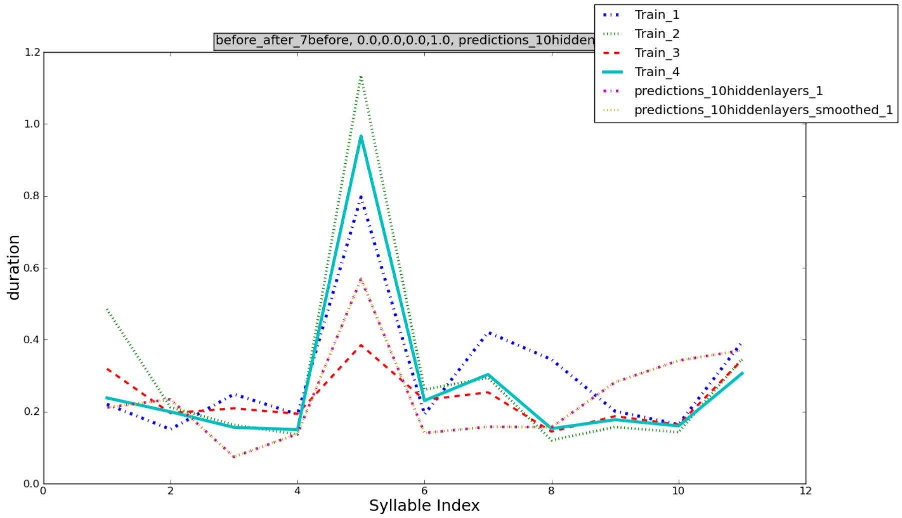


Figure 4.8: Syllable duration prediction for sentence 1: calm emotion with 0.25 sec/syll as speech rate, calm-fixed anchor

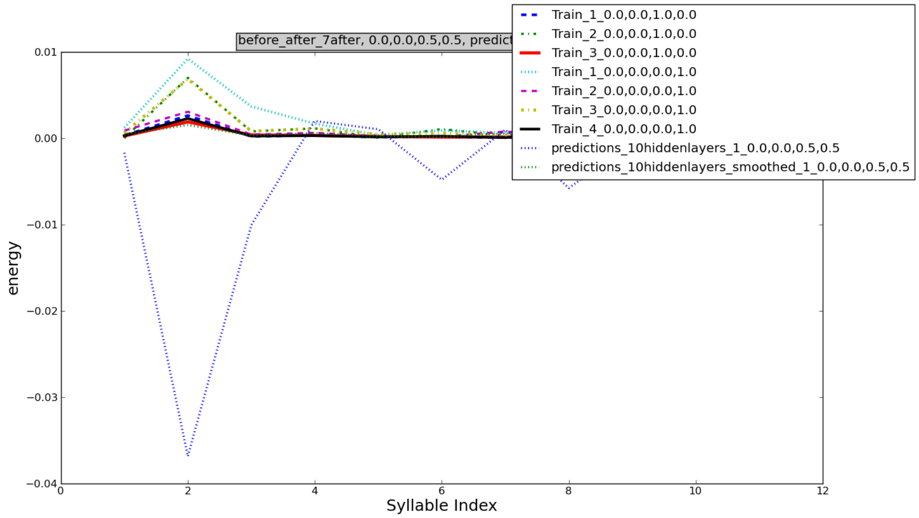


Figure 4.9: Syllable energy prediction for sentence 2: calm-slow emotion with 0.30 sec/syll as speech rate, calm-fixed anchor

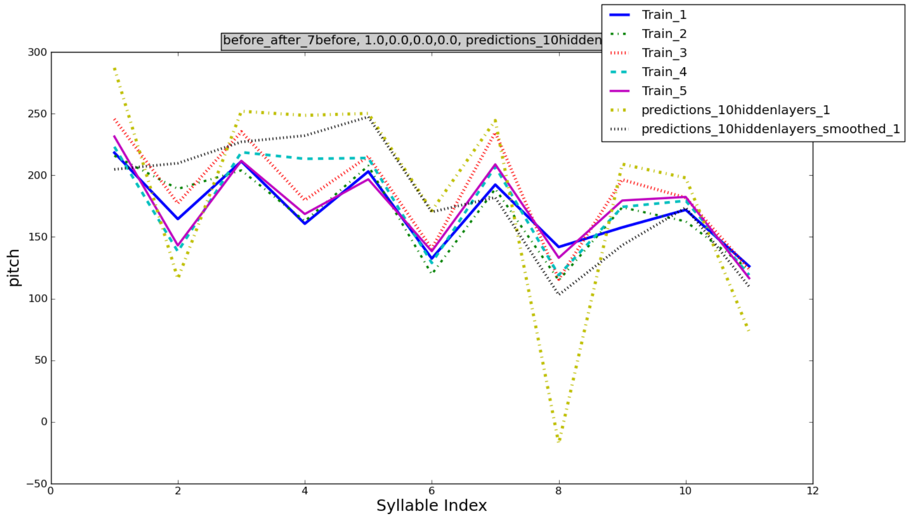


Figure 4.10: Syllable pitch prediction for sentence 1: excited emotion with 0.19 sec/syll as speech rate, calm-fixed anchor

Chapter 5

Evaluation and Results

This chapter focuses on the evaluation of the generated sentences and presents its results. A part from the objective tests performed presented in section 4.2.6, a perceptual evaluation has been carried out. More precisely, it has been evaluated if the same set of evaluated sentences in the objective test convey the intended emotion to the transformed sentence according to group of 10 evaluators.

5.1 Evaluation Methodology

5.1.1 Perceptual test

The perceptual test asked the listeners to rate how did they perceive each sentence in terms of activation and speech rate separately. Activation could be rated has been rated between 1 (calm) and 5 (excited). Speech rate has been rated between 1 (slow) and 5 (fast).

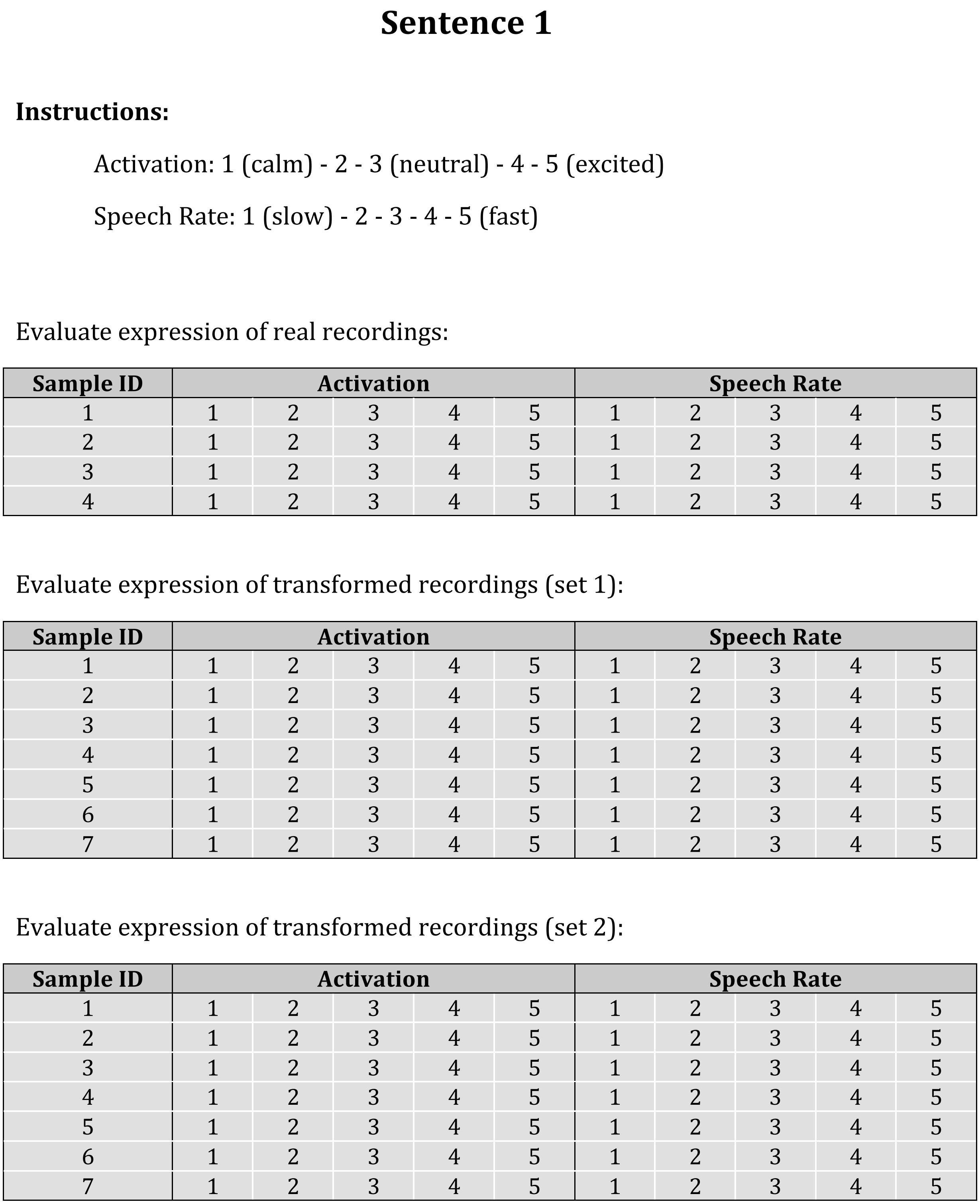
For each of the 2 sentences that have been evaluated, 3 subsets where rated:

- A first set with 4 original recordings corresponding to each of the 4 regions in the training set. This set has been evaluated in order to know if what has been considered as ground-truth corresponds to what the evaluators perceive.
- The second set corresponds to each of the 7 transformations presented in Table 4.1, using as anchor sentence a neutral-calm one.
- Similiarly to the previous one, the third set uses as anchor sentence a close one to the target in terms of activation and speech rate.

The dataset used to evaluate has been described in section 4.2.6

Test answer sheet

Figure 5.1 is a sample of the answer sheet for the perceptual evaluation of one of the 2 evaluated sets. Each section corresponds to one of the subsets. The real evaluation contained another page with the same charts to evaluate the second sentence transformations.



Sentence 1

Instructions:

Activation: 1 (calm) - 2 - 3 (neutral) - 4 - 5 (excited)

Speech Rate: 1 (slow) - 2 - 3 - 4 - 5 (fast)

Evaluate expression of real recordings:

Sample ID	Activation					Speech Rate				
1	1	2	3	4	5	1	2	3	4	5
2	1	2	3	4	5	1	2	3	4	5
3	1	2	3	4	5	1	2	3	4	5
4	1	2	3	4	5	1	2	3	4	5

Evaluate expression of transformed recordings (set 1):

Sample ID	Activation					Speech Rate				
1	1	2	3	4	5	1	2	3	4	5
2	1	2	3	4	5	1	2	3	4	5
3	1	2	3	4	5	1	2	3	4	5
4	1	2	3	4	5	1	2	3	4	5
5	1	2	3	4	5	1	2	3	4	5
6	1	2	3	4	5	1	2	3	4	5
7	1	2	3	4	5	1	2	3	4	5

Evaluate expression of transformed recordings (set 2):

Sample ID	Activation					Speech Rate				
1	1	2	3	4	5	1	2	3	4	5
2	1	2	3	4	5	1	2	3	4	5
3	1	2	3	4	5	1	2	3	4	5
4	1	2	3	4	5	1	2	3	4	5
5	1	2	3	4	5	1	2	3	4	5
6	1	2	3	4	5	1	2	3	4	5
7	1	2	3	4	5	1	2	3	4	5

Figure 5.1: Perceptual evaluation: Answer sheet

5.2 Results

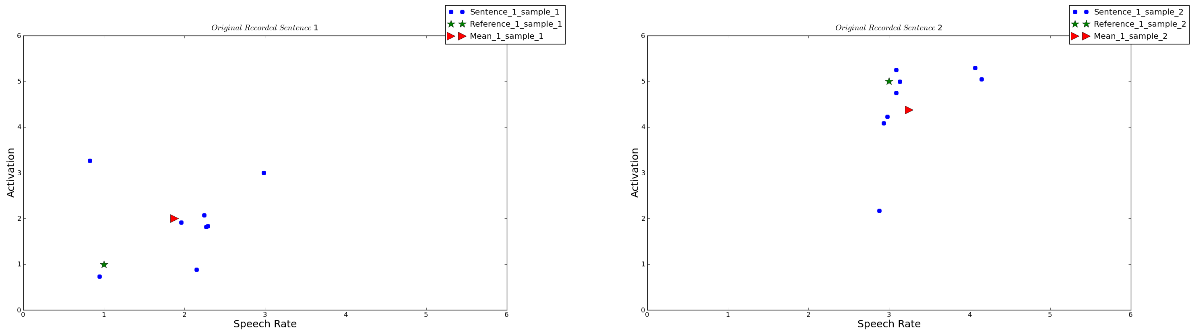
5.2.1 Perceptual test

The listeners' ratings have been processed and grouped according to the sentence these belong to. These have been compared to the manually annotated ground-truths ratings in order to check how these differ in a 2D plane.

Ratings of the real recordings

In order to understand the following figures, the x axis represents the speech rate and the y axis the activation. The ground-truth point can be compared to the rated values of each evaluator and the mean rated value.

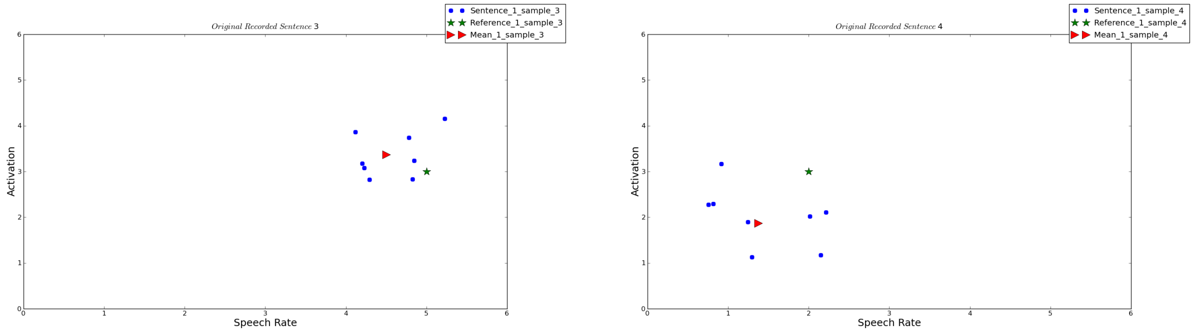
Figures 5.2 to 5.3 correspond to the ratings of the 1st sentence, and Figures 5.4 to 5.5 to the 2nd sentence.



(a) Calm

(b) Energetic

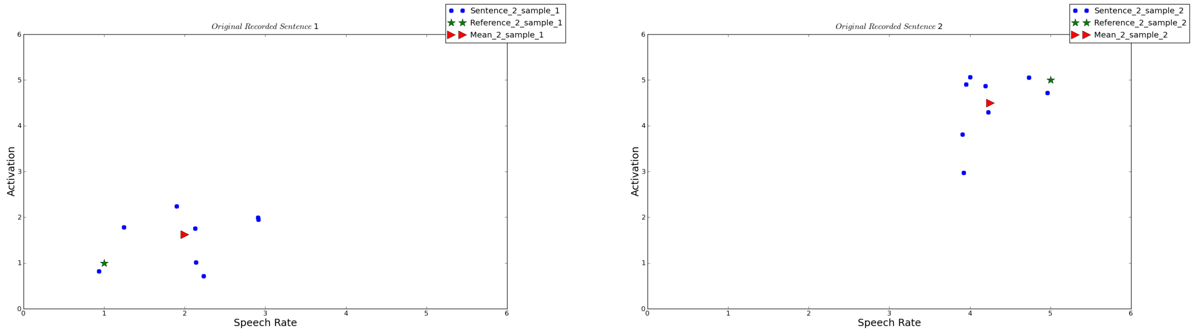
Figure 5.2: Perceptual evaluation: sentence 1 ratings of real recordings for calm and energetic samples



(a) Fast

(b) Slow

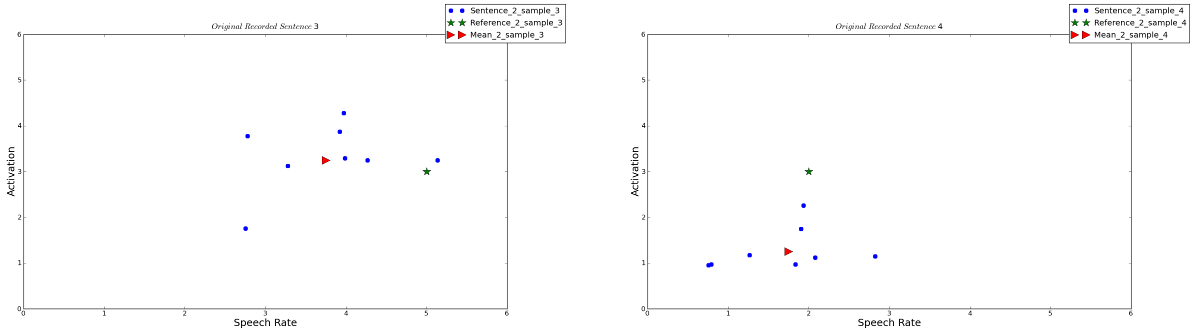
Figure 5.3: Perceptual evaluation: sentence 1 ratings of real recordings for fast and slow samples



(a) Calm

(b) Energetic

Figure 5.4: Perceptual evaluation: sentence 2 ratings of real recordings for calm and energetic samples



(a) Fast

(b) Slow

Figure 5.5: Perceptual evaluation: sentence 2 ratings of real recordings for fast and slow samples

Comparing ratings for transformations using neutral or close anchor sentence

This section compares the ratings of the transformations based on the anchor sentence. Each figure from 5.6 to 5.12 has 2 subplots, the one corresponding to the calm-fixed (neutal) anchor sentence and the close anchor transformation.

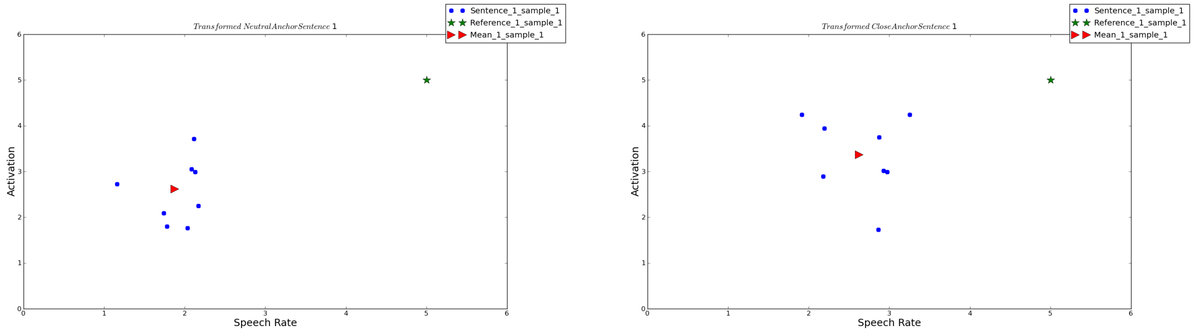
Summarizing the errors

To summarize the visualized information from Figures 5.6 to 5.12, the error (difference) between the mean rated values and the ground-truth has been computed. Separately for activation and speech rate, Figures 5.13 and 5.14 show the error per evaluated sentence (1 to 4 correspond to the training data evaluation, 5-11 to the first set and 12-18 to the second one).

Finally, to make it easier, the module of the error between the mean rated values and the ground-truth has been computed and is shown in Table 5.1.

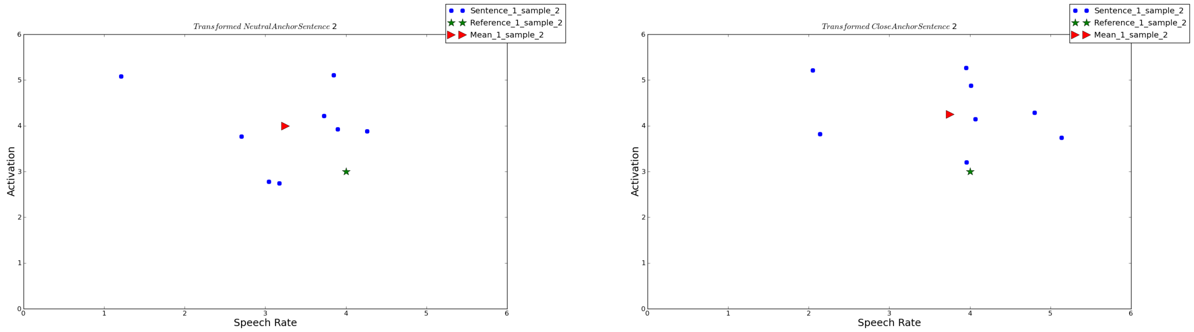
5.2.2 Correlation between perceptual and objective results

In order to check whether there is any relationship between the rated speech rate and real one from each sentence, this information has been represented in Figure 5.15. Although there are not enough



(**a**) Neutral (**b**) Close

Figure 5.6: Perceptual evaluation: Ratings for sample ID 1



(**a**) Neutral (**b**) Close

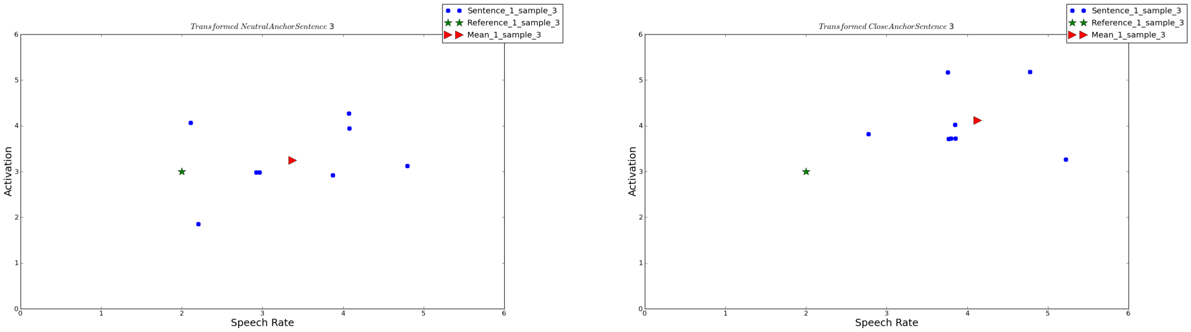
Figure 5.7: Perceptual evaluation: Ratings for sample ID 2

	Sentence 1		***Sentence 2***	
Sample ID	**Neutral Anchor**	**Close Anchor**	**Neutral Anchor**	**Close Anchor**
1	3.93	2.88	2.02	1.79
2	1.25	1.27	0.73	0.80
3	1.40	2.40	1.59	0.84
4	3.25	2.77	1.40	2.06
5	2.46	1.88	1.57	1.91
6	2.70	2.74	2.76	3.09
7	0.18	1.41	1.67	1.62
Mean	2.16	2.19	1.68	1.73

Table 5.1: Perceptual evaluation: error modules from the mean rating to ground-truth

sample points to take definite conclusions, there seems to be a slight tendency of extreme high and low speed rates to be evaluated as such. However, most sentences were rated in average between 3 and 4 points.

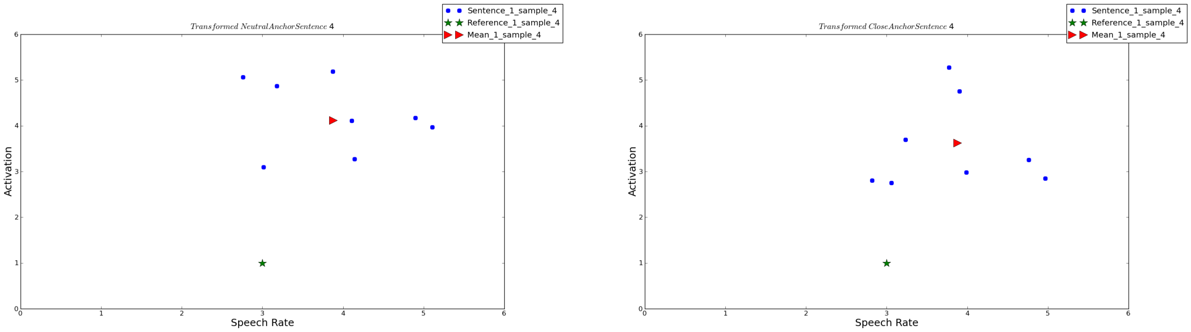
In a similar way, with respect to activation, it has been checked whether there is any kind of relationship between its ratings and the mean predicted energy of the syllables. Figure 5.16 shows there is a slight trend for high mean energy values and activation. However, there probably too few values to extract this conclusion, a part from the fact that most energies are located in the low part of the graph no matter what the rating is. These are mainly located between 2 and 4 points.



(a) Neutral

(b) Close

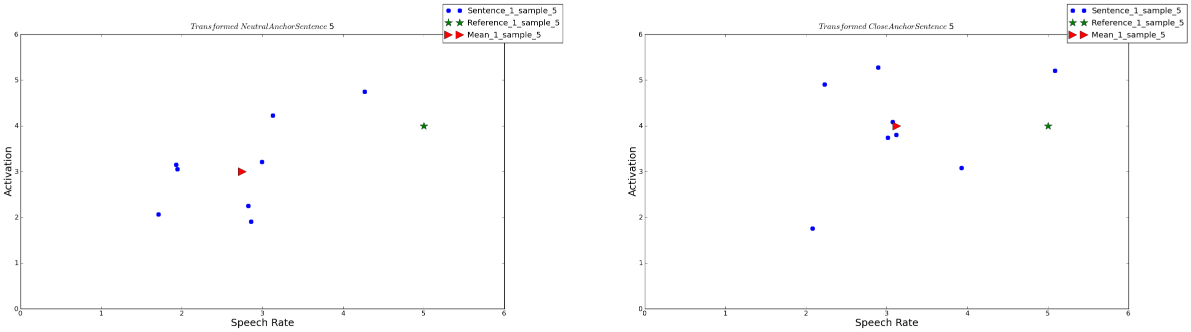
Figure 5.8: Perceptual evaluation: Ratings for sample ID 3



(a) Neutral

(b) Close

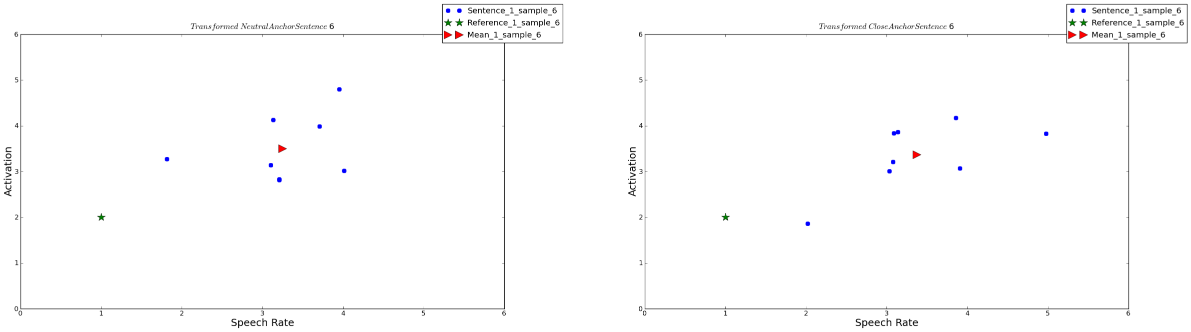
Figure 5.9: Perceptual evaluation: Ratings for sample ID 4



(a) Neutral

(b) Close

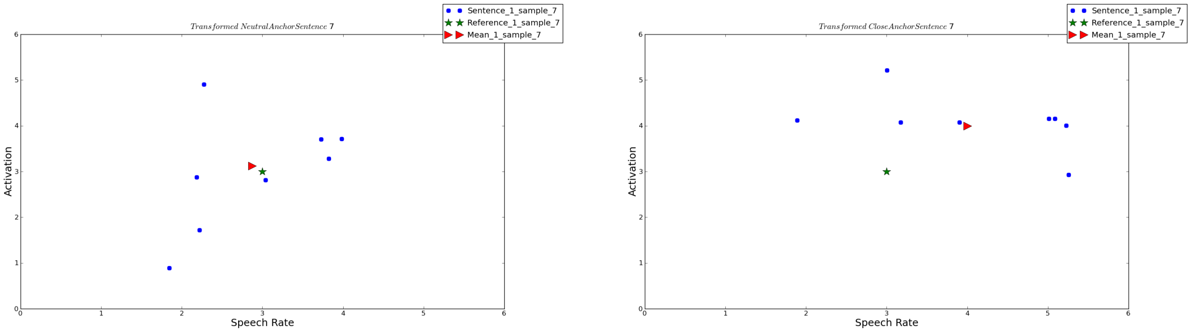
Figure 5.10: Perceptual evaluation: Ratings for sample ID 5



(**a**) Neutral

(**b**) Close

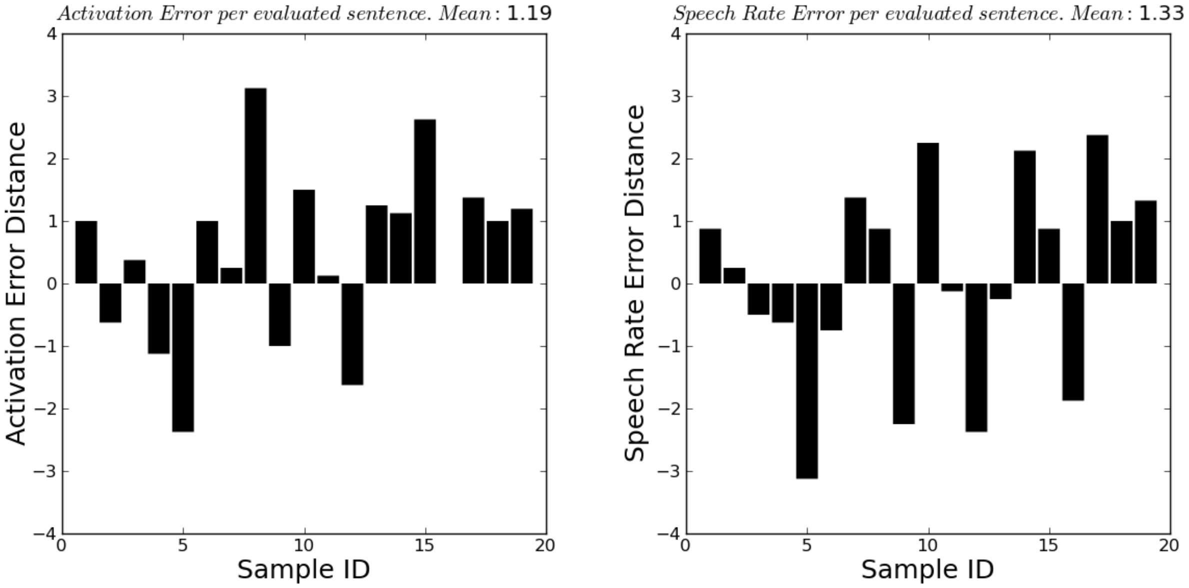
Figure 5.11: Perceptual evaluation: Ratings for sample ID 6



(**a**) Neutral

(**b**) Close

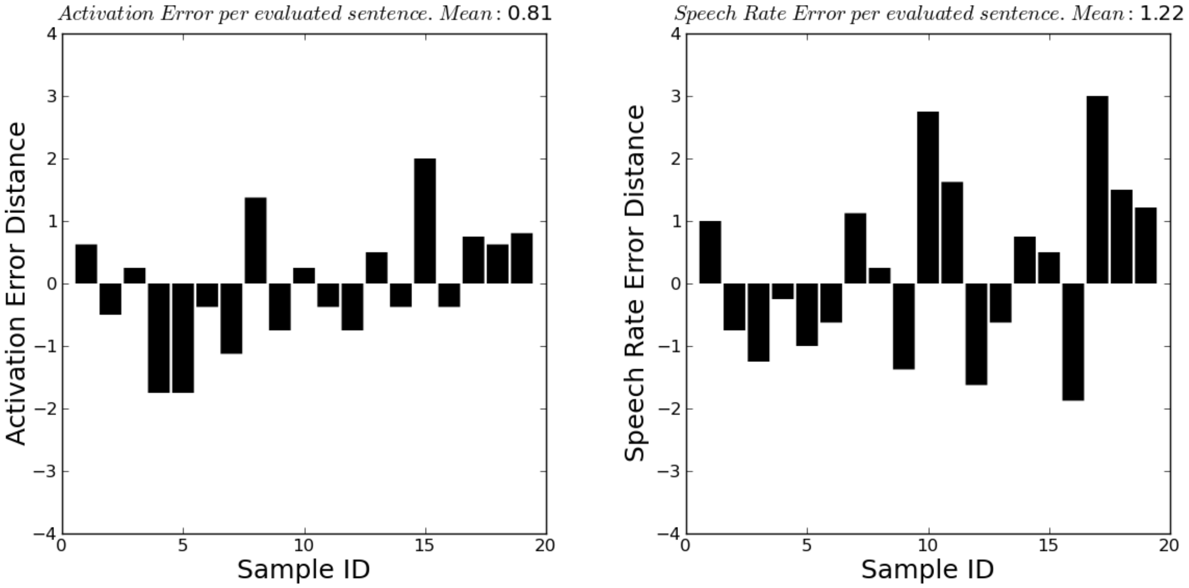
Figure 5.12: Perceptual evaluation: Ratings for sample ID 7



(a) Activation

(b) Speech Rate

Figure 5.13: Perceptual evaluation: sentence 1 error distance



(a) Activation

(b) Speech Rate

Figure 5.14: Perceptual evaluation: sentence 2 error distance

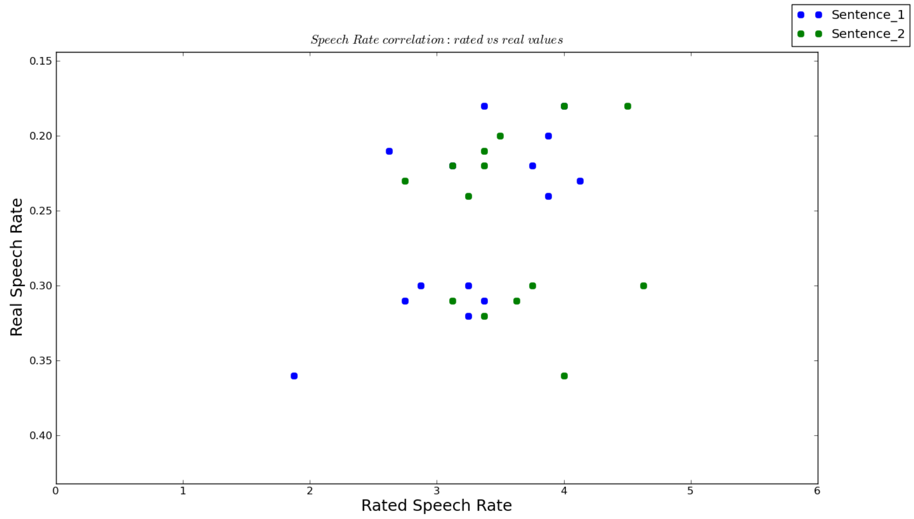


Figure 5.15: Speech rate correlation between rated and real values

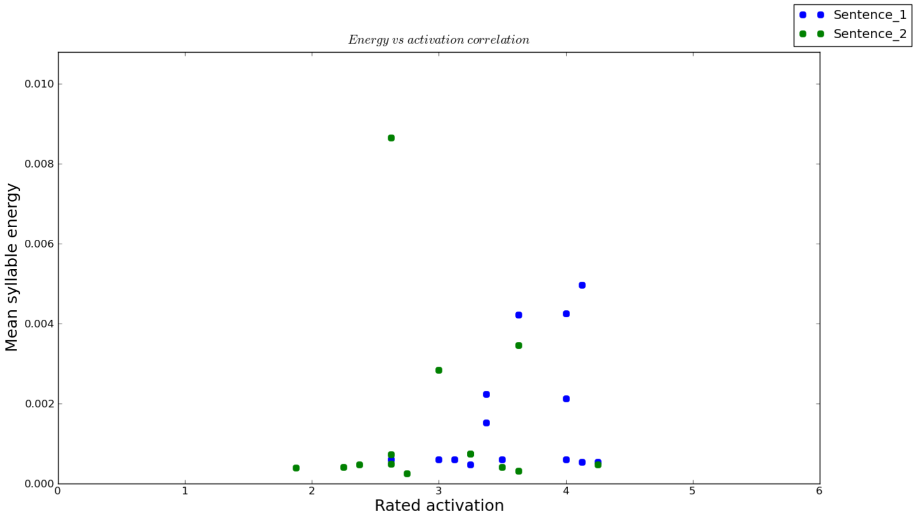


Figure 5.16: Activation vs energy correlation

Chapter 6

Future work and conclusions

6.1 Future Work

This section summarizes the research work that could be done to improve the current work. Also, from an application perspective, some ideas are given.

6.1.1 Research work

From the research point of view, the following aspects could be taken into account:

- Improve the models by labeling more data and tuning. The current model just takes around 3 recordings per class, which means an overall amount of 12 sentences. This is probably not enough to generate a reliable model. Recording other non-professional speakers would help to increase the amount of training data.

- Label ground-truth data in a perceptual way using tools such as FEELTRACE, described in [SCDc$^+$01] and introduced in section 2.3.1. This idea could be adapted by using evaluators' ratings to original sentences as ground-truth.

- Modeling emotions with other algorithms should be tried, such as support vector machines and gaussian processes. This would help to optimize and find better implementations of the final solution.

- Other control parameters of the model should be studied. It currently takes into account the speech rate in seconds per syllable and the expression activation. Maybe linguistic information concerning the keywords could be considered.

- A sentence independent model to generate the predictions would be also useful. This way, other sentences than the ones recorded in the initial database could be synthesized with some emotional content.

- The pitch envelope would also be modeled, a part from the current mean value for each syllable. This feature help to emphasize some words out of the sentence and, probably, increase intelligibility.

- Use a cost function in order to select the closest anchor sentence to the target predictions in order to apply to it the less transformations as possible.

- Also, test how the models improve performance by the fact of increasing the training data.

- Finally, the technologies and code used in this project belong to different operating systems. The binaries used for the pitch-shift and time-scale transformations were developed for Windows OS, meanwhile the rest has been developed under Mac OS. It would be a helpful and quicker to get results with an integrated framework under the same operating system.

6.1.2 Possible applications

From an application point of view, a couple of extensions could be done from the results of the thesis:

- UPF Radio: use the application in order to automatically generate radio shows.

- Music Player plug-in, in order to introduce the songs played in the computer.

6.2 Conclusions

The main conclusion of this project is that a proof of concept of what an emotional Radio DJ corpus could be has been set. Of course, a lot remains to be done in terms of data labeling, expression parameterization and variable model testing. Different conclusions can be considered with respect to the corpus design, the results, methodology as well as concerning the evaluation process.

Concerning the designed corpus, the structure of a basic Radio DJ show has been identified in terms of commonly used sentences and information given with respect to the presented songs. This corpus has been recorded in different emotional content. Once completely labeled, it could be used to control a speech synthesizer reproducing the same recorded sentences and moods, changing the relevant keywords to introduce any desired group, song, etc.

With regard to the methodology followed to model expression, more data would be desirable to train the models properly. Also, training other models than neural networks would be necessary in order to be able to compare their performances with different configurations. With regard to the emotion labeling, another emotion codification could have been used. This alternative coding could be -1 for calm emotion, 0 for slow and fast, and 1 for energetic, instead of a vector with 4 possible active coordinates. Concerning the basic unit for transformation, the syllable, other approaches could be tested, such as diphones, which is probably more related to the way speech synthesizer work in order to use control parameters during the unit selection instead of afterwards by means of transforming the synthesized utterance.

Since the speaker recorded was a professional singer, it would be more practical to record other speakers that would not represent a handicap in terms of cost. The differences that would appear in terms of pitch, energy, etc, should then be taken into account. Also, it has to be taken into account the fact that the recorded speaker was a singer but not an actor or a real radio speaker. Using a more familiarized speaker with the target task of the project would lead to better results. Besides, it has been particularly difficult to give the speaker concise instructions concerning how to express each emotion in the sentence's recordings, even with original radio samples as examples to listen to.

Concerning the results of the perceptual evaluation, 3 types of ratings have been asked. The first one checks if the concepts of activation and speech rate are similar to the design of the corpus. The mean values of the rated sentences in terms of speech rate and activation is closer in this type of evaluation than in the other two. This is reasonable since the recordings are real, non distorted and completely understandable.

The second and third ratings in the perceptual evaluation check whether the expressiveness pretended to produce with the model is also perceived by the listeners. The difference is the anchor sentence where the predictions are applied, which can be the same for all predictions, or the closer to the target emotion. Opposite to what was expected, the module of the error in Table 4.3 shows a slight lower value for evaluations with a neutral anchor sentence than with a close one to the target. Probably, the distortions that appear along the process with respect to energy, pitch and syllable duration are too much to allow the evaluation perceive the extra information given by the pitch curve uttered by the recording.

Another conclusion in the same topic is that the error on sentence 2 is lower. This is understandable since at this point evaluators were more familiarized with the evaluation process in the second sentence, knowing better what a 1 or 5 rate meant. Some of the listeners commented this fact after the task. Another explanation for this could be that, according to the objective test results in Tables 4.2 and 4.3, the ratios are closer to 1 for sentence 2 than for sentence 1, that is to say, closer to the mean values for each training samples.

From the objective evaluation test point of view, in sentence 1 it is slightly better to use a close anchor sentence to the target, where the energy feature makes the difference. In sentence 2, in terms of energy is better to use the close anchor, although in terms of pitch it is better the neutral sentence.

The correlation tests in order to check if evaluations and real values have similarities have shown that there is a slight tendency, although many ratings are centered around the rate 3. Extreme values confirm this tendency. In most cases, evaluators were not sure how to rate sentences, giving then a 3 rate in these situations.

Some other detail that could be improved with respect to the evaluation process is the fact that evaluators were in general non-native English speakers. Since the task was in this language, this could have been taken into account when selecting listeners. Some other feedback received from the evaluators is that in general the speech rate did not contain extreme values, so that it seemed too narrow to distinguish the concepts slow and fast.

Bibliography

[BLK03] J. Bonada, A. Loscos, and H. Kenmochi. Sample-based singing voice synthesizer by spectral concatenation. In *Proceedings of Stockholm Music Acoustics Conference*. Citeseer, 2003.

[BN08] M. Bulut and S. Narayanan. On the robustness of overall f0-only modifications to the perception of emotions in speech. *J. Acoust. Soc. Am*, 123:6, 2008.

[Bon02] J Bonada. Audio time-scale modification in the context of professional post-production, 2002.

[Bon08] J Bonada. Wide-band harmonic sinusoidal modeling, 2008.

[BS07] J Bonada and X. Serra. Synthesis of the singing voice by performance sampling and spectral models; ieee signal processing magazine. 24:67–79, 2007.

[CDPD$^+$04] S. Canazza, G. De Poli, C. Drioli, A. Roda, and A. Vidolin. Modeling and control of expressiveness in music performance. *Proceedings of the IEEE*, 92(4):686–701, 2004.

[CLB$^+$00] P. Cano, A. Loscos, J. Bonada, M. De Boer, and X. Serra. Voice morphing system for impersonating in karaoke applications. In *Proceedings of the 2000 International Computer Music Conference*. Citeseer, 2000.

[Coo91] P.R. Cook. Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing. *Electrical Engineering PhD Dissertation, Stanford University*, 1991.

[dA09] N. d' Alessandro. *Realtime and Accurate Musical Control of Expression in Voice Synthesis*. Phd thesis, University of Mons, 2009.

[GL08] H.Y. Gu and Z.F. Lin. Mandarin singing voice synthesis using ann vibrato parameter models. In *2008 International Conference on Machine Learning and Cybernetics*, pages 3288–3293, 2008.

[IT06] G. Ilie and W.F. Thompson. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, 23(4):319–330, 2006.

[JBB06] J. Janer, J. Bonada, and M. Blaauw. Performance-driven control for sample-based singing voice synthesis. In *Proc. DAFx*, volume 6, pages 42–44. Citeseer, 2006.

[KB09] Bernd J. Kröger and Peter Birkholz. Articulatory synthesis of speech and singing: State of the art and suggestions for future research. pages 306–319, 2009.

[KO07] H. Kenmochi and H. Ohshita. Vocaloid–commercial singing synthesizer based on sample concatenation. *Interspeech*, 2007.

[KOK+09] T. Kako, Y. Ohishi, H. Kameoka, K. Kashino, and K. Takeda. Automatic identification for singing style based on sung melodic contour characterized in phase plane. In *Proceedings of the 10th Int. Conference on Music Information Retrieval*, 2009.

[LB04] A. Loscos and J. Bonada. Emulating rough and growl voice in spectral domain. In *International Conference on Digital Audio Effects (DAFx'04), Naples, Italy*. Citeseer, 2004.

[Mae09] E Maestre. Modeling instrumental gestures: an analysis/synthesis framework for violin bowing, 2009.

[MBJ09] O Mayor, J Bonada, and J Janer. Kaleivoicecope: Voice transformation from interactive installations to video-games, 2009.

[MPR10] Luca Mion, Giovanni De Poli, and Ennio Rapanà. Perceptual organization of affective and sensorial expressive intentions in music performance. *ACM Trans. Appl. Percept.*, 7(2):1–21, 2010.

[NG] T. Nakano and M. Goto. Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation.

[PRP05] J. Posner, J.A. Russell, and B.S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(03):715–734, 2005.

[Rod02] X. Rodet. Synthesis and processing of the singing voice. In *Proceedings of the 1stieee benelux workshop on model based processing and coding of audio (mpca)*. Citeseer, 2002.

[Rus80] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[SB04] K.R. Scherer and T. Bänziger. Emotional expression in prosody: a review and an agenda for future research. In *Speech Prosody 2004, International Conference*. Citeseer, 2004.

[SCDc+01] Marc Schröder, Roddy Cowie, Ellen Douglas-cowie, Machiel Westerdijk, and Stan" Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. *IN: PROCEEDINGS EUROSPEECH*, 1:87–90, 2001.

[Sch95] K.R. Scherer. Expression of emotion in voice and music. *Journal of Voice*, 9(3):235–248, 1995.

[SGUA07] T. Saitou, M. Goto, M. Unoki, and M. Akagi. Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices. *Proc. WASPAA2007*, pages 215–218, 2007.

[SUA05] T. Saitou, M. Unoki, and M. Akagi. Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis. *Speech communication*, 46(3-4):405–417, 2005.

[TKL06] J. Tao, Y. Kang, and A. Li. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1145–1154, 2006.

[TSS$^+$06] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and Schröd. Modeling personality features by changing prosody in synthetic speech. In *Proceedings of Speech Prosody*. Citeseer, 2006.

[WG04] G. Widmer and W. Goebl. Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33(3):203–216, 2004.

Appendix A

Example: Song Label structure

This appendix shows the data structure used to label the keywords based on a syntax of 2 levels of labels, that could be formalized as *$label1.label2(keyword)*. In this example, the keywords belong to the *level 1* of the concept of *song*.

The information is presented with 3 indentation levels. The first one corresponds to the higher level label, *song*. The second level corresponds to the second label, which varies according to the specific aspect of level 1 that the keyword refers to. For example, one of these second level labels is *description*. In both cases, the first and second level labels, information about the number of times it appears (*n_elem*) and its conversion to frequency (*freq*) is given. Finally, the third level corresponds ti the actual tagged keywords and the number of times it appears.

```
song: n_elem(189) - freq(0.196261682243)
        amount: n_elem(3) - freq(0.00311526479751)
                200: 1
                4: 1
                50: 1
        description: n_elem(11) - freq(0.0114226375909)
                beautiful: 1
                brand new hot download: 1
                elegant: 1
                fantastic: 1
                full of incredible beauty and mystery: 1
                heart stopping: 1
                huge massive early nineties anthem: 1
                melodic: 1
                slamming: 1
                stunning: 1
                thrill ride: 1
        format: n_elem(1) - freq(0.00103842159917)
                mp3: 1
        genre: n_elem(2) - freq(0.00207684319834)
                impromptus: 1
                old time balades: 1
        name: n_elem(103) - freq(0.106957424714)
                Address the silence: 1
                All about our lives: 1
                All time low: 1
                Ascending bird: 1
                Bittersweet: 1
                Burning for you: 1
                Can't Stand It: 1
                Clock Radio: 1
                Concentrate: 1
                Cringe Or Smile: 1
```

```
Deborah: 1
Does not suffice: 2
Dr. Doctor: 2
Dramatas persona: 1
Dreamer's bay: 1
Emotional: 1
Everytime: 1
Fears wars the midnight storm: 1
Folling in to love: 1
Gem: 1
Give it up: 1
Glocken: 2
Green Withins Brook: 1
Growing up going down: 1
Have one me: 1
Hey: 1
Hey Kid: 1
Heys: 1
I am trying to break your heart: 1
I don't care: 1
I learned the hard way: 1
I'm Not Bad, I'm Just Drawn That Way: 1
I'm your goddess: 1
Impromptu number 4 in C# minor: 2
Leaves eclipse the light: 2
Liar Liar: 1
Like a Rocket: 1
Loud and clear: 1
Lovers: 1
Lythium: 3
Man in a box: 1
Many Is The Time: 1
Mercury: 1
Midnight directives: 2
Nicker sources: 1
No Matter: 1
Nourishment: 1
O3O5: 1
Oh Susanne: 1
One More Song: 1
Open Colour: 1
Oxfor Decor: 1
Parade Of Punk Rock T-Shirts: 1
Rebel Hero: 1
Red Alert 2010: 1
Revenge on the radio: 1
Right of your live: 1
Rock and Ball: 1
Rocking chair: 2
Room 410: 1
Samba Geladio: 2
Sentimental: 1
Sign your name: 2
Singles soundtrack: 1
Sky High: 1
Sometimes The Bear Eats You: 1
Sometimes When You Lose, You Win: 1
Souljack: 1
Still stuck: 1
Sugar Skulls: 1
Tales from a Dj: 1
The Splendour: 2
This is out town: 2
Three beats: 1
Void versus Bob: 2
Walk Out On Me: 1
Wendy naked: 1
Whole wide world: 1
```

```
        Worlds Away: 1
        You must be out of your mind: 1
        _Mabon_: 1
        automatic live: 1
        best time to say good bye: 1
        courtyard: 1
        like lovers often will: 1
        pretending to love you: 1
        six pound tracey: 1
        special boy: 1
        where ever you are: 1
reference: n_elem(67) − freq(0.0695742471443)
        _tune_: 1
        cover: 2
        cover version: 2
        final song: 1
        masterpiece: 1
        piece: 1
        remix: 4
        remixes: 1
        rendition: 1
        single: 4
        song: 33
        songs: 1
        track: 11
        tracks: 4
topic: n_elem(1) − freq(0.00103842159917)
        paganism, trees and car parks: 1
type: n_elem(1) − freq(0.00103842159917)
        title cut: 1
```

Appendix B

Example: Clustered data

This appendix shows the internal data structure used for the clustering of sentences according to the labeled keywords. The relevant variables to mention here are:

- *weight*: the mean frequency of the cluster labels,
- *n_tag_covered*: the number of labels in the set of sentences of this cluster,
- *tags*: the labels that define this cluster,
- and the set of sentences that belong to this cluster.

The cluster shown in this appendix are ordered in descending weight.

```
n_elem: 3
weight: 0.122533748702
n_tag_covered: 2
files: ['./svl/01FelaKutiShearwaterEluviumLauraGibsonMore.svl', './svl/RadioOrphansPodcast248.svl
     ', './svl/ReleaseYourselfPodcast102.svl']
tags: {'sentence.before': 1, 'group.name': 1}
1: $sentence.before(pos) And now to $group.name(Eluvium).
2: $sentence.before(pos) Here it is, they're the $group.name(Repeat Offenders).
3: $sentence.before(pos) We put the spotlight this week on $group.name(Chocolate Puma), while
     $transcription(_missing_) with $group.name(Danny Freakazoid).
```

```
n_elem: 4
weight: 0.122533748702
n_tag_covered: 2
files: ['./svl/01ChopinTheWhiteStripesJoannaNewsomMore.svl', './svl/
     TRIntroFreeIntroducingTracks08Mar10.svl', './svl/TRIntroFreeIntroducingTracks22Feb10.svl',
     './svl/XYRocks_084.svl']
tags: {'sentence.after': 1, 'group.name': 1}
4: $sentence.after(pos) They are amazing, $group.name(Sharon Jones and the Dap-Kings).
5: $sentence.after(pos) Really special, $group.name(Mackerel Jason) by $group.name(Hold your
     horse is).
6: $sentence.after(pos)  and, they say, they got bored with the whole $transcription(_missing_)
     with every gig of those not daring to dance, so, they say, in a pop withdrawl panic we formed
      $group.name(Town bike) with one simple rule to bring everything together: have fun.
7: $sentence.after(pos) $group.name(Gunpowder Sunset).
```

```
n_elem: 1
weight: 0.114745586708
n_tag_covered: 3
files: ['./svl/ReleaseYourselfPodcast102.svl']
```

```
tags: {'sentence.begin': 1, 'song.name': 1, 'group.name': 1}
8: $sentence.begin(pos) We are gonne kick things off with a $transcription(banger from) $group.
    name(Boza) taking a deep with $song.name(Tales from a Dj).
-------
n_elem: 13
weight: 0.114745586708
n_tag_covered: 3
files: ['./svl/RadioOrphansPodcast241.svl', './svl/RadioOrphansPodcast248.svl', './svl/
    TRIntroFreeIntroducingTracks22Feb10.svl', './svl/XYRocks_082.svl', './svl/XYRocks_082.svl',
    './svl/XYRocks_083.svl', './svl/XYRocks_083.svl', './svl/XYRocks_083.svl', './svl/XYRocks_083
    .svl', './svl/XYRocks_084.svl', './svl/XYRocks_084.svl', './svl/XYRocks_084.svl', './svl/
    XYRocks_084.svl']
tags: {'sentence.after': 1, 'song.name': 1, 'group.name': 1}
9: $sentence.after(pos) And after that, of course, your humble $transcription(host), $group.name(
    the Radio Orphans) with $song.name(Void versus Bob).
10: $sentence.after(pos) So here it was, it was $song.name(Rebel Hero) from $group.name(New
    Nobility).
11: $sentence.after(pos) $song.name(six pound tracey) $transcription(_missing_) four piece $group
    .name(the ball deep)
12: $sentence.after(pos) And that was $group.name(Better Than Toast) with $song.name(Cringe Or
    Smile).
13: $sentence.after(pos) That was $group.name(Maritime) with $song.name(Parade Of Punk Rock T-
    Shirts).
14: $sentence.after(pos) And that was $group.name(Plane Without A Pilot) with $song.name(
    Everytime).
15: $sentence.after(pos) And that was $group.name(Building Rome) with $song.name(Dr. Doctor).
16: $sentence.after(pos) And that was $group.name(I Hate My Ex) with $song.name(I'm Not Bad, I'm
    Just Drawn That Way).
17: $sentence.after(pos) And that was $group.name(Envy On The Coast) with $song.name(Sugar Skulls
    ).
18: $sentence.after(pos) And that was $group.name(Fall out boy) with $song.name(I don't care).
19: $sentence.after(pos) And that was $group.name(Never Shout Never) with $song.name(Liar Liar).
20: $sentence.after(pos) And that was $group.name(We the kings) with $song.name(This is out town)
    .
21: $sentence.after(pos) $group.name(My Lady Four) $song.name(Sometimes The Bear Eats You).
-------
n_elem: 12
weight: 0.114745586708
n_tag_covered: 3
files: ['./svl/01ChopinTheWhiteStripesJoannaNewsomMore.svl', './svl/662_Coverville.svl', './svl
    /662_Coverville.svl', './svl/RadioOrphansPodcast245.svl', './svl/ReleaseYourselfPodcast102.
    svl', './svl/XYRocks_082.svl', './svl/XYRocks_082.svl', './svl/XYRocks_082.svl', './svl/
    XYRocks_083.svl', './svl/XYRocks_084.svl', './svl/XYRocks_084.svl', './svl/XYRocks_084.svl']
tags: {'sentence.before': 1, 'song.name': 1, 'group.name': 1}
22: This one from $group.name(Pantha du Prince) is called $song.name(The Splendour). $sentence.
    before(pos)
23: $sentence.before(pos) Here's $group.name(Shiny toy guns) $song.name(Burning for you).
24: $sentence.before(pos) $song.name(I am trying to break your heart), here is $group.name(JC
    Brooks and the Uptown Sound).
25: $sentence.before(pos)$song.name(Like a Rocket), that's right, from $group.name(_missing_)
26: $sentence.before(pos)  It's $group.name(Clubshot) $song.name(No Matter).
27: $sentence.before(pos) This is $group.name(Karate High School) with $song.name(Sometimes When
    You Lose, You Win).
28: $sentence.before(pos) This is $group.name(Parachute Musical), $song.name(One More Song).
29: $sentence.before(pos) This is $group.name(Never Shout Never) with $song.name(Can't Stand It).
30: $sentence.before(pos) $group.name(Call the cops) $song.name(Room 410).
31: $sentence.before(pos) $group.name(Danger is my middle name), $song.name(Revenge on the radio)
    .
32: $sentence.before(pos) This is $group.name(We the kings) with $song.name(This is out town).
33: $sentence.before(pos) And here they are $group.name(Gunpowder Sunset) with $song.name(Singles
     soundtrack).
```

Appendix C

Recorded sentences

C.1 Sentences placed before the song

```
$sentence.before(pos) And now going to $group.country(Scotland)...
$sentence.before(pos) This is $group.name(Parachute Musical), $song.name(One More Song).
$sentence.before(pos) Here's $group.name(Shiny toy guns) with $song.name(Burning for you).
$sentence.before(pos) $song.name(I am trying to break your heart), here is $group.name(JC Brooks
    and the Uptown Sound).
$sentence.before(pos) The $group.reference(band) is the $group.name(Archie Bronson Outfit), they
    are $group.city(London) based $group.reference(band) with a $group.description(blown out
    psychedelic sound.
$sentence.before(pos) Our $song.reference(final track) is $song.name(I say hello), by the $group.
    city(Toronto) $member.role(bass producer and composer) $member.name(Julian Bach Low) taken
    from $album.reference(newly release 9 track EP) $album.name(Paradigm) which can be found via
    $group.site(myspace.com/julianbachlowlp).
$sentence.before(pos) $group.name(Danger is my middle name), $song.name(Revenge on the radio).
$sentence.before(pos) Here we go, $group.name(Coolhunter) and their $song.reference(song) $song.
    name(Deborah).
$sentence.before(pos) $group.name(Freeky Cleen) and their $song.reference(song) $song.name(Many
    Is The Time).
$sentence.begin(pos) we're gonna start off with a $group.ensemble(band) from $group.city(Illinois
    ).
$sentence.begin(pos)  the $group.ensemble(band) $group.name(Lousards from a far) and their $song.
    reference(song) $song.name(special boy).
$sentence.before(pos) $song.name(Like a Rocket), that's right, from $group.name(Guitarrist
    Slingers).
$sentence.sequence(prev) $sentence.before(pos) $group.genre(Rockabilly) right out of the $group.
    country(UK).
$sentence.before(pos) Here we are, it's $song.name(All time low).
$sentence.begin(pos) My name is $speaker.name(Marcus), and this is $group.name(Nice Peter).
$sentence.before(pos) This is $group.name(We The Kings), $song.name(All Again For You).
$sentence.before(pos) $group.name(Nice Peter).
$sentence.before(pos) we pop over the $group.state(california) at $group.city(long beach),
    actually.
$sentence.sequence(prev) $sentence.after(pos)  $group.city(long beach) $group.state(california),
    with the $group.ensemble(band) $group.name(the new fidelity).
$sentence.sequence(prev) $sentence.after(pos)  some tight good $group.genre(solid pop music)
    therefore, you know?
$sentence.sequence(prev) $sentence.after(pos)  $group.genre(pop rock sensibilities) as they say.
```

C.2 Sentences placed after the song

```
$sentence.after(pos) That was $song.name(Hey Kid) by $group.name(Make Spot).
$sentence.after(pos)  They $group.action(played) the $concert.place(Captain's Rest) in $concert.
    city(Glasgow) $concert.date(late yesterday), $concert.date(monday 22 of february) and
    $concert.city(Stirling uni) on $concert.date(tuesday).
```

```
$sentence.after(pos) Just see $group.site(myspace.com/makespotspan) for more info.
$sentence.after(pos) That's $song.name(All about our lives), by $group.city(Santa Barbara) $group.ensemble(duo) $group.name(watercolor paintings), aka $member.name(Rebecca and Joshua Redman).
$sentence.after(pos) oh, such fun, is $song.name(Right of your live) by $group.name(Town bike) who are a $group.ensemble(band) from $group.city(Liverpool).
$sentence.after(pos) and, they say, they got bored with those not daring to dance, so, they say, in a pop withdrawl panic we formed $group.name(Town bike) with one simple rule to bring everything together: have fun.
$sentence.after(pos) now we are having fun in $concert.city(Manchester) $concert.date(this coming friday february the 25th) at an undisclosed location which we'll soon reveal any day now at $group.site(myspace.com/townbike).
$sentence.after(pos) That was $group.name(Maritime) with $song.name(Parade Of Punk Rock T-Shirts).
$sentence.after(pos) That's $song.name(Lonely Lonely Lonely) by friend's of this show $group.name(The candle thieves), aka $member.name(Scott McEwan) and $member.name(the glockenshields).
$sentence.after(pos) $song.name(Millipede Stomps) is the $song.reference(debut single) $group.name(The Momeraths).
$sentence.after(pos) $song.name(Worlds Away) $group.name(From First To Last), the $group.reference(band) is kicks ass.
$sentence.after(pos) Huh, you've just heard the $song.reference(song) $song.name(Wendy naked).
$sentence.after(pos) So here it was, it was $song.name(Rebel Hero) from $group.name(New Nobility).
$sentence.after(pos) $song.name(Address the silence) from the $group.reference(band) $group.name(Von Bizmark).
$sentence.sequence(prev) $sentence.after(pos) Cool stuff there, $group.genre(alternative rock) out of $group.city(Dublin).
$sentence.after(pos) hey, we are back, and you've just heard $song.name(Void versus Bob)
```

C.3 Sentences placed before and after the song

```
$sentence.before(pos) That $album.reference(new record) is called $album.name(Ali and Toumani) and this $song.reference(song) is called $song.name(Samba Geladio).
$sentence.sequence(prev) $sentence.after(pos) $member.name(Ali Farka Tour ) and $member.name(Toumani Diabat ) from their $album.reference(record) called $album.name(Ali and Toumani) this $song.reference(song) was called $song.name(Samba Geladio).
$sentence.before(pos) The $album.description(new) $album.reference(album) by $group.name(Shearwater) is called $album.name(Golden Archipelago).
$sentence.sequence(prev) $sentence.after(pos) $group.name(Shearwater) from their $album.reference(album) $album.name(the Golden Archipelago).
$sentence.before(pos) And now to $group.name(Eluvium).
$sentence.before(pos) $group.name(Eluvium) is the $group.description(wind swept sound) of $member.role(musician) $member.name(Mathew Cooper).
$sentence.sequence(prev) $sentence.before(pos) This $song.reference(song) is called $song.name(Leaves eclipse the light).
$sentence.sequence(prev) $sentence.after(pos) The music of $group.name(Eluvium) from the $album.reference(album) $album.name(Similes) and a $song.reference(song) called $song.name(Leaves eclipse the light).
$sentence.before(pos) They've made an $album.description(magical) $album.reference(record) called $album.name(Bridge Carols) and this $song.reference(song) is called $song.name(Glocken).
$sentence.sequence(prev) $sentence.after(pos) $member.name(Ethan Rose) and $member.name(Laura Gibson) from a $album.description(gentle and wonderful) $album.reference(record) called $album.name(Bridge Carols).
$sentence.sequence(prev) $sentence.after(pos) That $song.reference(song) was called $song.name(Glocken).
$sentence.before(pos) This one is called $song.name(Green Withins Brook).
$sentence.sequence(prev) $sentence.after(pos) The sounds of $group.name(Richard Skelton) from a $album.reference(record) called $album.name(Landings), the entire $album.reference(record) is $album.description(just a jam).
$sentence.before(pos) This next $group.reference(band) is from $group.country(the UK), $group.name(The new rock chemists).
$sentence.sequence(prev) $sentence.after(pos) See and check them out at $group.site(myspace dot com forward slash the new rock chemists).
$sentence.before(pos) This is $group.name(We the kings) with $song.name(This is out town).
$sentence.sequence(prev) $sentence.after(pos) And that was $group.name(We the kings) with $song.name(This is out town).
```

```
$sentence.before(pos) And here they are, $group.name(Gunpowder Sunset) with $song.name(Singles
    soundtrack).
$sentence.sequence(prev) $sentence.after(pos) $group.name(Gunpowder Sunset).
$sentence.begin(pos) This is the opening track, it's called $song.name(You must be out of your
    mind).
$sentence.sequence(prev) $sentence.after(pos) $group.name(The Magnetic Fields) from their $album.
    reference(album) called $album.name(Realism).
$sentence.before(pos) This is the title $song.reference(track) form $group.name(Sharon Jones and
    the Dap-Kings), the $song.reference(song) is called $song.name(I learned the hard way).
$sentence.sequence(prev) $sentence.after(pos) They are amazing, $group.name(Sharon Jones and the
    Dap-Kings).
$sentence.sequence(prev) $sentence.after(pos)  That's from their soon to be released $album.
    reference(album) called $album.name(I learned the hard way).
$sentence.sequence(prev) $sentence.after(pos)  The $album.reference(record) comes out on $album.
    release(April).
$sentence.sequence(prev) $sentence.after(pos)  We'll have a $concert.reference(live concert) from
     $group.name(Sharon Jones and the Dap-Kings), a remarkable live band, on $concert.date(March
    17th) live.
$sentence.before(pos) This is $group.name(Building Rome) with $song.name(Dr. Doctor), this is
    from their $album.reference(album) $album.name(Nightmare).
$sentence.sequence(prev) $sentence.after(pos) And that was $group.name(Building Rome) with $song.
    name(Dr. Doctor).
$sentence.before(pos) I gotta get this show going, with $group.name(Death To Juliet).
$sentence.sequence(prev) $sentence.before(pos)  This $group.reference(band) right here is quickly
     become my favourite $group.reference(band), every $song.reference(song) that I listen to
    from them just totally rocks.
$sentence.sequence(prev) $sentence.before(pos)  Maybe check on iTunes daily for their new $album.
    reference(EP) because it's gonna blow your mind.
$sentence.sequence(prev) $sentence.before(pos) Here they are, $group.name(Death To Juliet).
$sentence.sequence(prev) $sentence.after(pos) $show.name(XY Rocks), $group.name(Death To Juliet).
```

C.4 Hello and goodbye sentences for a show

```
$sentence.end(pos) You can find us at $show.site(nprmusic.com).
$sentence.begin(pos)  You connected to $show.name(All Songs Considered).
$sentence.begin(pos) $speaker.hello(Hello everybody) from $show.station(NPR Music).
$sentence.after(pos) This concludes episode $show.episode(241) of the $show.name(Radio Orphans
    Podcast).
$sentence.sequence(prev) $sentence.end(pos) $speaker.goodbye(We thank you for listening).
$sentence.end(pos)  $speaker.goodbye(enjoy the rest of your night, have fun, stay save and I'll
    see you next weekend).
$sentence.end(pos) $speaker.goodbye(Enjoy the rest of your night. Have fun, stay safe, and I'll
    see you tomorrow).
$sentence.end(pos) I've been $speaker.name(Tom Robinson), $speaker.goodbye(thank you for
    listening) and do it join me again on $speaker.day(friday) from seven o'clock.
$sentence.begin(pos) $speaker.hello(Hello everybody) from $show.station(NPR Music).
$sentence.begin(pos) $speaker.hello(Welcome to the show).
$sentence.begin(pos) $speaker.hello(Thanks to all you friends out there).
$sentence.begin(pos) $speaker.hello(Greetings from) $show.name(Coverville).
```

Appendix D

Labels of the sentences

D.1 Labels for sentences placed before the song

Label	Occurrences	Label	Occurrences
album.name	1	group.state	2
album.reference	1	member.name	1
group.city	5	member.role	1
group.country	2	sentence.after	4
group.description	1	sentence.before	14
group.ensemble	3	sentence.begin	3
group.genre	3	sentence.sequence	4
group.name	13	song.name	11
group.reference	2	song.reference	4
group.site	1	speaker.name	1

Table D.1: Labels and occurrences for sentences placed before the song

D.2 Labels for sentences placed after the song

Label	Occurrences	Label	Occurrences
concert.city	3	group.reference	2
concert.date	4	group.site	2
concert.place	1	member.name	3
group.action	1	sentence.after	16
group.city	3	sentence.sequence	1
group.ensemble	2	song.name	11
group.genre	1	song.reference	2
group.name	10		

Table D.2: Labels and occurrences for sentences placed after the song

D.3 Labels for sentences placed before and after the song

Label	Occurrences	Label	Occurrences
album.description	4	group.site	1
album.name	11	member.name	5
album.reference	14	member.role	1
album.release	1	sentence.after	16
concert.date	1	sentence.before	16
concert.reference	1	sentence.begin	1
group.country	1	sentence.sequence	20
group.description	1	show.name	1
group.name	20	song.name	14
group.reference	3	song.reference	9

Table D.3: Labels and occurrences for sentences placed before and after the song

D.4 Labels for hello and goodbye sentences for a show

Label	Occurrences	Label	Occurrences
sentence.after	1	show.site	1
sentence.begin	6	show.station	2
sentence.end	5	speaker.day	1
sentence.sequence	1	speaker.goodbye	4
show.episode	1	speaker.hello	5
show.name	3	speaker.name	1

Table D.4: Labels and occurrences for hello and goodbye sentences

Appendix E

Objective test results

This appendix presents the sentence specific results for the objective test introduced in section 4.2.6.

E.1 Using an anchor sentence close to the target

Tables E.1 to E.6 show the ratios that compare the values used to transform a close sentence to the target with the mean value of each emotion class.. Different tables are shown for each feature.

	Syllable duration feature											
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Mean**
1	1.60	2.04	1.28	2.45	1.14	2.13	1.39	1.67	0.38	1.42	1.15	1.51
2	0.35	0.77	2.58	1.31	0.81	0.96	1.12	3.19	1.73	3.61	2.14	1.69
3	0.52	0.36	0.69	1.96	0.54	0.45	1.57	2.01	2.15	2.57	1.21	1.27
4	0.67	1.24	0.38	0.82	0.69	0.61	0.50	0.82	1.56	2.17	1.07	0.96
5	0.82	0.94	2.66	0.77	0.77	2.02	0.11	0.99	0.90	3.06	1.85	1.35
6	0.71	0.61	1.92	1.07	0.42	1.37	0.61	1.12	1.33	2.93	1.65	1.25
7	0.84	0.88	2.23	0.77	0.54	1.72	0.23	0.85	1.02	2.92	1.73	1.25
Mean	0.79	0.98	1.68	1.31	0.70	1.32	0.79	1.52	1.30	2.67	1.54	1.33

Table E.1: Objective evaluation with close anchor sentence: syllable duration feature for sentence 1

	Syllable duration feature												
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Syll 12**	**Mean**
1	1.12	0.94	1.66	1.19	1.30	0.88	1.20	0.68	1.29	1.02	1.08	0.78	1.09
2	2.76	0.20	0.45	1.92	0.48	1.51	1.00	0.51	1.36	1.03	0.91	1.03	1.10
3	1.36	0.73	0.38	1.20	1.23	0.69	0.64	0.67	0.26	0.78	1.04	1.27	0.85
4	1.15	1.00	0.66	0.56	1.44	0.77	0.61	0.55	0.22	0.68	0.90	1.13	0.81
5	2.40	0.26	0.58	1.89	0.58	1.44	1.26	0.71	1.77	1.11	0.98	0.87	1.15
6	1.33	0.58	0.27	1.14	1.39	0.46	0.43	0.34	0.02	0.66	1.00	1.40	0.75
7	2.38	0.30	0.23	1.33	0.96	0.59	0.28	0.66	0.33	0.64	0.82	1.28	0.82
Mean	1.78	0.57	0.60	1.32	1.05	0.90	0.77	0.59	0.75	0.84	0.96	1.11	0.94

Table E.2: Objective evaluation with close anchor sentence: syllable duration feature for sentence 2

	Energy feature											
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Mean**
1	0.90	1.33	1.68	1.15	1.12	1.84	0.39	0.81	1.39	2.31	1.62	1.32
2	1.64	2.09	1.76	2.05	1.58	1.13	1.78	1.76	1.20	1.59	0.85	1.58
3	1.04	1.60	1.93	2.50	3.04	2.25	1.45	1.60	2.35	2.75	2.91	2.13
4	0.24	0.59	0.76	0.69	0.56	0.81	0.33	0.57	1.00	0.65	0.53	0.61
5	1.81	0.94	2.05	2.14	1.12	0.86	1.55	0.92	1.04	0.72	0.51	1.24
6	2.34	2.99	2.14	2.48	3.67	2.88	2.67	3.94	2.62	3.12	3.50	2.94
7	1.96	1.09	1.35	1.04	0.91	0.95	0.54	0.70	0.81	0.83	0.45	0.97
Mean	1.42	1.52	1.67	1.72	1.72	1.53	1.24	1.47	1.49	1.71	1.48	1.54

Table E.3: Objective evaluation with close anchor sentence: Energy feature for sentence 1

	Energy feature												
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Syll 12**	**Mean**
1	0.60	1.79	1.37	1.14	0.89	1.15	1.03	1.45	1.09	1.25	1.14	0.61	1.12
2	0.46	0.77	0.26	1.16	0.83	0.86	1.17	0.87	0.55	0.83	0.61	0.71	0.76
3	0.82	0.34	0.34	0.57	0.92	0.33	0.34	0.16	0.31	0.26	0.56	0.26	0.43
4	1.00	0.42	0.24	0.65	0.79	0.30	1.01	0.25	0.28	0.22	0.53	0.29	0.50
5	0.37	0.68	0.31	1.09	0.42	0.90	1.15	0.66	0.60	0.85	0.77	0.48	0.69
6	0.85	0.33	0.34	0.76	1.04	0.36	1.73	0.34	0.32	0.25	0.63	0.26	0.60
7	0.21	0.14	0.11	0.23	0.20	0.09	0.13	0.04	0.15	0.07	0.14	0.03	0.13
Mean	0.61	0.64	0.42	0.80	0.73	0.57	0.94	0.54	0.47	0.53	0.63	0.38	0.60

Table E.4: Objective evaluation with close anchor sentence: Energy feature for sentence 2

	Pitch feature											
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Mean**
1	0.99	1.05	1.13	1.22	1.30	1.21	0.95	0.73	0.89	0.95	0.85	1.03
2	1.50	1.66	1.29	1.11	1.60	1.55	1.65	1.79	1.62	1.49	1.34	1.51
3	1.65	1.59	1.64	1.24	1.43	1.46	1.73	2.51	1.81	1.59	1.34	1.63
4	1.23	1.73	1.48	1.45	1.54	1.42	1.10	1.98	1.08	1.21	1.13	1.39
5	1.48	1.46	1.20	1.22	1.39	1.59	1.70	1.67	1.60	1.44	1.18	1.45
6	1.60	1.89	1.53	1.27	1.64	1.94	1.82	3.76	1.95	1.75	1.30	1.86
7	1.56	1.44	1.52	1.09	1.65	1.66	1.76	1.89	1.75	1.52	1.27	1.55
Mean	1.43	1.55	1.40	1.23	1.51	1.55	1.53	2.04	1.53	1.42	1.20	1.49

Table E.5: Objective evaluation with close anchor sentence: Pitch feature for sentence 1

E.2 Using the same neutral-calm anchor sentence for all transformations

Tables E.7 to E.12 show the ratios that compare the values used to transform a neutral and calm sentence to the target with the mean value of each emotion class.. Different tables are shown for each feature.

	Pitch feature												
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Syll 12**	**Mean**
1	1.20	1.30	1.20	1.13	0.96	1.00	1.19	1.51	1.29	1.46	1.42	1.21	1.24
2	1.13	1.64	1.05	1.44	0.83	1.12	0.88	1.65	0.91	1.20	0.70	0.58	1.10
3	0.59	0.40	0.68	1.14	1.45	1.20	1.14	0.92	0.00	1.32	1.52	1.57	0.99
4	1.09	0.63	1.21	1.25	1.56	1.44	2.20	1.00	0.00	1.35	1.56	1.70	1.25
5	1.08	1.58	1.06	1.37	0.71	1.07	0.83	1.68	1.00	1.29	1.00	0.85	1.13
6	0.77	0.36	1.06	1.26	1.41	1.08	1.72	0.91	0.00	1.37	1.46	1.49	1.07
7	1.05	1.18	1.01	1.25	1.20	1.37	1.19	0.97	0.00	1.17	1.47	1.53	1.12
Mean	0.99	1.01	1.04	1.26	1.16	1.18	1.31	1.24	0.46	1.31	1.30	1.28	1.13

Table E.6: Objective evaluation with close anchor sentence: Pitch feature for sentence 2

	Syllable duration feature											
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Mean**
1	1.60	2.04	1.28	2.45	1.14	2.13	1.39	1.67	0.38	1.42	1.15	1.51
2	0.35	0.77	2.58	1.31	0.81	0.96	1.12	3.19	1.73	3.61	2.14	1.69
3	0.52	0.36	0.69	1.96	0.54	0.45	1.57	2.01	2.15	2.57	1.21	1.27
4	0.67	1.24	0.38	0.82	0.69	0.61	0.50	0.82	1.56	2.17	1.07	0.96
5	0.82	0.94	2.66	0.77	0.77	2.02	0.11	0.99	0.90	3.06	1.85	1.35
6	0.71	0.61	1.92	1.07	0.42	1.37	0.61	1.12	1.33	2.93	1.65	1.25
7	0.84	0.88	2.23	0.77	0.54	1.72	0.23	0.85	1.02	2.92	1.73	1.25
Mean	0.79	0.98	1.68	1.31	0.70	1.32	0.79	1.52	1.30	2.67	1.54	1.33

Table E.7: Objective evaluation with neutral anchor sentence: syllable duration feature for sentence 1

	Syllable duration feature												
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Syll 12**	**Mean**
1	1.12	0.94	1.66	1.19	1.30	0.88	1.20	0.68	1.29	1.02	1.08	0.78	1.09
2	2.76	0.20	0.45	1.92	0.48	1.51	1.00	0.51	1.36	1.03	0.91	1.03	1.10
3	1.36	0.73	0.38	1.20	1.23	0.69	0.64	0.67	0.26	0.78	1.04	1.27	0.85
4	1.15	1.00	0.66	0.56	1.44	0.77	0.61	0.55	0.22	0.68	0.90	1.13	0.81
5	2.40	0.26	0.58	1.89	0.58	1.44	1.26	0.71	1.77	1.11	0.98	0.87	1.15
6	1.33	0.58	0.27	1.14	1.39	0.46	0.43	0.34	0.02	0.66	1.00	1.40	0.75
7	2.38	0.30	0.23	1.33	0.96	0.59	0.28	0.66	0.33	0.64	0.82	1.28	0.82
Mean	1.78	0.57	0.60	1.32	1.05	0.90	0.77	0.59	0.75	0.84	0.96	1.11	0.94

Table E.8: Objective evaluation with neutral anchor sentence: syllable duration feature for sentence 2

	Energy feature											
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Mean**
1	0.21	0.11	0.35	0.18	0.09	0.16	0.05	0.15	0.23	0.10	0.04	0.15
2	0.20	0.35	0.27	0.17	0.14	0.24	0.06	0.36	0.28	0.29	0.09	0.22
3	0.12	0.35	0.64	0.66	0.55	0.68	0.23	0.69	1.14	0.69	0.60	0.58
4	0.24	0.59	0.76	0.69	0.56	0.81	0.33	0.57	1.00	0.65	0.53	0.61
5	0.21	0.15	0.31	0.17	0.10	0.18	0.05	0.19	0.25	0.13	0.05	0.16
6	0.26	0.59	0.77	0.67	0.56	0.83	0.39	0.84	1.32	0.67	0.56	0.68
7	0.23	0.23	0.43	0.27	0.16	0.29	0.08	0.30	0.40	0.21	0.09	0.25
Mean	0.21	0.34	0.50	0.40	0.31	0.46	0.17	0.44	0.66	0.39	0.28	0.38

Table E.9: Objective evaluation with neutral anchor sentence: Energy feature for sentence 1

	Energy feature												
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Syll 12**	**Mean**
1	0.32	0.12	0.26	0.33	0.18	0.07	1.34	0.06	0.20	0.05	0.10	0.01	0.25
2	0.78	0.21	0.24	0.33	0.52	0.07	1.98	0.12	0.27	0.06	0.14	0.03	0.40
3	2.14	0.40	0.69	1.00	1.75	0.33	2.70	0.32	0.40	0.29	0.85	0.24	0.93
4	1.00	0.42	0.24	0.65	0.79	0.30	1.01	0.25	0.28	0.22	0.53	0.29	0.50
5	0.45	0.15	0.25	0.33	0.26	0.07	1.61	0.08	0.23	0.05	0.12	0.01	0.30
6	0.85	0.33	0.34	0.76	1.04	0.36	1.73	0.34	0.32	0.25	0.63	0.26	0.60
7	0.55	0.16	0.22	0.41	0.37	0.10	1.28	0.09	0.20	0.08	0.22	0.03	0.31
Mean	0.87	0.26	0.32	0.55	0.70	0.18	1.66	0.18	0.27	0.14	0.37	0.12	0.47

Table E.10: Objective evaluation with neutral anchor sentence: Energy feature for sentence 2

	Pitch feature											
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Mean**
1	0.90	1.29	1.05	1.31	1.19	1.29	0.88	0.83	0.81	0.99	0.90	1.04
2	1.24	2.07	1.27	1.09	1.52	1.60	1.47	1.66	1.62	1.44	1.48	1.50
3	1.33	2.09	1.36	1.48	1.20	1.53	1.71	2.80	1.86	1.52	1.39	1.66
4	1.23	1.73	1.48	1.45	1.54	1.42	1.10	1.98	1.08	1.21	1.13	1.39
5	1.19	1.79	1.15	1.21	1.28	1.58	1.50	1.58	1.64	1.40	1.32	1.42
6	1.45	2.16	1.43	1.45	1.48	1.86	1.99	3.47	2.15	1.69	1.42	1.87
7	1.27	1.91	1.26	1.32	1.40	1.71	1.67	2.05	1.77	1.50	1.34	1.56
Mean	1.23	1.86	1.29	1.33	1.37	1.57	1.47	2.05	1.56	1.39	1.28	1.49

Table E.11: Objective evaluation with neutral anchor sentence: Pitch feature for sentence 1

	Pitch feature												
Sample ID	**Syll 1**	**Syll 2**	**Syll 3**	**Syll 4**	**Syll 5**	**Syll 6**	**Syll 7**	**Syll 8**	**Syll 9**	**Syll 10**	**Syll 11**	**Syll 12**	**Mean**
1	1.54	1.01	1.47	1.00	0.83	0.88	1.93	0.74	0.00	0.87	1.07	1.09	1.04
2	1.78	1.14	1.51	1.11	0.80	0.68	1.28	0.67	0.00	0.71	0.68	0.67	0.92
3	1.01	0.35	0.96	1.21	1.39	0.94	1.73	0.86	0.00	1.32	1.38	1.41	1.05
4	1.09	0.63	1.21	1.25	1.56	1.44	2.20	1.00	0.00	1.35	1.56	1.70	1.25
5	1.88	1.20	1.57	1.08	0.71	0.72	1.46	0.73	0.00	0.77	0.93	0.95	1.00
6	0.77	0.36	1.06	1.26	1.41	1.08	1.72	0.91	0.00	1.37	1.46	1.49	1.07
7	1.41	0.78	1.37	1.23	1.12	0.97	1.70	0.86	0.00	1.14	1.30	1.35	1.10
Mean	1.35	0.78	1.31	1.16	1.12	0.96	1.72	0.82	0.00	1.08	1.20	1.24	1.06

Table E.12: Objective evaluation with neutral anchor sentence: Pitch feature for sentence 2