A Study on Correlates of Acoustic Features to Emotional Singing Voice Synthesis

Nguyen Thi Hao

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
September, 2018

Master's Thesis

A Study on Correlates of Acoustic Features to Emotional Singing Voice Synthesis

1610058 Nguyen Thi Hao

Supervisor : Masato Akagi
Main Examiner: Masato Akagi
Examiners : Masashi Unoki
Jianwu Dang
Atsuo Yoshitaka

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information science]

August, 2018

ABTRACTS

Singing voice analysis and synthesis become an interesting topic recently. In this context, expressiveness plays an important role in obtaining a high quality of singing voices. Expression control conducts a set of acoustic features that are related to emotions, styles or singer individualities. Listeners, depending on their moods and situations, would like to hear a song with different emotional expressions. It therefore needs to have a computer-based application in singing voice performances, in order to satisfy such purpose.

M. Alonso tries to generate emotional singing voices using a rule-based approach. The advantage in this method is that the obtained rules are relatively simple and deterministic. However, it takes a long time for the analysis and synthesis phases to get the rules, and the system is yet difficult to control. Another singing voice synthesizer is proposed by M. Umbert, which applies a unit-selection methodology. It constructs implicit rules based on a number of units (each unit is a short time segmentation) of a singing voice. Hence, it requires a huge unit database to process, in order to obtain a high quality voice. Besides, HMM-based approaches (Hidden Markov Model) are also taken into account for synthesizing expressive singing voice, by using statistical model to model important features from database. This method, nevertheless, has a parameters over-fitting problem. A novel speech-to-singing system, proposed by Saitou et al., has applied the performance-driven approaches, which can produce a singing voice from simple resources: (i) a speaking voice reading a song's lyric, and (ii) its musical score. It succeeded in synthesizing a neutral singing voice. Nonetheless, the expressiveness was not taken into consideration.

Our work aims to investigate the correlations of acoustic features to emotional singing voices. Two sub-goals are consisted in this study: (i) analyzing a set of acoustic features that are strongly related to emotions, and (ii) conducting experimental examinations to evaluate the importance of each acoustic feature in the emotional singing voice. By achieving the first sub-goal, we determine which features are most significant to emotional expressions in singing voices. Regarding the second sub-goal, we propose a method to modulate the amplitude envelope based on the entire F0 contour, to have a higher naturalness in a singing voice. Our experimental results show that the spectral feature is the most affected acoustic feature to the emotions of a singing voice. However, in order to obtain high naturalness and singing-ness in synthesized voices, it is necessary to manipulate all three features, including F0 contour, amplitude envelope and spectral sequences.

Keywords: expressive singing voice, emotional, acoustic features, subjective test.

ACKNOWLEDGEMENT

Personally, I felt very lucky when I was assigned this project. During the last two years, there have been moments of everything, good and bad, but at all times I have received help and good advice from those who surround me (my professors, my family, my friends). In this sense, I feel fortunate.

First of all, I would like to express my appreciation to my supervisor, Professor Masato AKAGI. During the time doing my master study at Japan Advanced Institute of Science and Technology (JAIST), he has given me strong support and has guided me with his academic knowledge. He always made me believe that I could succeed by encouraging me to foster my ideas and research motivation. Working with Professor AKAGI, I have learned the value of research and, above all, how to become a good researcher.

I would like to express my sincere thanks to Professor Masashi UNOKI for his supports and guidance. With his enthusiastic advice, precise comments and encouragement, I can achieved the good results from this study. In addition, his contributive revisions for my weekly reports and presentations that help me to improve my writing and presenting skill a lot. Without his help, I would had have many difficulties to finish this thesis.

I would like to express my special thanks to Professor Elbarougy REDA, Doctor Rieko KUBO, Doctor Maori KOBAYASHI, for their helpful advice, comments, and contributions to my study. They help point out my mistakes and give me comments to deal with many arisen problems during the time doing my thesis.

I am grateful to Assoc. Prof. Le Minh NGUYEN for his enthusiasm to guide me when I doing my minor research project with him.

I would like to also thank Mr. Van Thuan NGO, Mr. Kim Dung TRAN and Mr. Tuan Vu HO who have helped me a lot in answering many of my research questions and sharing their experience with me.

I greatly appreciate the following organizations: Acoustic and Information Science laboratory, JAIST; JASSO scholarship organization, Ishikawa scholarship organization; the Student Welfare Section, JAIST.

Finally, I want to give the best thank to my parents, my relatives and my friends who always encourage, take care of me and raise me up during the studying and researching period.

This thesis was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026).

Sincerely,
Nguyen Thi Hao

Contents

Abstract **i**

Acknowledgement **ii**

1 Introduction **1**
1.1 Research Background 1
1.2 Problems 2
1.3 Research Aims 3

2 Literature review **5**
2.1 Singing voices synthesis 5
2.2 Emotional speech synthesis 10
2.3 Valence - Activation domain 12

3 Acoustic Feature Analysis **13**
3.1 Corpus 13
3.2 Determining of significant acoustic features 14
3.3 Extraction of acoustic feature 15
3.4 Acoustic feature modification 20
3.4.1 Fundamental Frequency (F0) 20
3.4.2 Spectral Sequence 21
3.4.3 Amplitude Envelope 21
3.5 Experimental Examination 23
3.5.1 Subjective test 25
3.5.2 Results 27
3.5.3 Discussion 30
3.6 General Discussion 31

4 Conclusion **32**
4.1 Summary 32
4.2 Contribution 33
4.3 Remaining works 33

Bibliography **34**

Publication **36**

Appendix **38**

List of Figures

2.1 General framework blocks for expression control [1]. 5
2.2 Overview of HMM-based singing voices synthesis system [2]. 7
2.3 Mean vocal intensity across the 11 emotions for speech (a), and song (b). The figure illustrates the two main effects of Domain and Emotion, with song being louder overall than speech, and with speech emotions appearing to show greater variability in intensity than song [3]. 11
2.4 Position of emotions in V-A domain. 12

3.1 Melody component and extracted F0 fluctuations. 17
3.2 Detected 4 types of F0 fluctuations: overshoot, vibrato, preparation. Fine fluctuation along the entire contour. 17
3.3 Block diagram of the F0 control model for singing voices was proposed by Saitou et al. [4]. 17
3.4 Amplitude Envelope detected from singing voice. 19
3.5 The spectral sequences of different emotions. 20
3.6 Spectral of anger speaking voice. 21
3.7 Lengthen spectral of anger speaking voice. 22
3.8 Modulating amplitude envelope from F0 contour. 23
3.9 Synthesis diagram for emotional singing voice synthesis. 24
3.10 User interface used for activation evaluation. 26
3.11 User interface used for valence evaluation. 27
3.12 Position in V-A domain of stimuli of female singer. Green: happy; Blue: angry; Black: sad; Red: neutral. 28
3.13 Position in V-A domain of stimuli of male singer. Green: happy; Blue: angry; Black: sad; Red: neutral. 28
3.14 Direction of emotional stimulus position in V-A domain. 30

List of Tables

3.1 Parameters value for F0 control model. 20
3.2 Average distance and direction between the emotional stimulus position to neutral position on V-A domain - Female singer. 29
3.3 Average distance and direction between the emotional stimulus position to neutral position on V-A domain - Male singer. 29

Chapter 1

Introduction

In this chapter, we provide an overview of research backgrounds, challenges, our motivations and objectives, and finally is the contribution of this study.

1.1 Research Background

Alongside with "speaking a word", "sing a song" is another effective way of communication to express human emotions and feelings. It is interesting when there has been a song named "A Song is Worth a Thousand Pictures (author Greenlight Promise), and meanwhile, another song named "A Picture is Worth a Thousand Words". It suggests that a singing voice with expressions bring much more impression to human being than a neutral speech. For example, a professional singer uses her own singing laws and timbres in singing a song, to make listener feel inspired. How can they create such a singing voice? In the process of finding an answer for this question, people face with a big problem: how do people listen and vocalize a singing voice? Until now, they are yet to obtain a clear answer about the mechanism of the singing voice perception and generation.

In the context of "speak the language", a multilateral standpoint has been carried out for many years, such as psychology and physiology. Many findings are being obtained and actively applied to speech synthesis. This approach is not only develop a synthesis system by making use of knowledge from psychology and physiology in speech synthesis, but also aim to synthesize the speech of higher quality. People are still wondering: "What are the characteristics of acoustic important for speech perception ?","How to control those acoustic features?", and "How it contributes to the elucidation of these mechanisms". Actually, a part of the speech perception (or generation) mechanism is gradually clarified by realizing spontaneous speech synthesis. It is done by considering: (i) the knowledge of human vocal tract control mechanism, and (ii) the combination of emotional speech synthesis and knowledge about emotional perception.

Regarding "singing songs", various knowledge has been gained by psychology, physiology, and even efforts from singing studies. Then, by using the same framework as speech synthesis, these findings have been applied to singing voice synthesis. This approach should enable efforts leading to the elucidation of singing voice perception and generation mechanism, as well as development of singing voice synthesis research. However, currently, it is almost impossible to clarify what physical quantity is necessary to synthesize singing voice, not to mention the construction of a singing voice

synthesizing system.

So, what is the physical quantity necessary for singing voice synthesis? Even with a singing voice within a word, its acoustic features show a wide variety of properties due to differences in singing techniques, such as opera, and other singing songs. However when we are listening to singing voices with big differences in singing skills, human beings can easily distinguish the difference in skill while they are perceived as singing voices. Also, when listening to a singing voice without familiarity, even if it is difficult to identify the singing law, it can be perceived as a singing voice rather than a speech voice.

In other words, singing voice has contained the acoustic features that common among singing voices which do not depend on the difference in singing skill and singing law, and this characteristic is the one that necessary for synthesizing singing voices in order for singing voices to be perceived as singing voices. Therefore, the singing voice synthesis system focusing on the acoustic features commonly included in this singing voice and the approach to clarify the relationship between each feature and singing voice perception among them is the base of clarifying the singing voice perception and generation mechanism [5].

In the topic of singing voice analysis and synthesis, expressiveness, especially emotion, plays an important role in obtaining high quality of singing voices. Expression control manipulates a set of acoustic features that are related to emotion, style or individuality of a singer. Even though the quality of synthesized singing voice nowadays has been achieved the acceptable result, it is still not contain the expressiveness yet in compare with the real singer. Meanwhile, listeners, depending on their moods and situations, would like to hear a song with different emotional expressions. It is therefore needed to have a computer-based application in singing voice performances for fulfilling such purpose. It will be interesting for the end users because they can listen to the generated singing voice having the speaker's voice timbre also different emotions embedded inside. This system can also be used for computer-based music productions, supporting for those mood-based retrieval tools as music recommendation, automatic play-list generation. Those effective singing voice also can use for health care as a tool for stress management or for the game development industry and entertainment.

1.2 Problems

Even though expression control and especially emotion control in singing voice has become a trending topic recently, it is still having some challenges and problems that we need to solve out. In this section, we discuss about significant challenges in trying to obtain a more human-like naturalness in the synthesis voices. Then, some of the difficulties in modifying acoustic features (especially is F0 contour) to make the singing voice not only natural but also bring the information of emotions. In the other words, it is about keeping the singing-ness along side with emotional in the synthesized voice.

Trying to make the synthesized singing voice that can not be distinguishable from natural human singing voice is one of the fundamental goals of singing synthesis technologies. Although the naturalness of synthesized singing voices has been increasing, perfect human-like naturalness has not yet been achieved. It is required the more dynamic, complex, and expressive changes in the voice pitch, loudness, and timbre in synthesized singing voice. For example, voice quality modifications could be

related to emotions, style, individuality or lyrics. In addition, according to the "Uncanny Valley" hypothesis, the higher human-like naturalness is generated, the more creepiness of the synthesized voice is felt [6]. Although this hypothesis often correlates with robot and computer graphics, it suites also for singing voices. However, regardless of this discussion, the topics of singing voice analysis and synthesis are more and more attractive nowadays.

Regarding the acoustic features, although F0 can be easily control to have emotion in synthesized product like speech, in the domain of singing voice synthesis, it has many constraints due to their special characteristics. The fundamental frequency of the voice is as important in singing voice as in speech, but it is obviously different about their roles. There have been some difference as below:

- A singer is expected to sustain the mean fundamental frequency at a constant value over the time interval of a note, at least in classical music. In the other words, they have to follow the musical notes (the melody) of the song to make it like "singing voice", on the other hand, a speaker is never expected to do so during an utterance in speech. Hence, there is no F0 declination within a note, while F0 declination is quite common in speech.

- Unlike in the speech when the speaker is not required to consciously controlled the rate of F0; in the singing voice, a singer is asked to control the rate of F0 transition from one note to an other according to a specific manner of a performance.

Therefore, it is interesting to study about the F0 control in singing voices, but it is also relevant to the study of F0 control in speech, since it involves the same control mechanisms as are used in speech production [7].

Furthermore, there are only few research investigate about emotional singing voice until now due to the difficulties they have. Therefore, the appropriate results about the importance of the acoustic features in emotion of singing voice are still unclear. It is essential to look over this crucial topic to contribute for the singing voice synthesis industry.

1.3 Research Aims

Our final goal is to create a computer-based application that embeds different emotions into a song by using a simple input as a the speaking voice of the user. In order to do so, the aim of this thesis is to study about contributions of significant acoustic features to emotional singing voices. With this aim, there are two sub-goals needed to be achieved: (i) analyzing the acoustic features that are strongly related to emotion, (ii) conducting the experimental examination to discuss importance of each acoustic feature in the emotional singing voice.

My thesis will be organized as follow:

- Chapter 1 gives an overview about emotional singing voice synthesis, and discusses about some challenges of emotional singing voice synthesis.
- Chapter 2 mentions about the theory of singing voice synthesis and emotional speech synthesis and introduces the basic knowledge about emotional domain Valence-Activation.
- Chapter 3 investigates about the role of different acoustic features in emotional singing voice and presents the discussion about those obtained results.
- Chapter 4 gives some conclusions and our remaining works to get the better results.

Chapter 2

Literature review

In this chapter, we provide a literature review about emotional singing voice synthesis. First, the history, characteristics, applications and potential difficulties of synthesizing singing voices are presented in Section 2.1. We then introduce about emotional speech synthesis systems in Section 2.2. Finally, in Section 2.3, we present the Valence-Activation domain, which is used for the experimental examinations in our research.

2.1 Singing voices synthesis

According to a review about expression control in singing voices synthesis of [1], a general framework for controlling expressions in singing voices can be displayed as in Figure. 2.1.

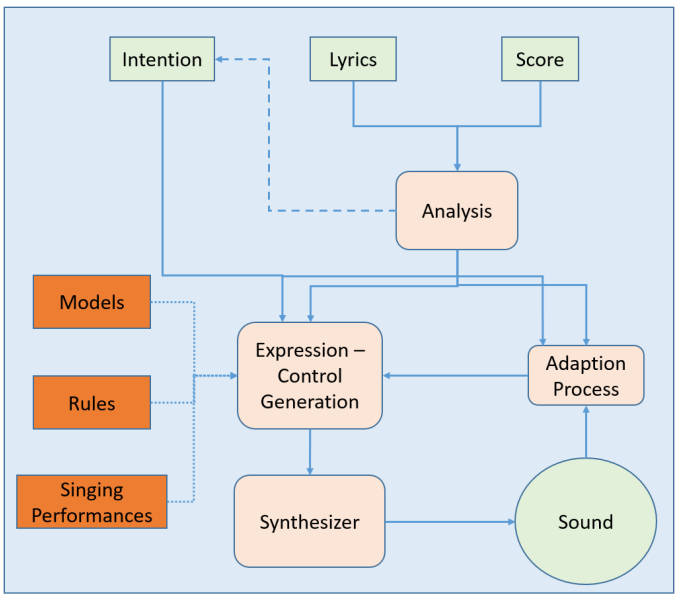


Figure 2.1: General framework blocks for expression control [1].

In this diagram, the score (melody, musical notes), the lyrics or the intension (the singer individ-

uality, style or emotion) will be used as inputs of the system. The intension can be inputed either directly or by deriving from the lyrics and score content (shown by the dash line). The implicit or explicit knowledge of the system (a set of reference singing performances, a set of rules, or statistical models) is presented by the expression control generation block. After that, the output of Expression Control Generation block will be used by the synthesizer to generate the sound.

Based on this diagram, there are some successful approaches in generating the singing voices with expression as below:

- **Rule-based approaches**: Alonso [8] has tried to generate emotional singing voices using rule-based approach. His research is based on KTH, a very famous system [9]. In this method, a set of rules that reflects a singer's cognitive process will be derived. These rules will be examined in both analysis by synthesis phase by synthesizing singing voices performances. By observing the singing voices contour and trying to imitate their characteristics, the corpus-derived, rule-based approaches will generate expression controls.

  While KTH system concentrate on making the singing voices synthesizer that generating the singing voices having good performance and style, they not yet considering about the emotion expression in singing voices, the research about emotional singing voices of M. Alonso [8] after trying to applied a selection of the KTH rules to the synthesizer, they have not done the evaluation test to show their ability yet. No proven about any subjective or objective test were shown.

  The advantage in this approach is that they are somewhat straightforward and completely deterministic. The new synthesis of the same score can be generated different contours due to the random variations can be easily brought in, therefore, we can having many distinct interpretations as well.

  However, if the models are based on only few observations, it can not fully express a given style. Meanwhile, when the system is created more elaborate, it is time-consuming for both the analysis and synthesis phases to get the rules, and the system is yet difficult to control and also become unmanageable due to the complexity of the rules.

- **Unit-selection approaches**: Another singing voices synthesizer by Umbert has applied unit-selection methodology [10]. This method constructs implicit rules based on units of a singing voices and after that concatenating them. For more details, this approaches using database of singing recordings that were segmented into units and based on the selected score, a sequences of phonemes with specific features such as pitch of duration is retrieved from the database. Based on the very important criteria is the definition of the target and concatenation cost functions, the unit selection will be built. The overall cost of the unit sequence will be calculated by the sum of the cost functions' contribution weight.

  One typical case of unit-selection approaches, which is a famous singing synthesizer system, is VOCALOID developed by Yamaha Corporation [11]. This work try to model the singer's performance with heuristic rules. By using VOCALOID software, you can create songs on your computer just by inputting lyrics and a melody. This method based on the database named Voice Banks that are recordings of professional singer, and one more software named

VOCALOID Editor for adjusting the detailed settings to change the singing style however you like.

Same like KTH system, VOCALOID also does not considering about the emotion of the singing voices due to they does not have the emotional singing voices in the Voice Banks. In addition, the control rules for adding emotion to the voice also complicated as well.

As unit-selection methods use a training database of expressive singing that have been labeled, it requires a huge database of units for processing when achieving higher singing voices quality with expressiveness. In addition, the sub-cost weights need to be specified. Nonetheless, the voice quality and naturalness are high because of the implicit rules applied by the singer within the units [1].

- **Statistical approaches**: Statistical approaches are also taken into account for synthesizing expressive singing voices by using a statistical methodology such as Hidden Markov Models (HMMs) [12] to model the important features from database. In the domain of speech synthesis they normally using the jointly model for some important parameter as spectrum, F0 and state duration. The overview of HMM-based singing voices synthesis system is shown in the Figure. 2.2

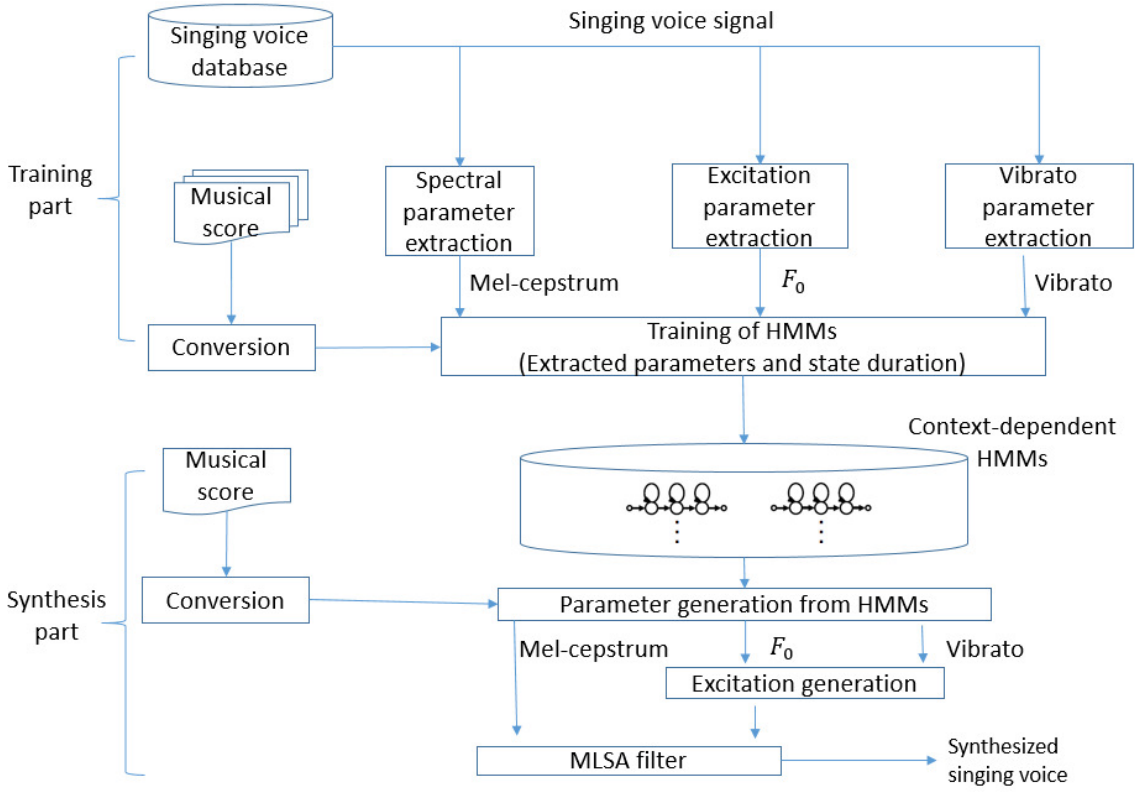


Figure 2.2: Overview of HMM-based singing voices synthesis system [2].

According to this diagram, the system consists of two parts are training part and synthesis part. In the training part, the spectrum (mel-cepstral coefficients), the excitation and the

vibrato are extracted from the singing voices database. In the next step, these information will be used for training of HMMs. The context-dependent models of state durations are also estimated. In the training part, they also use the musical score that is converted to a context-dependent label sequence as one input for the training of HMMs.

Regarding the synthesis part, by concatenating the context-dependent HMMs, an HMM corresponding to the song is constructed. Then, the spectrum, excitation and vibrato parameters are generated by an algorithm to obtain the speech parameters. Finally, by using a Mel Log Spectrum Approximation (MLSA) filter, a singing voice is synthesized directly from the generated spectrum, excitation and vibrato parameters.

Since using the statistical model to model the features, the quality of synthesized voice is lower than using unit-selection method but they only need a small amount of database to train the system. Plus, by changing the HMM parameters like rule-based method, we can easily generate the new voice characteristics. However, this method has to face with the problem of over-fitting of parameters.

- **Performance-driven approaches**: This approaches control the synthesizer by using a real performances of the singer, hence, it can take the advantage of the implicit rules has been applied to interpret a score. These implicit rules will be used to both control the parameter such as F0, intensity of timing and transform the speech audio containing the target lyrics in order to match the pitch and timing of the input score.

  The synthesis engine is the main part in this approach. Based on a reference singing performance of a target score, it extracts the expression parameters controls for pitch, timing, dynamics, and timbre. After that, the unit-selection based synthesizer will use these parameters to modify the selected units. In addition, a combination of sinusoidal modeling (SM) with time domain pitch synchronous overlap add (TD-PSOLA) called SM-PSOLA will be also used [1].

  By directly applying the real knowledge of the singer that implicitly contain in the input, the expression control performances are very high. These approaches are especially convenient for creating parallel database recordings, which are used in voice conversion approaches [13]. However, it is time consuming when they needed to do the phonetic segmentation task manually and correctly because it may cause timing errors.

  A novel speech-to-singing system proposed by Saitou et al. [4], which can produce a singing voices from simple resources: (i) a speaking voice reading a song's lyrics, and (ii) its musical score. It succeeded in synthesizing a neutral singing voices. Nevertheless, the expressiveness has not been taken into consideration.

In order to choosing suitable approach for synthesizing the singing voices, we have several considerations like: the disadvantages of each approach; the available of singing voices database; the reason for synthesizing a song; or the requirements of the study. For more detail the brief information about each approach can be summarized as below:

- Due to the necessary of the implicit rule, applied in singer expression for controlling the synthesizer, it will be suitable to use the Performance-driven approaches when the target performance is available. This approach also can be used for creating the parallel databases, which cover different purposes such as voice conversion [8]. In addition, speech to singing synthesis system is an other example for the applicability of this approach. This application can be used for untrained singers, when it takes their timbre from the speech recording and uses the professional singer's expression for pitch and dynamics.

- Rule-based approaches are suitable to be applied if there are no recordings available. This approaches are convenient to examine the defined rules and to see how these are combined to reveal an emotion. Even without the speech recordings, they still can define the rules with the help of an expert, so that these approaches are not thoroughly dependent to the singing voices databases.

- Statistical modeling approaches have an advantage of being flexible when it is possible to interpolate models and also generate new voice characteristics. Especially, in some complete singing voices synthesis systems, which the input is just the score and they can output both the expression parameters and the voice.

- In the same way as rule-based and statistical modeling approaches, unit selection approaches do not require the goal performance. Despite that, these approaches also take the implicit knowledge of the singer as the performance driven approaches, however, it just extracts the hint rules from unit of audio (short time segments). There is no training step in this approach, so that, in order to increase the expression, it is needed to include more new labeled singing voices recordings to the database.

As we have stated in the first chapter, the aim of this study is to investigate contributions of significant acoustic features to emotional singing voices. There are two sub-goals need to be achieved: (i) analyzing the acoustic features that are strongly related to emotion, (ii) conducting the experimental examination to discuss the importance of each acoustic feature in the emotional singing voices.

Hence, it needed to synthesize the singing voice to carry out the evaluation test in the second step. Based on the reviewed knowledge, we decided to combine the two approaches are rule-based method and performance-driven in order to not only achieve the highest quality but also get used of the advantages of these methods.

2.2 Emotional speech synthesis

Even though the expression singing voices synthesis is a very trending topic, the studies about emotion in singing voices are limited. However, in the research about acoustic similarity and differences in the singing voices and speaking voice of Livingstone (2013), they have stated that speech and song may once have existed as a coupled means of vocal communication, a central goal of which was the expression of emotion. Similarly, speech and song have long been considered to share a common 'acoustic code' in the expression of emotion [14]. An other researcher also concluded that in a major review of the subject, music performance, under which singing was classified, shared many of the same acoustic features of speech for the expression of emotion.

Thus, in this section, we briefly discuss about some significant researches about emotional speech, from that having the preliminary knowledge on the main topic of emotional singing voices synthesis. Similar to singing voices synthesis method, the emotional speech synthesis approaches also use the below methods:

- Formant synthesis (Rule-based synthesis): this method provides processing modules for modifying an input neutral speech, following different emotions. Each emotion is described by an acoustic profile, which is obtained from previous works. Montero et al. [15] have succeed to generate Spanish with three basic emotions (hot anger, happy, and sad). In this study, they have used two acoustic profile are: global prosodic and voice quality parameter.

  An other successful research that also used the rule-based approaches to synthesize the emotional speech was proposed by Xue et al. [16]. Similar to the Montero's work, in this research, list of acoustic features was considered includes: F0, Power envelope, Power spectrum, Duration and Voice quality features. They are all the global prosodic acoustic features. In addition, the three-layered model that using for a dimensional approach was introduced. This method help they can generated not only categorical emotion but also the emotion with different intensity, make the synthesize more dynamic.

- Diphone concatenation (concatenation synthesis): this method uses the recordings of a human speaker database and then concatenates them in order to generate the synthetic speech. The common process of this method is they have tried to create the diphone by stretching the speech signal from the middle of one sound (so called 'phone') to the middle of the later sound. This method successes in giving the high natural output quality in compare with formant synthesis. However, they only can control the F0 and timing (sometime also intensity) meanwhile the voice quality control is unachievable.

- Unit selection: This approach has been recognized as giving the highest naturalness of output. It used the similar processing technique as Diphone synthesis, however, instead of using a minimum speech data database, a large database is taken in to account. Beside, based on the required output speech according to different parameters like: duration, phoneme string, and F0, the most suitable size of units will be selected from the database.

However, speech and song have been considered as an entwined form of vocal expression for a long time. Despite their long association, until now there have been no direct comparisons of the acoustic

similarities of speech and song in their expression of emotion. According to the results of preliminary data from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). They showed that speech and song shared many of the same acoustic features in their expression of emotion, while also exhibiting differences that distinguish speech from song. Concurrently, these data support the notion that speech and song may once have emerged from a common vocal origin [17].

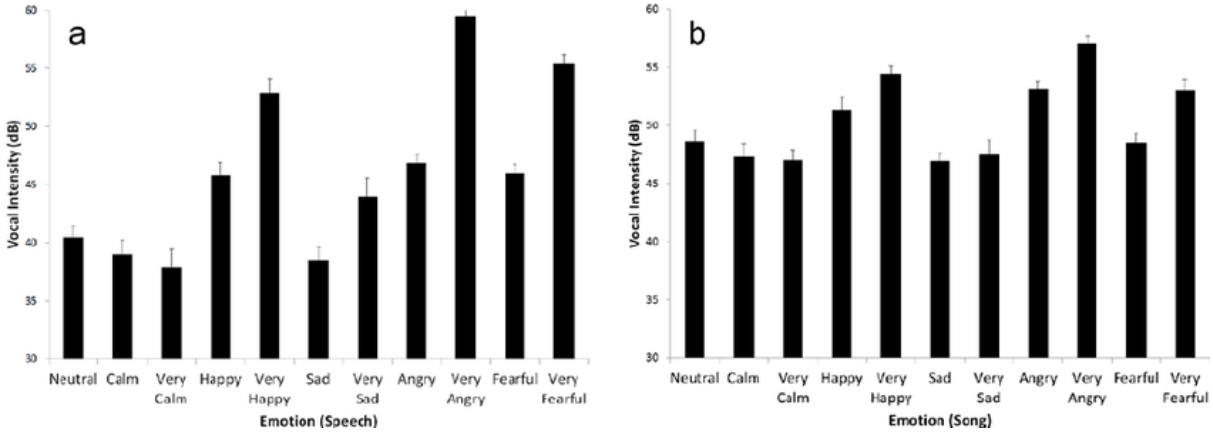


Figure 2.3: Mean vocal intensity across the 11 emotions for speech (a), and song (b). The figure illustrates the two main effects of Domain and Emotion, with song being louder overall than speech, and with speech emotions appearing to show greater variability in intensity than song [3].

2.3 Valence - Activation domain

In this research, a method for representing emotion and its degree is very importance. As we know, human emotional states can be simply represented as categorical items such as neutral, happy, angry, sad. However, deciding the number of category and the standard for each category in the real-life emotion is a touch work. Besides, we can not have the information about degree of this emotion.

An other way to represent the emotion is that we can display them as a point in n-dimensional space, such as two dimension space as Valence-Activation space or three dimension space when they added one more dimension is Dominance. The definition about dimensional domain was first introduced by Russell in 1980 [18] and it was widely used in many research about emotion. Valence is represented from positive to negative scale while activation is represented from calm to exited one and dominance giving the information about the power of the voice that is represented from weak to strong scale.

They are normally using Dominance domain to distinguish Fear and Anger because it related to power, as this research just concentrate on the four basic emotions are Neutral, Happy, Sad and Angry, so that we have decided to use the two-dimension space.

According to Rusell's research, emotion categories can be represented by regions in the V-A space, where the neutral state locates in the center, and the other emotions locate in a specific region as in the Figure. 2.4. Using dimensional approach not only give us the result about category but also the result about the degree of emotion. Plus, they have stated that it is easier for the listener to evaluate the emotion by Valence and Activation instead of categories. Therefore, we decided to use V-A domain to examine the effect of acoustic features in our study.

Figure 2.4: Position of emotions in V-A domain.

Chapter 3

Acoustic Feature Analysis

In this chapter, we first present the corpus that has been used to analysis the acoustic features. Then, the characteristics, important and significant of some acoustic features in singing voices synthesis and emotional speech synthesis will be investigated. In the section 3.3, the method has been used for extracting the acoustic features will be presented. The next section 3.4 is about how we control and modify the acoustic feature to obtain the emotional singing voice. Finally, the experimental examination and subjective test results will be presented in section 3.5.1 and section 3.5.2 respectively.

3.1 Corpus

- Emotional singing voice database:

  For F0 and amplitude envelope analysis, the Ryerson Audio - Visual Database of Emotional Speech and Song (RAVDESS) was created by Livingstone [19] has been used. The RAVDESS, which is being prepared for public release, consists of speech and singing voices of 23 professional actors (12 males and 11 females), with a large range of different emotions. The emotional speech database contains 8 different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) while the emotional song has 6 different emotions (neutral, calm, happy, sad, angry, fearful), each with two intensities. The total size of the database is 24Gb and the size of singing voice only is 500MB, and will be released with perceptual validation data, acoustic analyses, and facial motion analyses. The purpose for creating the RAVDESS was to provide researchers with an open-access repository of high-quality, audio-visual recordings of speech and song in North American English. Perceptual accuracy of the acoustic recordings used in the present analysis was confirmed in a separate pilot experiment.

  Regarding the musical score, three isochronous melodies were used: F3 - F3 - G3 - G3 - F3 - E3 - F3 for the neutral condition, F3 - F3 - A3 - A3 - F3 - E3 - F3 for the positive emotions (calm and happy, and F3 - F3 - Ab3 - Ab3 - F3 - E3 - F3 for the negative emotions (sad, angry, and fearful). These melodies were presented to the actors as piano MIDI tones of fixed acoustic intensity, includes six eighth notes (300ms) and finally is a quarter note (600ms).

- Emotional vowel speaking voice database

The reason we do not analysis the spectral feature from the RAVDESS database is this spectral feature includes both the vowel and consonant part. In addition, they have stated that it is having the similar about tendency of spectral feature in speaking voice and sining voice, therefore, we decided to use the emotional speech vowel database. This database were recorded from 10 speakers (5 males and 5 females) and included 8 different emotions (Neutral, Afraid, Anger, Disgust, Joy, Relax, Sad, Surprise), each emotion has 3 different intensities of expression are: little, normal and strong. The utterances of normal degree of 4 typical emotions (Neutral, Joy, Anger and Sad) are used for the analysis step.

3.2 Determining of significant acoustic features

This study aims to investigate the correlates of acoustics feature to the emotional singing voice. Hence, a selection of distinctive features of emotional singing voice was considered. Specifically, we consider the analysis of the basic acoustic features as F0, amplitude envelope and spectral feature, which represent the differences between an emotional and neutral singing voice.

- **F0:** F0 features generally have an important effect on the emotional expression of speech. Also, F0 fluctuations have the most influence on the singing-ness of synthesized voice in Saitou's model [20]. This implies that F0 contours possibly contribute significantly to the emotion of a singing voice.

- **Amplitude envelope:** This is a prosodic feature that contributes to the emotional expression [16]. Oncley [21] found that a singer-formant amplitude of a singing voice is modulated in the synchronization with the frequency modulation of each vibrato in the F0 contour. However, we realize that there exists a high correlation between the entire F0 contour and the amplitude envelope. We therefore propose a rule to re-synthesize such the amplitude envelope, for the purpose of synchronizing with the F0 contour, not only in the vibrato part but also in the overshoot and preparation ones

- **Spectral sequence:** A spectral sequence contains two parameters, including spectral tilts and spectral balance, also bring substantial information of emotional expressions [22]. The preliminary results show that the re-synthesized voices generated using a typical spectral sequence of each emotion, have presented different impressions to listeners.

3.3 Extraction of acoustic feature

- Preprocessing database

  Each utterance in singing voice database was segmented and labeled carefully, this is the most important step to have the dataset using for analysis, therefore, it took a lot of time for doing this step carefully and exactly. The Praat software [23] had been used in this step to visualize the data for segmentation and labeling them. We extracted the note changing points of each utterance and then labeled them in order to use it later.

  Regarding vowel speaking voice database, only the voiced part of an utterance is used, the silence parts are cut out.

- F0 contour

  After finishing the pre-processing tasks, the crucial properties of 4 components of F0 contour that well reflect the singingness including overshoot, vibrato, preparation, fine fluctuation will be thoroughly studied among different emotions.

  - F0 estimation method

    There exists several approaches for F0 estimation as below:

    1. YIN pitch detection was proposed by A. Cheveigne and H. Kawahara in 2002 [24]. YIN is well appropriate with the high-pitched speech and singing voice because there is no upper constrain among the frequency range. According to the autocorrelation method and different modifications, this method is quite simple and efficiently. Especially, only a small number of parameters will be tuned.
    2. A Sawtooth Waveform Inspired Pitch Estimator (SWIPE) was proposed by A. CAMACHO in 2008 [25] in order to suit with voices and music. This method chooses the sawtooth waveform that has the 'best matches' spectrum as the input one, and then returns the fundamental frequency of this waveform as the estimated pitch for input signal. In order to have the smooth peaks with decaying amplitudes for correlating with harmonics of the signal, a decaying cosine kernel will be used (based on the characteristic of giving an extension to older frequency).
    3. WORLD was proposed by M. Morise et al. in 2016 for the purpose of giving the proficient sound value in a real-time processing. It is a vocoder-based speech synthesis system [26] that return a quickly and definitive F0 estimated result, however, the estimation execution is not as good as YIN or SWIPE. The F0 estimation algorithm of WORLD is consist of 3 steps. After passing the low-pass-filtering with number of cut off frequencies at a first step, possibility F0 will be calculated and finally the one with the highest reliability is selected.
    4. Even STRAIGHT - TEMPO was proposed by Kawahara from very early (1997) [27], it still considered as a universal tool for manipulate speech parameter as pitch, speaking rate or vocal tract length. By implementing the instantaneous frequency method, TEMPO was confirmed that having the best result on extracting the fine fluctuation in F0 contour than others.

Although the first 3 methods are good at extracting the F0 contour, they sometime give the result with suddenly change in the curve, especially at the position of node change. We thought that a suitable combination of several approaches might give us more robust pitch estimation.Therefore, we have decided to remove all the strange parts of previous methods and combine with the F0 detected by TEMPO.

- F0 contour characteristics

  Figure 3.1 shows an estimated F0 contour of an utterance in the RAVDESS database. The coordinate was displayed in log-frequency scale. This figure also shows a melody component that represents note changing in the extracted F0. Figure 3.2 displays the four F0 fluctuations have been observed in the F0 contours:

  * **Overshoot**: the deflection exceeding (over) in compare with the target note after note changes. It consists of 2 following components:
    - Portamento: fundamental frequency change with slope at pitch change.
    - Overshoot Extent: instantaneous vibration component immediately after pitch change.
  * **Vibrato**: quasi-periodic frequency modulation (5-8 Hz).
  * **Fine fluctuation**: inconsistent fine fluctuation related to the modulation frequency that higher than 10Hz.
  * **Preparation**: deflection in the opposite direction of a note change observed just before the note changes

- F0 control model

  The outline of the proposed singing voice F0 control model is shown in Figure 3.3. Each component of the diagram is described as follow:

  * **Melody Component**: a superposition of step functions. Created from the score, it represents a rectangular melody change.
  * **Overshoot**: it is described by the damped second order model.
  * **Vibrato**: it is described by the steady-state vibration of the second order oscillation model. Controls the periodic vibration that occurs when the same tone continues.
  * **Preparation**: it is described by damped second order model. The component which instantaneously fluctuates in the opposite direction to the change immediately before the node change is controlled.
  * **Fine-fluctuation**: Describe based on white noise. Apply irregular and fine vibration to the entire fundamental frequency change.

- Parameters Value

  In [20], Saitou et al. had proposed the model to synthesize proper F0 contour that using for speech-to-singing synthesis system. Based on this model, after a score-based melody contour were inputted, the four types of F0 fluctuations will be added to this input as the block diagram on Figure. 3.3. Regarding the melody contour, the sum of consecutive step functions will be used to described them, each step function corresponding to a music note. As in the figure 3.3, the transfer function of a second-order system is used

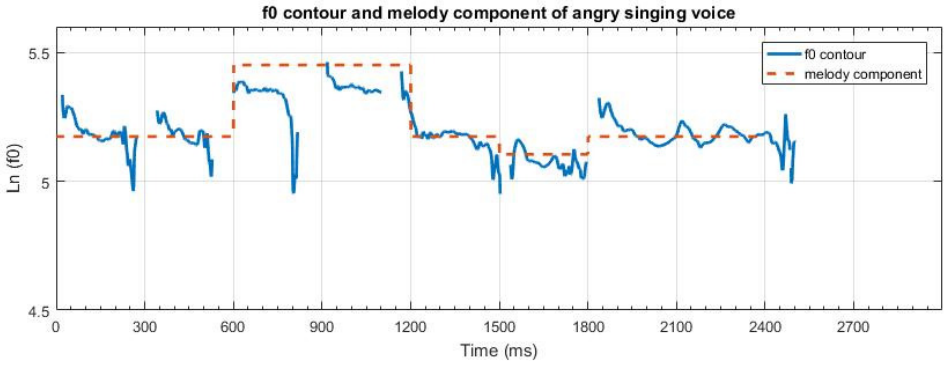


Figure 3.1: Melody component and extracted F0 fluctuations.

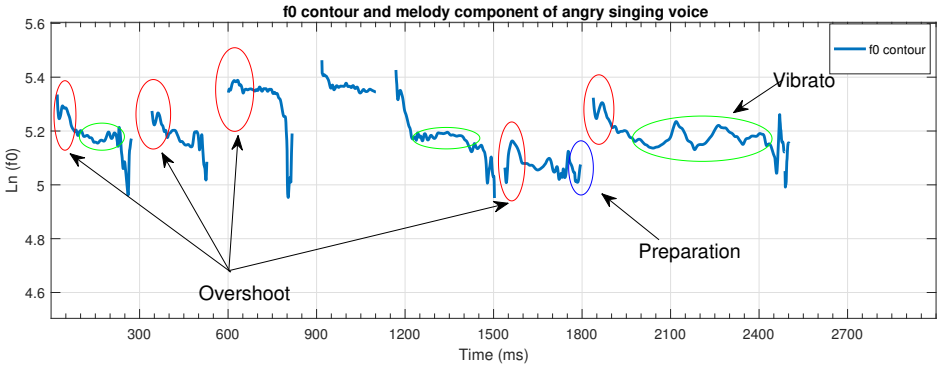


Figure 3.2: Detected 4 types of F0 fluctuations: overshoot, vibrato, preparation. Fine fluctuation along the entire contour.

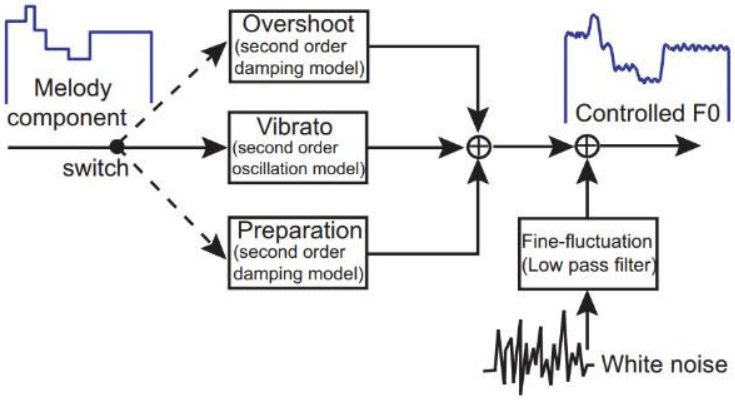


Figure 3.3: Block diagram of the F0 control model for singing voices was proposed by Saitou et al. [4].

to represented the overshoot, vibrato and preparation. This function is as shown in the

equation 3.1

$$H(s) = \frac{k}{(s^2 + 2\zeta\omega s + \omega^2)} \tag{3.1}$$

Where ω is the natural frequency and ζ is the damping coefficient and k is the proportional gain of the system. The impulse response of $H(s)$ can be obtained as:

$$h(t) = \begin{cases} \frac{k}{(2\sqrt{\zeta^2-1})}(exp(\lambda_1\omega t) - exp(\lambda_2\omega t))), & |\zeta| > 1 \\ \frac{k}{\sqrt{1-\zeta^2}}exp(-\zeta\omega t)sin(\sqrt{1-\zeta^2}\omega t), & 0 < |\zeta| < 1 \\ ktexp(-\omega t), & |\zeta| = 1 \\ \frac{k}{\omega}sin(\omega t), & |\zeta| = 0 \end{cases} \tag{3.2}$$

Where $\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$, $\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$

Each fluctuation will be represented by equation 3.2 as follows:

1. **Overshoot:** the second-order damping model ($0 < |\zeta| < 1$)
2. **Vibrato:** the second-order oscillation model ($|\zeta| = 0$)
3. **Preparation:** the second-order damping model ($0 < |\zeta| < 1$)

The constructed F0 contour will be controlled by changing the value of system parameters ω, ζ and k. To minimize the errors between the generated F0 and actual ones, the nonlinear least-square-error method [28] with the calculate equation as the equation 3.3 will be used. Regarding fine fluctuation, an irregular frequency fluctuation higher than 10 Hz, it will be generated from white noise that pass through high-pass filter first and then normalized the amplitude.

$$E = \sqrt{\frac{a}{N}\sum_{M+N}^{m=M+1}(x(mT) - y(mt))^2} \tag{3.3}$$

- Amplitude Envelope

  Instead of using Hilbert transform to get the Instantaneous Amplitude, we have taken the power envelope of signal by picking those peaks of the absolute value of signal and interpolate them later. The reason we have not directly used the amplitude envelope from singing voice is this envelope contain the consonant parts, therefore, we decided to modify the envelope by taking the peaks of absolute value of power and then interpolate them, finally using low pass filter to get the power envelope as equation. 3.4

$$powEnv = LPF[interpolate[peaksPicking[|x(t)|]]] \tag{3.4}$$

The extracted amplitude envelope is showed in the Figure. 3.8

However, the results seem like not fit with the synthesized voices when the naturalness of voice was decreased a lot. It comes with the other idea is that, we tried to use directly F0 contour as the amplitude envelope of synthesized data, and the synthesized voices are more natural at all. So, the precises method to generate the amplitude envelop of synthesizing from F0 contour is taken into account.

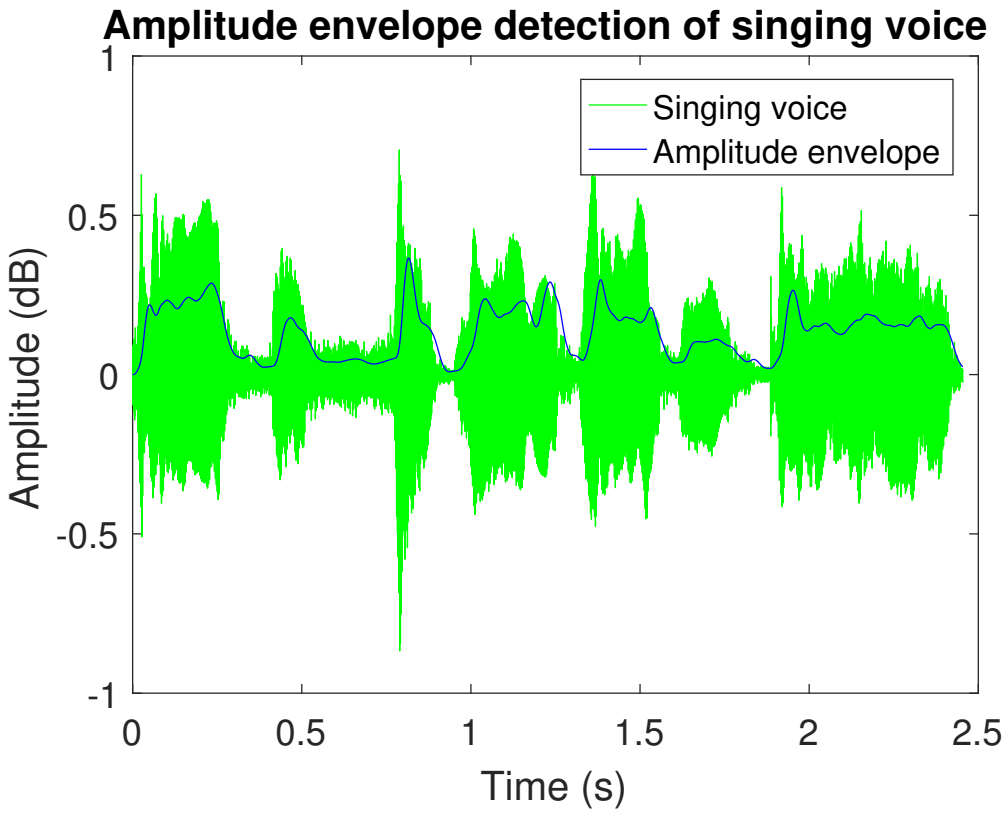


Figure 3.4: Amplitude Envelope detected from singing voice.

- Spectral Sequence

  The normal emotion's intensity utterance of each emotion has been selected as the typical spectral sequence for each emotion. Similar to F0 contour, the spectral feature is extracted by using the analysis function of STRAIGHT. After extracted the spectral of different emotional speaking voice. We realize that the fluctuation of the spectral are different among emotions as in the Figure 3.5

  According to the figure, we can see that there exist a clearly different about the fluctuation (shape) of spectral sequences among emotions, specifically, the spectral sequence of Angry voice with the special formant trajectory at the end of the voice.

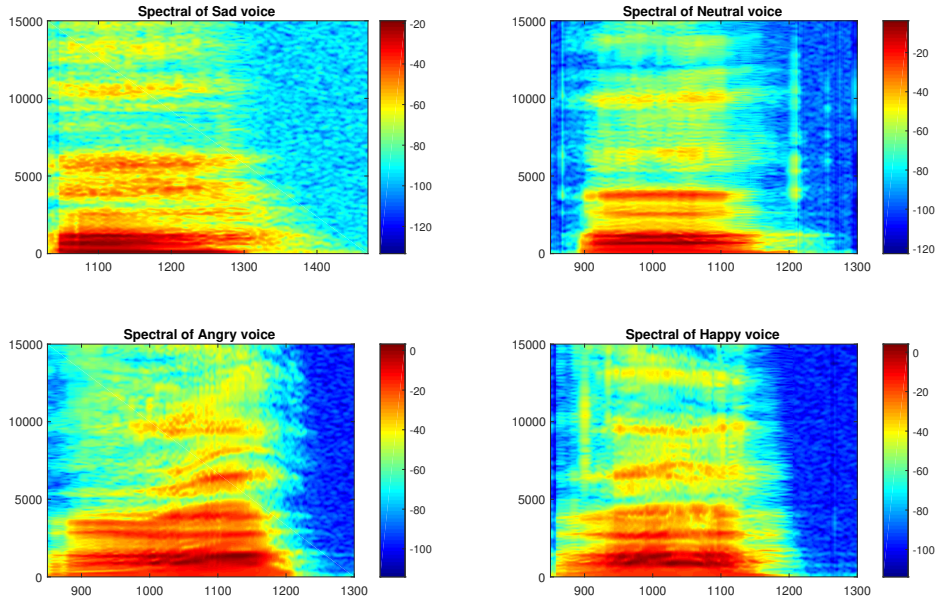


Figure 3.5: The spectral sequences of different emotions.

3.4 Acoustic feature modification

After finishes selected and extracted the acoustic features, the appropriate rules for modifying them will be proposed to have the singing voices with emotions.

3.4.1 Fundamental Frequency (F0)

As we have stated in the Section 3.2, the constructed F0 contour will be controlled by changing the value of system parameters ω, ζ and k. The value of parameters for each emotion will be set as table. 3.1

Table 3.1: Parameters value for F0 control model.

	Overshoot		Preparation		Vibrato	
	ω	ζ	ω	ζ	ω	k
Neutral	0.05393	0.521	0.03135	0.3445	0.035	0.0008
Sad	0.0448	0.5898	0.04628	0.4772	0.038	0.0009
Happy	0.0443	0.6156	0.04048	0.4670	0.045	0.0011
Angry	0.03778	0.521	0.03135	0.3445	0.045	0.0013

3.4.2 Spectral Sequence

The spectral sequences have been used for synthesizing singing voices are extracted from the emotional vowel speaking voices using STRAIGHT. After obtaining the spectral sequences, which representing the emotions, we carefully lengthen it to the desired duration. To preserve the fine fluctuation in the spectral sequences and to keep naturalness of the synthesized voices, we repeat each frame of the vowel sounds with the same number of repeating times as in the Figure 3.6 and 3.7.

The other element of voice that was extracted from STRAIGHT is aperiodicity component (ap) will also be lengthen in the same way as spectral sequences in order to match with the new spectral sequences.

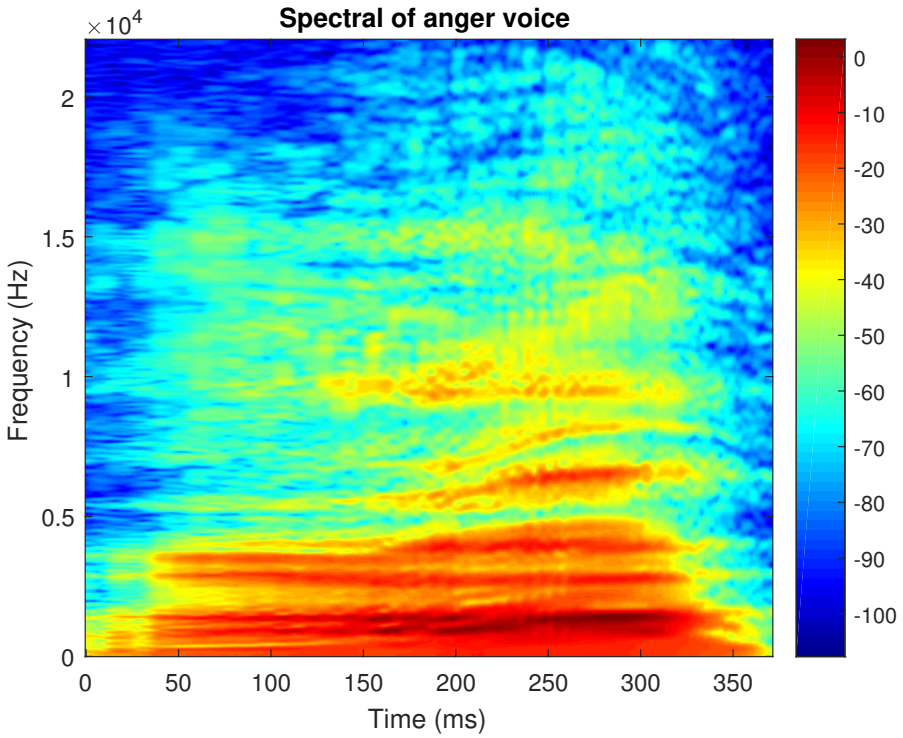


Figure 3.6: Spectral of anger speaking voice.

3.4.3 Amplitude Envelope

Oncley [21] had stated that there exist the synchronization between the formant amplitude modulation of singing voice and frequency modulation of each vibrato in F0 contour. Beside, after examine and analysis the database, we found that there is also a high synchronization between the entire F0 contour and amplitude envelope. Hence, we have decided to modify the amplitude envelope by using the F0 fluctuation. As the fluctuation of amplitude is larger than F0's, we modulate the

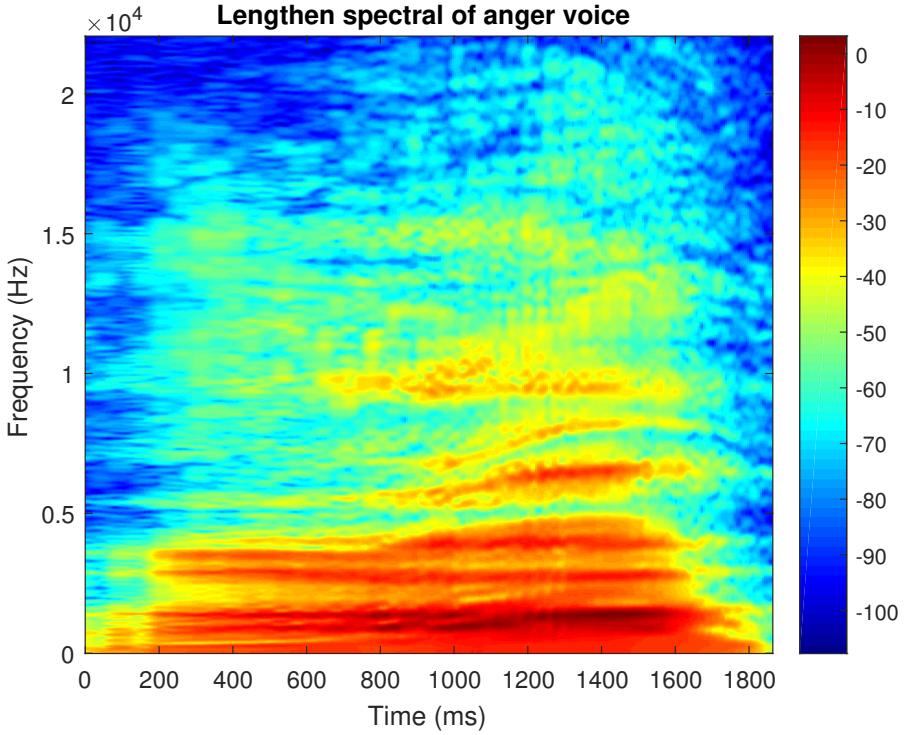


Figure 3.7: Lengthen spectral of anger speaking voice.

amplitude envelope according to the equation 3.5

$$ampEnv = [\mathrm{k} \cdot F_0] \tag{3.5}$$

With each specify part, we use the different coefficients. The modification of overshoot and preparation follows the equation 3.6.

$$ampEnv = interpolate(\mathrm{k} \cdot peakPicking(F_0)) \tag{3.6}$$

Where $\mathrm{k} = [\mathrm{k}_{ovs}\mathrm{k}_{pre}]$ with k_{ovs} for overshoot part and k_{pre} for preparation part ($1 < \mathrm{k}_{ovs},\mathrm{k}_{pre} < 2$).

Regarding vibrato part, we use Saitou's results as in the equation 3.7 to modify it.

$$ampEnv = (1 + k_{am}sin(2\pi f_{am}t))F_0(t) \tag{3.7}$$

Where f_{am} is the rate of Amplitude Modulation (AM) and k_{am} is the extend of AM.
All the set of K, k_{am} and f_{am} are different among emotions.
The modulated amplitude envelope has been shown in the Figure 3.8

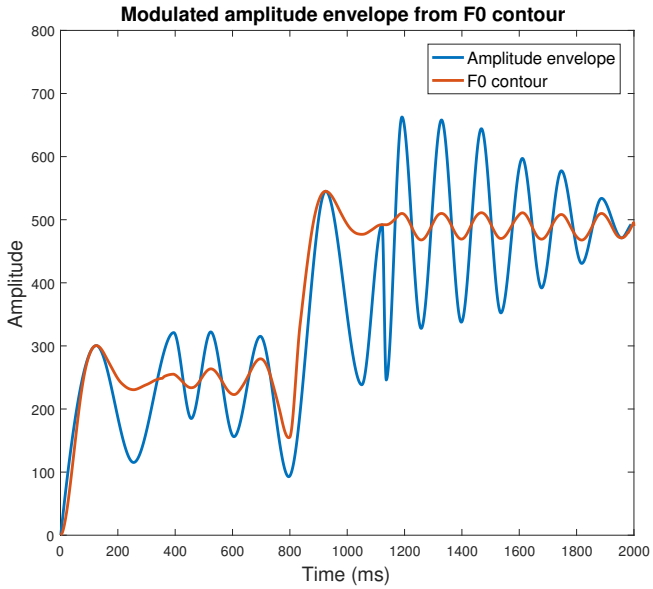


Figure 3.8: Modulating amplitude envelope from F0 contour.

3.5 Experimental Examination

In this section, the listening experimental will be conducted to examine the correlates of not only each acoustic feature but also the combination of acoustics features to the emotional singing voice.

As we have stated before, the singing voices with vowel 'a' only will be synthesized in order to remove the linguistic information of the voice, help the listener concentrate on the affect of the acoustic features to the voice. We synthesized the singing voice by using the diagram as in Figure 3.9.

The synthesized voices for each emotion include 7 types as below:

- **All:** the three acoustic features are modified.
- **F0 and Spec:** F0 contour and Spectral Sequences are modified.
- **F0 and Amp:** F0 contour and Amplitude Envelope are modified.
- **Spec and Amp:** Spectral Sequences and Amplitude Envelope are modified.
- **F0:** Only F0 contour is modified.
- **Spec:** Only Spectral Sequence is modified
- **Amp:** Only Amplitude Envelope is modified.

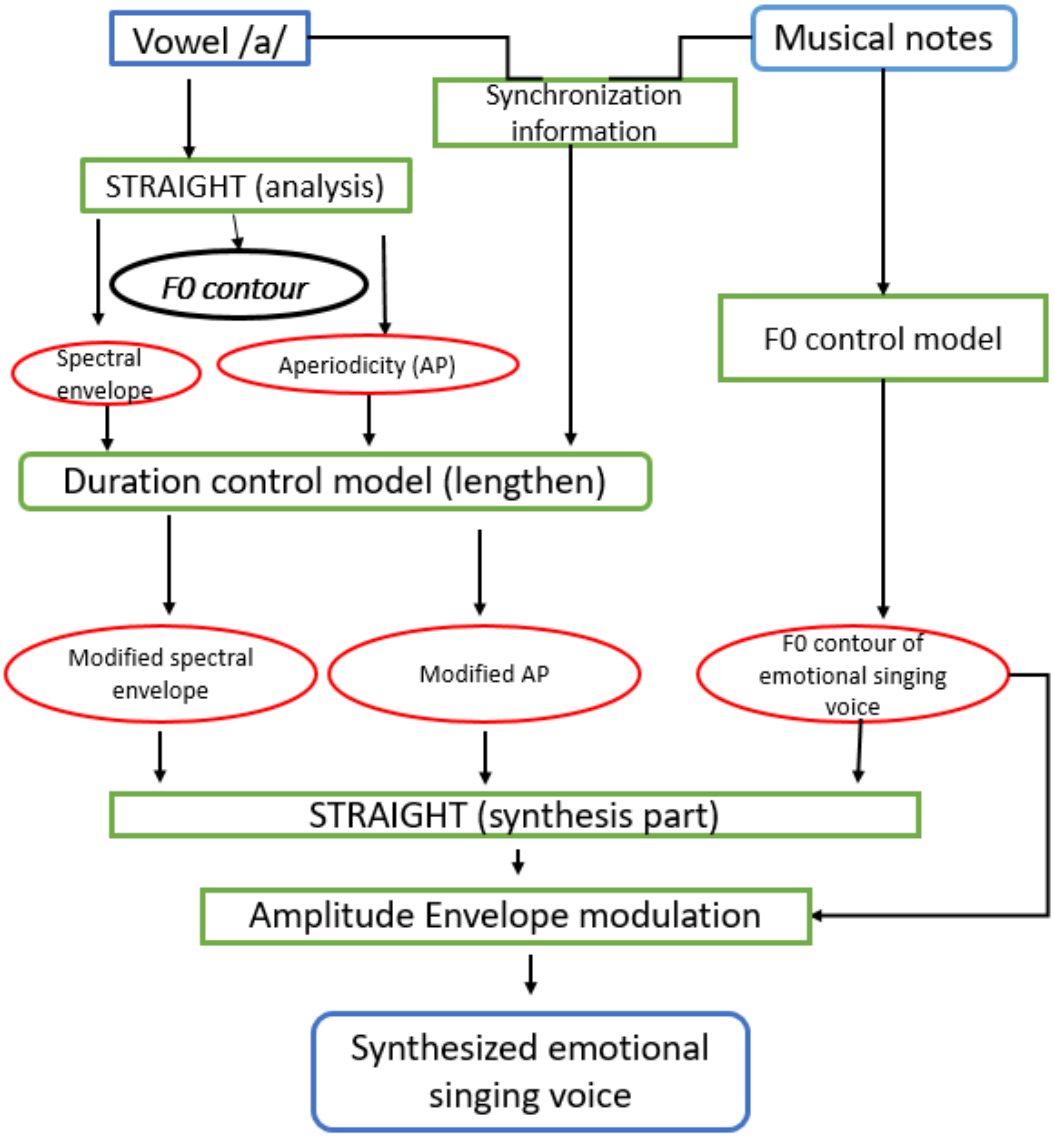


Figure 3.9: Synthesis diagram for emotional singing voice synthesis.

3.5.1 Subjective test

This experiment is conducted to investigate the importance of each acoustic feature in the emotional singing voice. It is also to confirm that whether the expressiveness of singing voice can be perceived or not with the vowel /a/ only. The listeners are required to listen to the synthesized singing voice and evaluate the degree of Valence and Activation of this voice.

Experimental Description:

- Subject: Ten students (eight males and two females; average age: 25 years old) who have normal hearing level without any hearing loss are invited to do the listening tests. Among the listeners, 3 people are having the knowledge about Valence - Activation domain and 1 people has the special training about music before. All the listeners listen to music everyday.

- Stimuli: In the listening test, the total 44 stimuli (22 female voices and 22 male voices) will be synthesized for the listening test. Each stimulus have been synthesized by modifying the different combination of acoustic features. Among 22 stimuli for each gender, there are seven utterances for each emotion: sad, angry and happy and one neutral voice. The stimuli were synthesized by using the first seven musical notes of a children 'ABC song'.

- Procedure: In a soundproof room, subjects will be invited to listen to the stimuli, which are presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER). The average sound volume is about 60-69 dB. Before the test, the volume of stimuli will be carefully checked by 4 ways: normal hearing, pure tone and pink noise, artificial ear, amplifier level. The detail volume for each side of ear (left and right) also be carefully checked.

  - Instruction: before doing the experiments, listener is asked for their personal information (name, age, student ID, nationality) and also some other information about their experience with V-A domain and music before.

    An instruction file is given to help listener understand about this test.

  - Familiarize test: the familiarize listening test will be carried out to help listener get familiar with the stimulus, the user interface and how to evaluate the sound. Any question is welcome after this test.

  - Main test: in the main test, for each dimension (valence and activation), subjects will listen to the stimuli four times. The last three times result will be used. The reason for this is that they are supposed to have an impression of the stimulus first, and then doing the evaluate for this dimension from -2 to 2 in 21 value scales. In addition, in order to support for the statistical analysis of results, we need to have the results from more than two times of a listener to have the precise outcomes.

  In addition, the listener are required to do the evaluations of valence and activation separately. It is needed for the break of at least one day. By then, they will not misunderstand the definition of valence and activation.

- User interface: each dimension will be evaluated using 21 scales (Valence: from [Very Negative] to [Very Positive]; Activation: from [Very Calm] to [Very Excited], all of them will be evaluated within the range from -2 to 2 by 0.2 step). Subjects will evaluate the degree by using the graphic user interface as Figure 3.10 and Figure 3.11. We allow the subjects to listen to the stimuli repeatably in each evaluation.

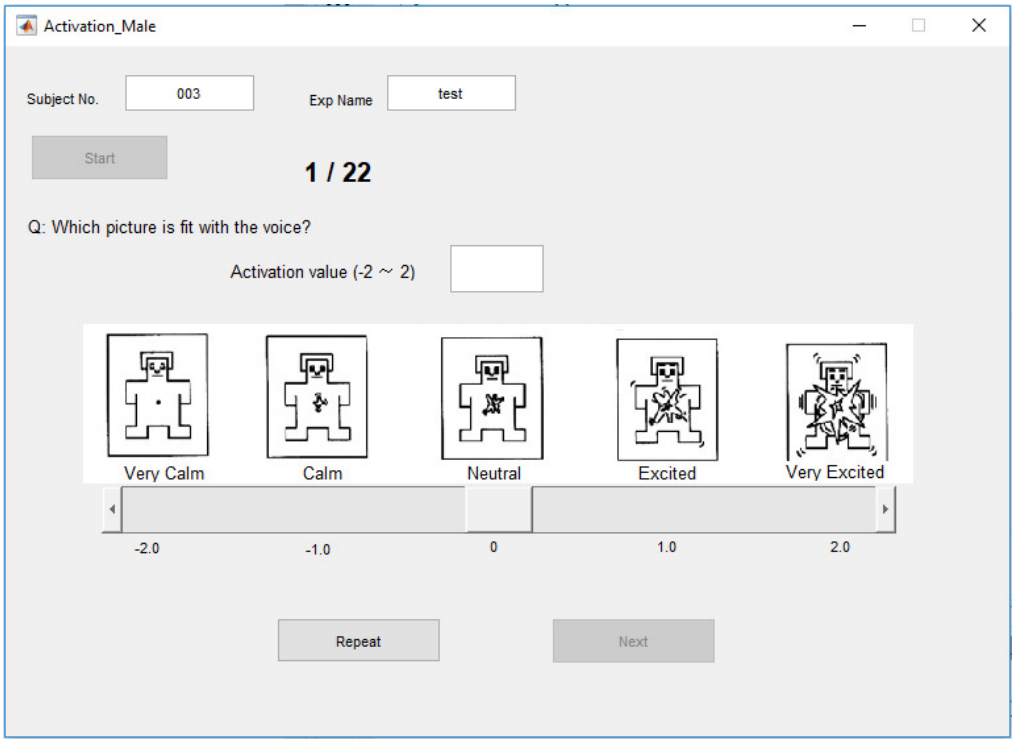


Figure 3.10: User interface used for activation evaluation.

- Estimation time for the test: 2 hours
  - Each stimulus will last for 6.5 seconds
  - 22 stimuli per session
  - There are total 16 sessions (Valence and Activation; Male voices and Female voices; 4 times)
  - Break time after each session: 3 minus

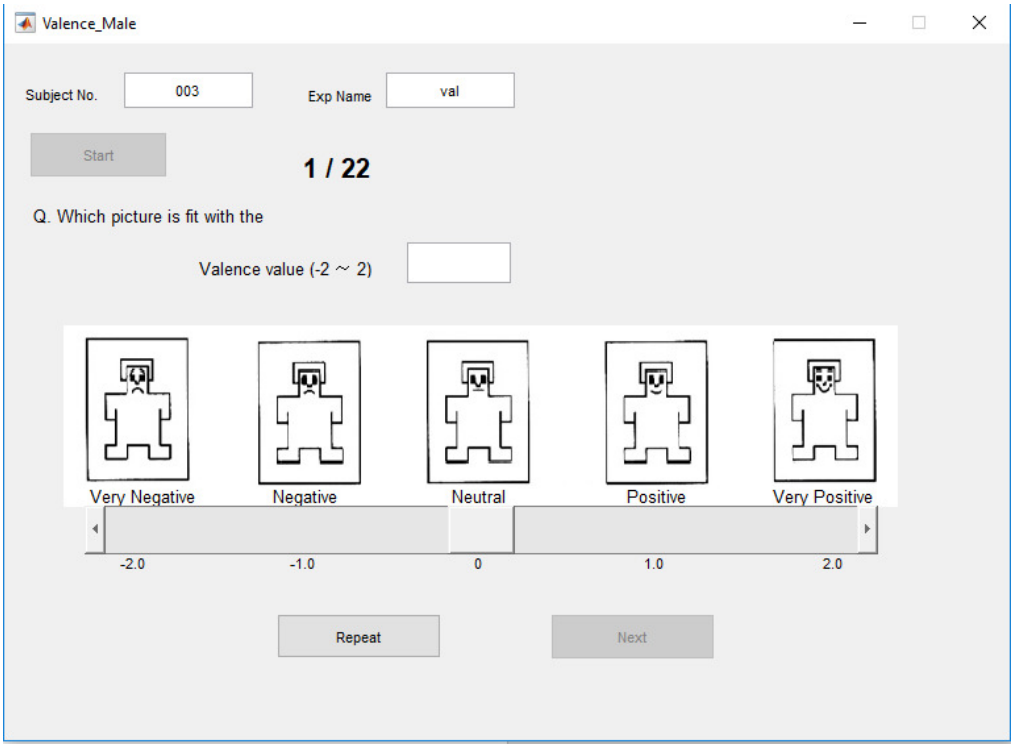


Figure 3.11: User interface used for valence evaluation.

3.5.2 Results

Averages of evaluation positions in V-A space of all the listeners are shown in Figure. 3.12 for female voice and Figure. 3.13 for male voice.

As we can see in the figures, the angry and happy stimuli are correctly distributed in its own region in V-A space while sad voices are mostly selected as the neutral voice.

According to section 2.3, the emotion categories can be represented as regions in the V-A space, where the neutral state locates in the center, and the other emotion locates in a specific region as shown in Figure. 3.14. Using the dimensional approach not only gives us the result about category but also the result about the degree of the emotion. Therefore, to investigate the effect of these acoustic features to emotion in singing voice, we calculate the distance and direction between the synthesized voice and neutral one.

Regarding the direction, it will be the angle created by the line go through this point and the origin point and the positive x-axis as described in the Figure. 3.14. The calculated results are displayed in the table 3.2 and 3.3

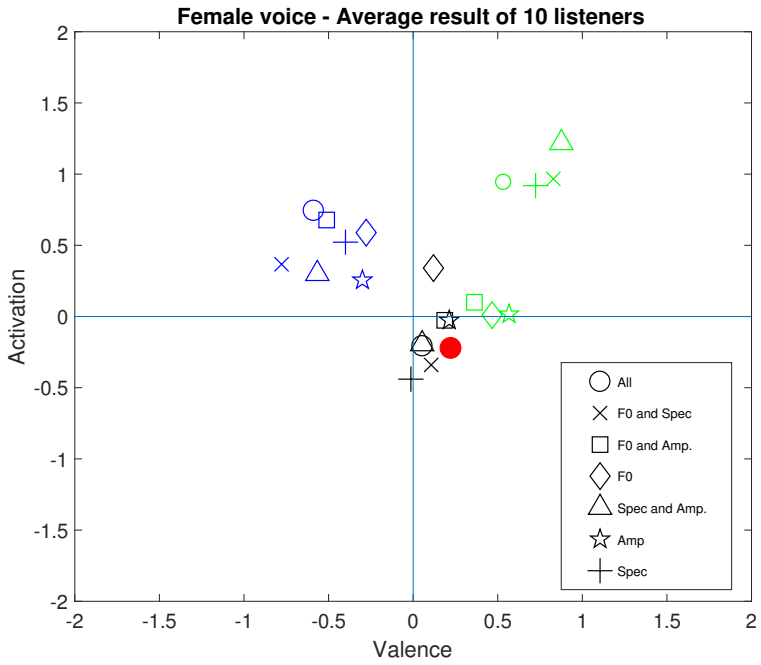


Figure 3.12: Position in V-A domain of stimuli of female singer. Green: happy; Blue: angry; Black: sad; Red: neutral.

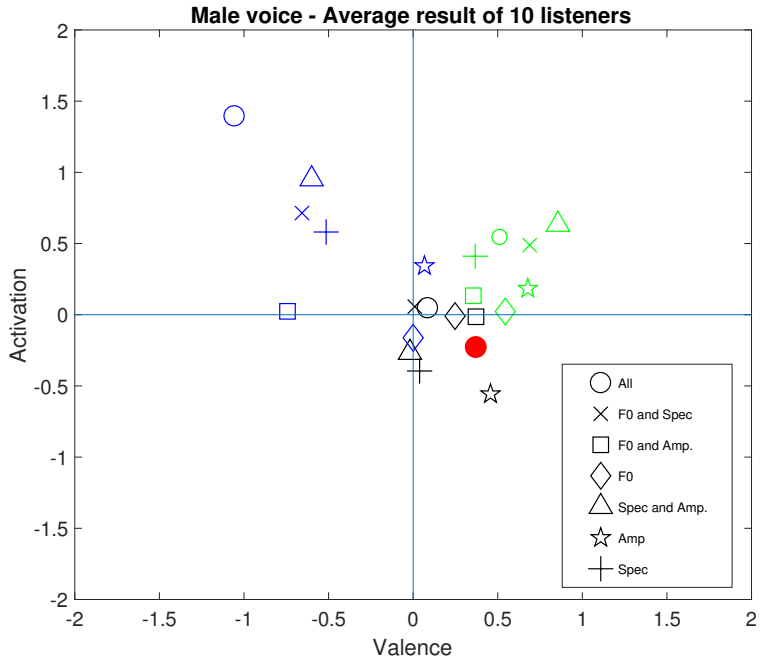


Figure 3.13: Position in V-A domain of stimuli of male singer. Green: happy; Blue: angry; Black: sad; Red: neutral.

Table 3.2: Average distance and direction between the emotional stimulus position to neutral position on V-A domain - Female singer.

	Emotion	F0	Spectral	Amplitude Envelope	F0 and Spectral	F0 and Amplitude Envelope	Spectral and Ampitude Envelop	F0, Spectral and Amplitude Envelope
Distance	Happy	0.34	1.25	0.42	1.33	0.35	1.58	1.11
	Sad	0.57	0.32	0.2	0.17	0.2	0.17	0.17
	Angry	0.95	0.97	0.7	1.16	1.16	0.95	1.26
Direction	Happy	43	66	35	63	67	66	73
	Sad	100	223	93	225	100	170	174
	Angry	122	130	138	150	129	146	130

Table 3.3: Average distance and direction between the emotional stimulus position to neutral position on V-A domain - Male singer.

	Emotion	F0	Spectral	Amplitude Envelope	F0 and Spectral	F0 and Amplitude Envelope	Spectral and Ampitude Envelop	F0, Spectral and Amplitude Envelope
Distance	Happy	0.3	0.64	0.51	0.78	0.36	1	0.79
	Sad	0.25	0.37	0.34	0.46	0.21	0.4	0.4
	Angry	0.38	1.2	0.65	1.4	1.14	1.53	2.2
Direction	Happy	55	90	53	66	93	61	80
	Sad	119	206	285	141	90	186	136
	Angry	169	137	118	137	167	129	131

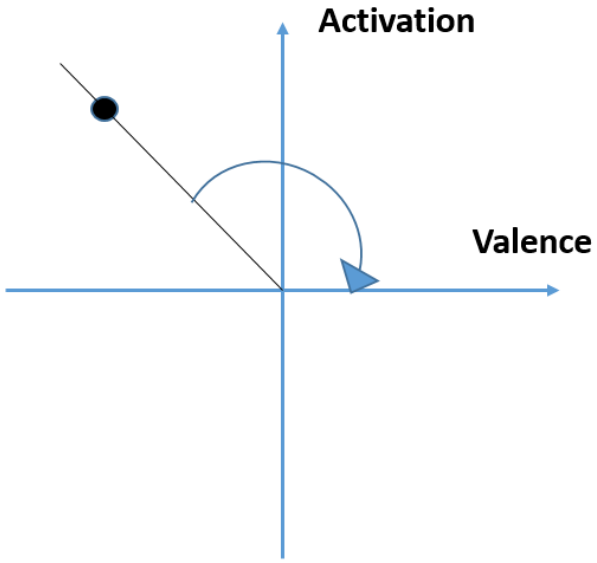


Figure 3.14: Direction of emotional stimulus position in V-A domain.

The farthest distance from the position of emotional stimulus to the neutral one is 1.58 for Happy voice (modify Spectral and Amplitude Envelope) of female singer and 2.2 for Angry voice (modify all three acoustic features) of male singer. The distance is high enough for recognition of emotion for listeners. Besides, all the happy and angry voices are having the same direction with the definition as Figure. 2.4, meanwhile, the sad stimulus are scattered around the Neutral one.

3.5.3 Discussion

From the results of the experimental examination we can see some significant points as below:

- Listener can distinguish difference emotion among stimuli, especially Anger voice and Happy voice.
- The combination having 'spectral' modification giving the better result than others.

  This can be stated that spectral feature is the most important acoustic feature for giving emotion expression in the singing voice.
- The combination having highest degree of emotion is different among emotions (Spectral and Amplitude Envelope for Happy; Spectral for Sad and All of three acoustic feature for Angry).
- The results of Sad voice is still not obtained the good result. Almost all the Sad voices are evaluated as neutral voice.

3.6 General Discussion

According to the listening test results, we can see that each acoustic feature has a different contribution to the emotional expression of a singing voice. In addition, the contribution also depends on the emotion and gender of the singer. We discuss in detail our results for each emotion as below:

- **Happy**: this emotion obtains the highest degree when modifying the Spectral Sequences and the Amplitude Envelope for both genders. After that, the combinations of (i) Spectral Sequences and F0 contour, and (ii) all three features, have achieved the second rank of emotion degree. Meanwhile, the stimuli which are generated by modifying F0, Amplitude Envelope or both of them produce very low results.

  Regardless of different results achieved, listeners are able to detect all the Happy stimuli in all the combinations above. Therefore, it can be stated that all the three basic acoustic features (Spectral Sequences, F0 and Amplitude Envelope) are significantly contributed to the emotional expression of singing voice.

- **Angry**: different from Happy, the listeners feel most the Angry expression in the stimuli when we modify all of the three acoustic features. In particular, the male voice has achieved a very high results when having the furthest distance (2.2) and the most appropriate direction (131 degree).

  Is is very interesting as that the outcomes of male and female voices are relatively different from each other, even though they are synthesized by using the same method. Regarding the female voices, modifying Spectral Sequences (All, Spec and Amp, F0 and Spec, Spec) or not, there is not too much difference among 7 stimuli. In contrast, it produces diverse results between these two situations in male voices.

  Nonetheless, all of the Angry stimuli have been efficiently detected in the listening test. It is again confirmed that all of the three acoustic features we have chosen play very important roles in the expression of singing voice.

- **Sad**: Sad voice is the most difficult one to synthesize, due to its characteristics of F0 and power. Therefore, the listening test results are not as good as the results of Angry and Happy voices. The listeners succeed in recognizing this emotion when we modify the Spectral sequences for both genders. Besides, the results that are generated by changing a combination that contains Spectral sequences also achieve a good emotion degree.

In conclusion, by modifying different groups of the three acoustic features, we have obtained the good quality emotional singing voice. In addition, we have evaluated the contribution of not only individual acoustic feature, but also their combinations, to the emotional expression of singing voices. It will help for the further study about singing voices, especially in terms of emotions. Nevertheless, these results here are limited to modifying three important acoustic features. Hence, it is needed to investigate about other features, such as duration, power spectrum, voice quality feature, etc.,to improve the quality of synthesized singing voices.

Chapter 4

Conclusion

4.1 Summary

The final goal of this study is to create a computer-based application for synthesizing singing voices with emotions. Given a neutral speech in reading a song's lyric, and its corresponding musical score, it produces a singing voices with a specific emotion (e.g., sad, happy, or angry, etc.). In order to achieve that, we have analyzed different acoustic features in a singing voice, in terms of their contributions to the emotional expression. This analysis includes three main tasks as below:

1. Acoustic feature extraction: A list of features are extracted to analyze, based on their representations in the singing-ness and emotion of the voice.

   Basic acoustic features, including F0 and spectral sequences, which are significantly related to the singing-ness and emotion are extracted first. After that, amplitude envelope, another feature linked to singing-ness and naturalness, is also analyzed.

2. Acoustic feature modification: Important properties of each basic acoustic feature among different emotions are carefully investigated. Based on our analysis results, a set of suitable rules for modifying that acoustic feature were proposed.

   Next, singing voices are synthesized by using the above proposed rule set. Moreover, to remove the linguistic information, especially the one related to emotions that can be hinted by the lyric, we choose to synthesize the voices with only the vower /a/. This helps listeners concentrate on the emotional expression, which is produced by the acoustic features.

3. Experimental examination: Different combinations taken from the three acoustic features, including F0, spectral sequences, and amplitude envelope, will be used to synthesize a singing voice. After that, we perform a subjective test, in order to evaluate the effect of each combination to the degree of the emotion in the synthesized voice. The experiment results have been displayed in the V-A domain and then calculated the distances and directions, compared to the position of the neutral singing voice.

Our experiments illustrate that all the synthesized voices with modified spectral sequence obtain very good results. Moreover, it shows that if a voice had synthesized by modifying all of three acoustic features even get the better results.

With the analysis results about three significant acoustic features and proposed rules for modifying them to having emotional singing voices, the synthesized voices include consonants need to be investigated in the next step. Nonetheless, the other acoustic features, which may affect to the emotion of singing voice (duration, formants, etc), also need to be considered in the future work.

4.2 Contribution

This research succeeds in extracting the three appropriate acoustic features and manipulating them, in order to obtain a singing voice with different emotions. The three acoustic features are F0 contour, spectral features and amplitude envelope.

The synthesized singing voices, even without the linguistic information, contain emotional expression, and listeners can distinguish them adequately. In addition, by carrying out a subjective test and analyzing the results, the spectral feature is determined as the most affected acoustic feature to the emotion of singing voices. However, the analysis results also show that it is necessary to modify all the three acoustic features to obtain the highest naturalness and singing-ness.

4.3 Remaining works

Since the results in producing a sad singing voice is relatively low, we need to find a suitable method to modify the acoustic features for sad voice. Besides, another important task is to study about other features that help determine information about emotions in a singing voice. Nonetheless, the study finishes with synthesizing the voices with vowel only, therefore the synthesizing method for both vowel and consonant needs to be further improved.

Bibliography

[1] M. Umbert, J. bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis," *IEEE signal processing magazine*, 2015.

[2] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system - Sinsy," *Proc. the 7th ISCA Tutorial and Research Workshop on Speech Synthesis*, pp. 211–216, 2010.

[3] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," *The Journal of the Acoustical Society of America*, 2013.

[4] 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. Interspeech2007, Singing Challenge*, 2007.

[5] T. Saitou, "Study on construction of singing voice synthesis system for elucidation of singing voice perception / generation mechanism," *Doctoral dissertation*, 1997.

[6] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robot Automatic Magazine*, vol. 19, pp. 98–100, 2012.

[7] P. F. MacNeilage, *The Production of Speech.* 1983.

[8] M. Alonso, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," *Master's thesis*, 2005. Online available: http://mtg.upf.edu/node/2223.

[9] G. Berndtsson, "The KTH rule system for singing synthesis," *STL-QPSR*, vol. 36, 1995.

[10] M. Umbert, J. Bonada, and M. Blaauw, "Generating singing voice expression contours based on unit selection," *Proc. Stockholm Music Acoustics Conference (SMAC)*, pp. 315–320, 2013.

[11] Y. Corporation, "VOCALOID - The modern singing synthesizer," 2014.

[12] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, "Pitch adaptive training for hmm-based singing voice synthesis," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5377–5380, 2012.

[13] H. Doi, T. Toda, T. Nakano, M. Goto, , and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," *Asia-Pacific Signal and Information Processing Association (APSIPA)*, pp. 1–6, 2012.

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

[15] J. M. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, and J. M. Pardo, "Emotional Speech Synthesis: From Speech Database to TTS," *International Conference on Spoken Language Processing*, pp. 235–248, 1995.

[16] Xue, "Emotional speech synthesis system based on a three-layered model using a dimensional approach," *Asia-Pacific Signal and Information Processing Association (APSIPA)*, pp. 505–514, 2015.

[17] M. Schrder, "Emotional speech synthesis: a review," *Proc: INTER-SPEECH*, pp. 561–564, 2001.

[18] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, pp. 1161–1178, 1980.

[19] Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song," *the 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS), Kingston, ON*, 2012.

[20] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using straight," *Proc. Interspeech2007, Singing Challenge*, 2007.

[21] P. B. Oncley, "Frequency, amplitude, and waveform modulation in the vocal vibrato," *The Journal of the Acoustical Society of America(JASA)*, p. 136, 1971.

[22] R. Elbarougy and M. Akagi, "Speech emotion recognition system based on a dimensional approach using a three-layered model," *Proc. Int. Conf. APSIPA ASC*, 2012.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Online available: http://www.praat.org.

[24] D. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

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

[26] M. MORISE, F. YOKOMORI, and K. OZAWA, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. E99, p. 1877, 2016.

[27] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, pp. 187–207, 1999.

[28] W. H. Press, B. P. Flannery, S. Teukolsky, and W. T. Vetterling., "Numerical Recipes in C," *Cambridge University Press*.

Publications

1. Nguyen Thi Hao, Masato Akagi, **"Synthesis of expressive singing voice by F0, amplitude envelope and spectral feature conversion,"** (*The 2018 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'18)*, Hawaii, US)

APPENDIX

Questionnaire for participant

Name:

Student ID:

Participant ID:

Gender:

Age:

Nationality:

Have you had any problem with hearing ability?

Have you ever known about Valence – Activation domain?

Have you ever had any special training about singing voice before?

How often do you listen to music?

*** Your personal data is only for this experiment and it will never be sent to others.**

Thank you very much for your cooperation!

Total amount of sessions: 16
This instruction file will be using for the 8 first sessions.

Thank you for your cooperation today for the experiment.

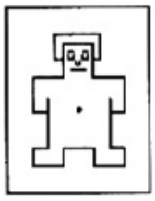
In this experiment:

- You will listen to the singing voices.
- You will be asked to answer the impression of the singing voice.

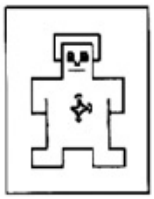
In this time, you will be required to evaluate the degree of Activation of the listened voice.

- The value of Activation is evaluated in a scale of 20 values (range -2 ~ 2 by 0.2 step).

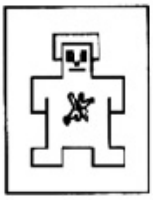
Degree of Activation:



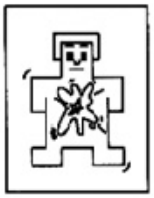
Very Calm
-2



Calm
-1



Neutral
0



Excited
1

Very Excited
2

Flow of experiment

- **Familiarize task**

You will be listen to some samples of singing voice and try to get familiar with this test.

1. When you are ready, please click “Start”.

2. A sound of singing voice will be played.

3. Listen to the sound and drag the slider to the proper position.

In each evaluation, you can listen to the sound repeatedly.

4: Click “Next” button to move to the next sample.

5: There are total 3 samples, therefore, this flow will repeat 3 times.

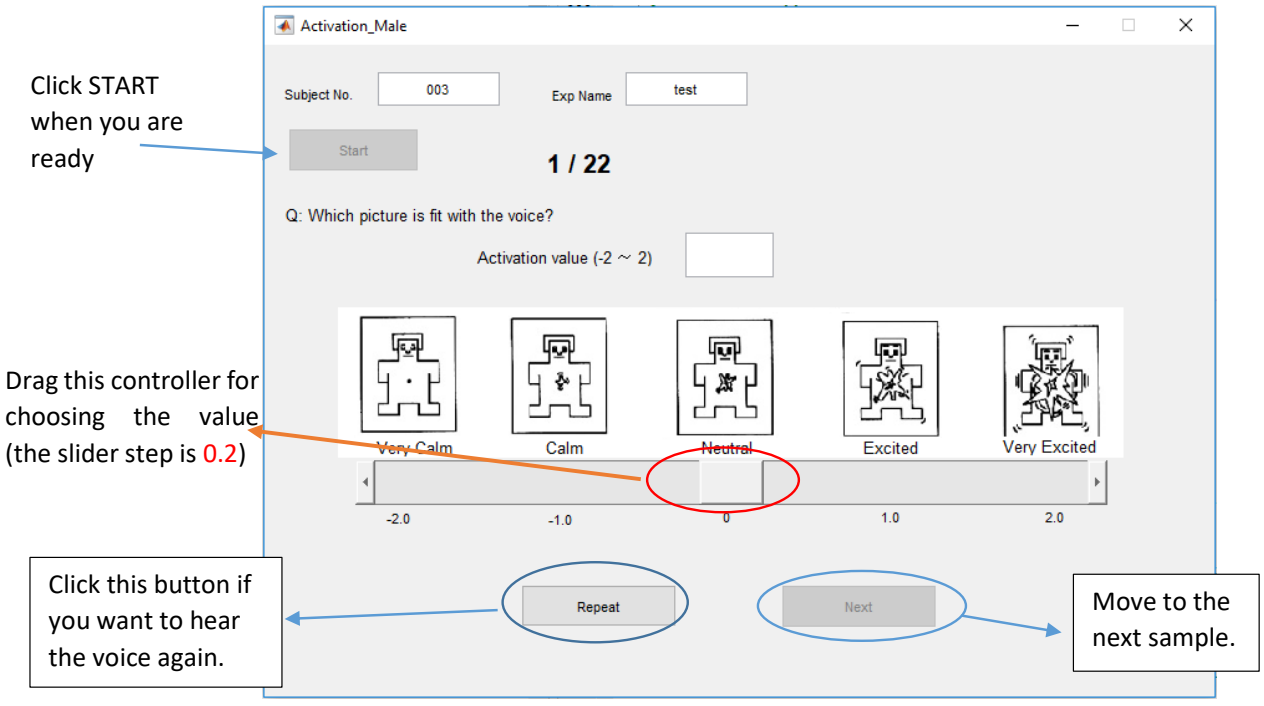
6: Please let me know when a window with "Finished" pops up.

After finish this familiarize task, please feel free to ask me any question about this test.

- **Main task**

Please follow the same procedure as familiarize task, but this time, the total sample is 22, so the flow will repeat 22 times.

GUI explanation:



* Caution:

- Do not press other buttons until one voice is over.

- Decision does not need to be consistent among evaluations. Please answer your impression of that time.

Total amount of sessions: 16
This instruction file will be using for the 8 last sessions.

Thank you for your cooperation today for the experiment.

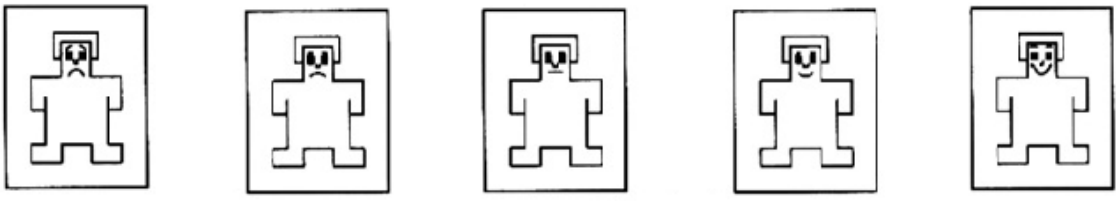
In this experiment:

- You will listen to the singing voices.
- You will be asked to answer the impression of the singing voice.

In this time, you will be required to evaluate the degree of Valence of the listened voice.

- The value of Valence is evaluated in a scale of 20 values (range -2 ~ 2 by 0.2 step).

Degree of Valence:



Very Negative	Negative	Neutral	Positive	Very Positive
-2	-1	0	1	2

Flow of experiment

- **Familiarize task**

You will be listen to some samples of singing voice and try to get familiar with this test.

1. When you are ready, please click “Start”.

2. A sound of singing voice will be played.

3. Listen to the sound and drag the slider to the proper position.

In each evaluation, you can listen to the sound repeatedly.

4: Click “Next” button to move to the next sample.

5: There are total 3 samples, therefore, this flow will repeat 3 times.

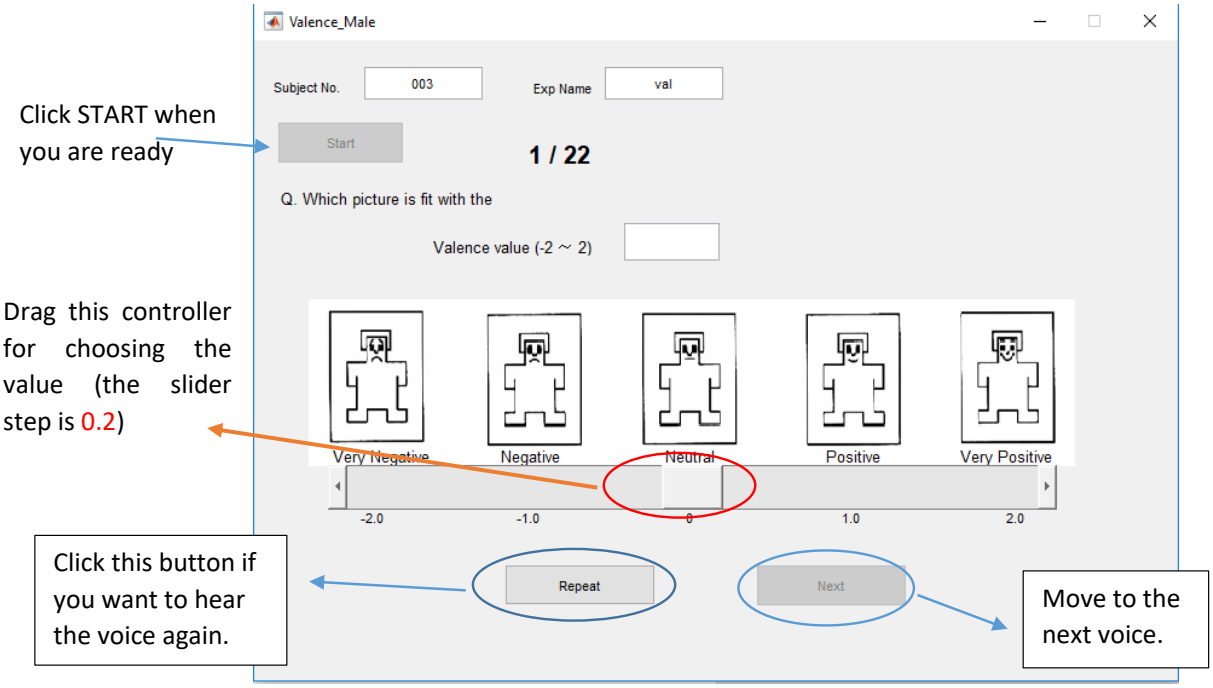
6: Please let me know when a window with "Finished" pops up.

After finish this familiarize task, please feel free to ask me any question about this test.

- **Main task**

Please follow the same procedure as familiarize task, but this time, the total sample is 22, so the flow will repeat 22 times.

GUI explanation:



* Caution:

- Do not press other buttons until one voice is over.

- Decision does not need to be consistent among evaluations. Please answer your impression of that time.