Master's Thesis

Study on Relations between Emotion Perception and Acoustic Features using Speech Morphing Techniques

1710034 Zi Wang

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)

February 2019

Abstract

Analyses and synthesis of emotional sounds is an exciting research direction. Moreover, a sophisticated emotional sound synthesis system can significantly improve experiences of human-computer communication. Although humans can express subtle emotional changes in their voices, most researches on emotional speech synthesis focus on the categorical approach to emotional states expressions, such as synthesizing speech to joy, sadness or anger. Besides categorical approach, some studies tried to control speech emotion continuously as humans do. As the study of Y. Xue has constructed an emotional speech conversion system using a rule-based approach and a three-layer model, following the emotional perception and production of human being. However, the system has rooms to improve in continuous emotion control, especially on Valence scale.

Whether categorical or continuous emotional speech synthesis, it is necessary to face a common problem, which researchers can only get categorical emotional voice data in the most case, and it is impossible for asking a human actor to record emotional voices data in a regular gradient variation, whether respecting physical acoustic features or emotional perception. Therefore, categorical data determines that many studies focus on categorical approaches. Even study as Xue's emotion conversion system, which focuses on continuous emotional speech synthesis, is trained by categorical data. Consequently, discontinuous training data had distorted the mapping rules between acoustic features and emotional impression to a certain extent. Also, limited and discontinuous training data makes studies as Xue's system fail to clarify the correspondence between some important acoustic features variations and emotion impression.

For the purpose to obtain emotional voices continuously spanned on the V-A space, discuss what acoustic features are important to emotional impressions and how those features relate to emotion perception in a more detailed way, this study has two sub-goals: (i) Obtaining emotional speech samples continuously spanned on the V-A space by morphing techniques and collect the impressions of synthesized voices. (ii) Examining how acoustic features related to perceptions of emotional speech. Therefore, this study interpolates voices from pairs of typical emotions with a morphing method, collects emotion scores on Arousal-Valence space by a listening test, and analyzes which acoustic features significantly influence emotion perception and how those features vary changes emotion impression.

Analyses based on acoustic features and evaluation scores show that Arousal perception can be stably described by merely using fundamental frequency (F0). Power related features have a significant influence on Arousal perception, however, limited on sad-related voices. Comparing to Arousal, this research found that F0 and formants significantly influence Valence perception simultaneously, and how acoustic features correspond to Valence perception vary with different morphing references. Considering the correspondence and significances vary across different acoustic features for different morphing groups, this study proposed an assumption that how acoustic features relate to Valence perception depends on different areas of V-A space, and it is necessary to manipulate formants-related features in order to obtain high quality of Valence control in synthesized emotional voices.

Keywords: morphing voices, emotional speeches, acoustic features, relation analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of the whole thesis, including research backgrounds, problems of previous researches, research aims, and the structure of this thesis.

1.1 Research Background

Analyses and synthesis of speech is an interesting research direction. Moreover, a highly sophisticated emotional sound synthesis system can significantly improve experiences of human-computer communication. Although, speech synthesis system is currently used in advanced applications such as text to speech systems, translation systems, or intelligent assistants, most of these systems are still only focused on text information and designed to generate natural sounding synthetic speech. However, only *linguistic information*, which a set of discrete symbols and their combination, cannot completely convey the intelligence in speech. Beside linguistic information, *paralinguistic information* and *nonlinguistic information* are also important to encompass information expressed by human speech [1]. The paralinguistic information means information can not infer from the written counterpart but is deliberately added by the speaker to modify or supplement the linguistic information, while the nonlinguistic information means those factors like the age, gender, idiosyncrasy, or emotional states of the speaker. Therefore, affective synthesized speech comprehend nonlinguistic information is increasingly required [2], and emotional speech synthesis can substantially contribute to the acoustic manifestation of the spoken language.

The most common method for emotional voice conversion is the categorical approach. Previous researches utilize Gaussian Mixture Model (GMM) [3] or deep Neural Network (DNN) [4] to synthesize emotional speech from an-

other. Comparing to prior studies converting voices to other simply categorized of emotions such as joy, anger, and sadness, studies by Tao *et al.* attempts to subtly synthesize speech by using "strong," "medium," and "weak" degrees [5].

However, emotions conveyed by humans are mild and not interrupted from one emotion to another, but can be described as a continuum of incessant states [6] [7]. Based on this idea, a rule-based voice conversion system for emotional speech is proposed by Xue *et al.* [8], in order to control the degree of emotion on dimensional space, which adopted to express emotions as points in dimensional space. Therefore, emotion with degrees can be described by changing the position in the emotion dimension continuously.

1.2 Problem

Although the categorical approach of emotional speech synthesis has its shortcomings, it is still used in lots of researches. Because it is impossible for asking a human actor to record emotional speech data in a regular gradient variation, whether respecting physical acoustic features or emotional perception; Unquestionably, researchers can allow the listeners to mark a large number of emotional voices on continuous axes by listening tests, in order to obtain an emotional voice database spanning with continuous dimensional representation; however, the listeners' perception of emotion may be affected by the personality of recorders of each voice. Besides, since the variation of acoustic features caused by personalities and emotional dissimilarities are difficult to distinguish, it creates obstacles to subsequent analysis. As a result, most studies can only get the categorical database, which leads to more researches focusing on categorical approaches.

This problem does not only exist in studies related to categorical emotion speech synthesis but a problem that is widely present in related research using categorical data. Even study as Xue's emotion conversion system, which focuses on continuous emotional synthesis voices, is trained by categorical data, i.e., Japanese Fujitsu database, which a professional female actress was asked for acting the uttered sentence with 5 categorical emotions involving joy, cold anger, hot anger, neutral, and sadness [8]. Figure 1.1 shows the evaluation scores of voices in Fujitsu database labeled from research of Li *et al.* [9], which 'Activation' is a synonymous term of 'Arousal'. Obviously, the emotional voices of training data are not continuously distributed on the V-A space. In studies focus on continuous emotion conversion system, this discontinuous training data had distorted the mapping rules between acoustic features and emotional impression to a certain extent. For example, although

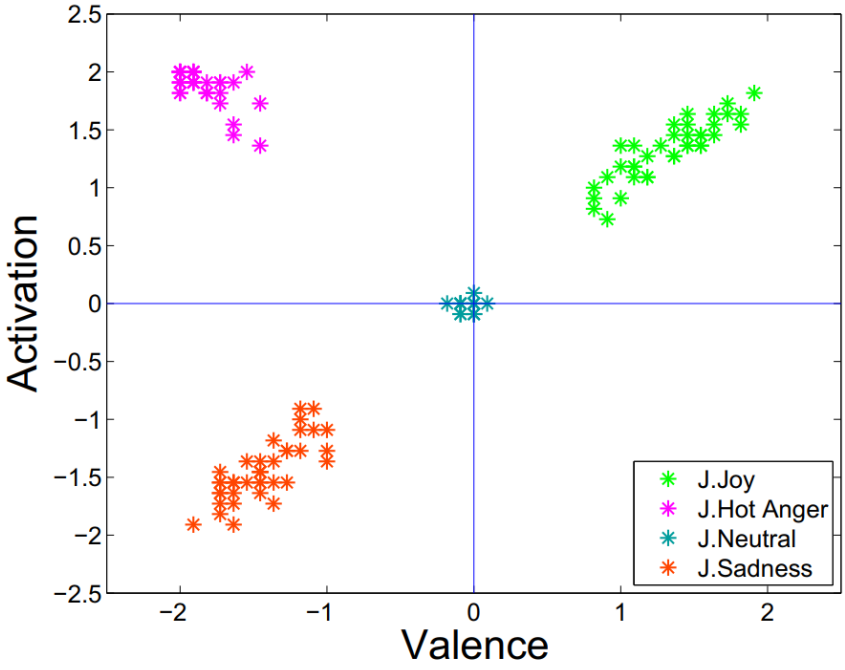


Figure 1.1: V-A evaluation scores of Fujitsu database [9]

the goal of Xue's system is to synthesize emotional voices anywhere in the V-A space, there exist some certain areas with any training sample on the V-A space (i.e., point (0.5,0.5) and its surrounding area as Fig. 1.1 shown). Moreover, limited training data also restrict researchers to explore the more corresponding relationship between acoustic features and emotional. This problem is reflected in the fact that previous studies have found that features like spectral sequences significantly influence on valence scale, by replacing the spectral sequences information of neutral voices with others emotional voices [10] [11] [8], limited and discontinuous training data makes study as Xue's system fail to clarify the correspondence between continuously spectral sequences variations and emotion impression. Therefore the system did not propose any spectral sequence modification model.

1.3 Research Aims

In the previous section, two problems related to emotion speech synthesis were pointed out. The first is that the emotional voice samples with continuous distribution cannot be obtained by the traditional recording method, and the second is that the corresponding relationships between acoustic fea-

tures and emotion impression training by categorical data are insufficient. Therefore, this study has two goals corresponded with those two problems as following.

Firstly, this study aims to obtain emotional speech samples by continuously interpolating the acoustic features between categorical reference voices. Then, this study is going to carry out the listening evaluation tests to verify whether synthesized morphed voices, which continuously distribute with acoustic features, are also continuously spanned on the V-A space. By synthesizing morphed voices that are continuously distributed with both acoustic features and dimensional emotion space, subsequent analyses are able to analyze how the emotional voices with different acoustic features located on the V-A space, especially those blank areas on the V-A space where no voice samples. Besides, since each reference voices used in morphing are recorded by one recorder (the topic of the corpus is put in Chapter 3, Section 3.1), the morphed voices can avoid the differences of acoustic features caused by personality for the most part. Therefore, this study can observe which and how feature variations correspond to alterations of emotion perception

Secondly, this study discusses which acoustic features are important to emotional impressions and how those features relate to emotion perception by achieving the first goal. Based on morphed voices and evaluation scores of listening test, this study extracts multiple acoustic features which are regarded as possibly related to emotion impress and examines how those acoustic features related to perceptions scores form listening test. Therefore, this study is able to discuss what acoustic features are important to emotional impressions and how those features relate to emotion perception. Those results can be used to adjust the existing modification rules of previous systems, or help researches to propose new modification models to construct a more complete emotional speech conversion system.

1.4 Structure of the Thesis

This thesis is going to be organized with following elements as the structure shown in Fig. 1.2:

- Chapter 1 gives an overview of emotional voice synthesis, discusses some challenges of emotional voice synthesis, and state the objectives of this research.

- Chapter 2 reviews the previous researches related to this topic, includes the emotional speech conversion system, morphing techniques, and the dimensional emotion representation.

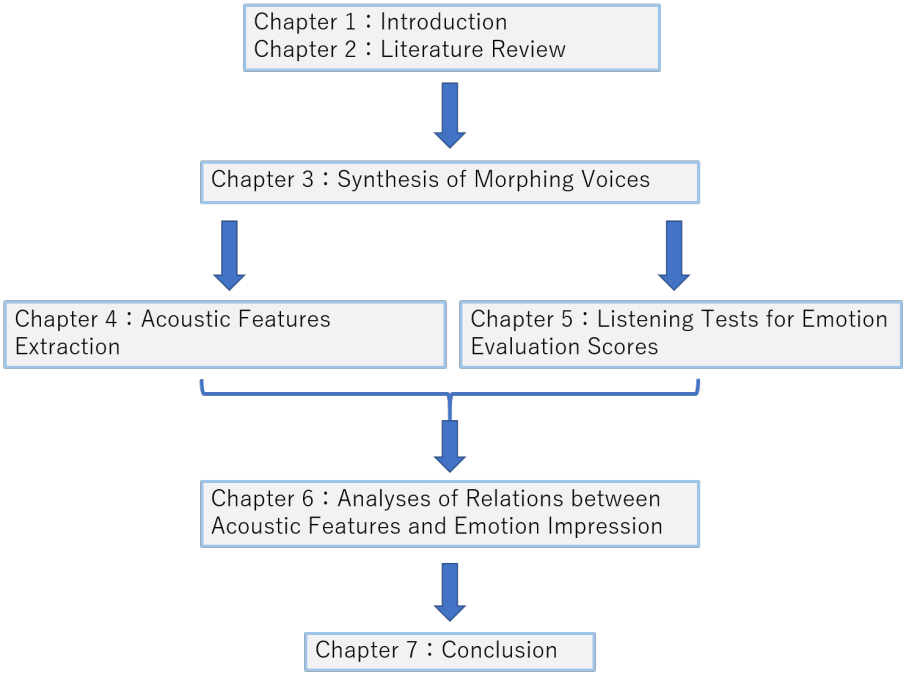


Figure 1.2: Structure of this Thesis

- Chapter 3 gives some details about how this research interpolates voices from reference emotional voices.
- Chapter 4 gives the acoustic features used in this study and how to extract those features.
- Chapter 5 introduces the listening test for collecting the emotion evaluation scores of morphed voices.
- Chapter 6 illustrates the investigation results of how different acoustic features influence emotional perception in different morphed voices and presents the discussion about those obtained results.
- Chapter 7 summaries the conclusions, contributions, and our remaining works of this study.

Chapter 2

Literature Review

This chapter provides a literature review about some important previous researches related to Emotional speech synthesis system. First, section 2.1 talk about some successful approaches in emotional voices synthesis area. Then, section 2.2 introduce what speech morphing techniques are and how to synthesize morphing voices using morphing toolbox. Finally, section 2.3 discuss what dimensional emotion representation is and how this study uses Valence-Arousal dimensional space to express emotion continuously. The above three sections are going to introduce the current status of emotional speech synthesis researches, then discuss the importance and necessity to use morphing techniques and dimensional emotion representation in this study.

2.1 Emotional Speech Synthesis System

This section provides a brief review about successful approaches in synthesizing emotional voices with some significant researches.

- **Unit selection approaches:** Unit selection approaches choice the units of variable size from recording database and then concatenate those unites in order to generate desired target utterance [12] [13]. This synthesis method often gives very natural results. However, the appropriate units of target utterance need to exist in the database; otherwise, the synthesized voices can be very bad [11]. Therefore, a highly sophisticated unit-selection synthesis system needs a vast database to cover all required prosodic, phonetic, and stylistic variations [14].

- **Statistical approaches:** Statistical approaches are widely studied in emotional voices synthesizing area, using statistical methods such as Hidden Markov Model (HMM) [12] [15], GMM [3], or DNN [4]

to model the important features from database. Those systems normally use the jointly model for some important parameter such as spectrum, F0, duration, *etc.* Figure 2.1 shows a typical structure of HMM-based voices synthesis system [15]. Comparing to the unit selection approaches, statistical approaches are more complex but general solutions, because statistical parametric synthesis systems do not require a complete database of any phonetic or prosodic contexts [14].

- **Rule-based approaches:** The rule-based method modifies the acoustic features of concatenating synthetic speeches or neutral speeches with rule-based simulation, and output others emotional speeches [16] [17]. Like Montero *et al.* successfully synthesize Spanish with three basic emotions (hot anger, happy, and sad), using the acoustic profile of global prosodic and voice quality parameter [18]. Compared with the previous two methods, although the rule-based method needs stand speeches as references, it can generate good emotional synthetic voices with smaller training data.

  However, those previous researches modified the related acoustic features separately, and there are only a few rule applications with basic emotions. The problems are modifying one acoustic feature influence other related features, and there is any suitable order for modification, and several basic categorical emotions are not enough for the sophisticated emotion impression as mentioned in section 1.1. For those reasons, Xue *et al.* proposed a rule-based emotional voice conversion system with degrees in dimensional emotion space shown in Fig. 2.2 [8]. This system adopted the Fujisaki model [1], STRAIGHT system [19], target prediction model [20] [21] *etc.* to modify related acoustic features with a more integrated way. Besides, Xue introduced a three-layered model for a dimensional approach, which a method help the system can generate more dynamic emotional speeches with different intensity.

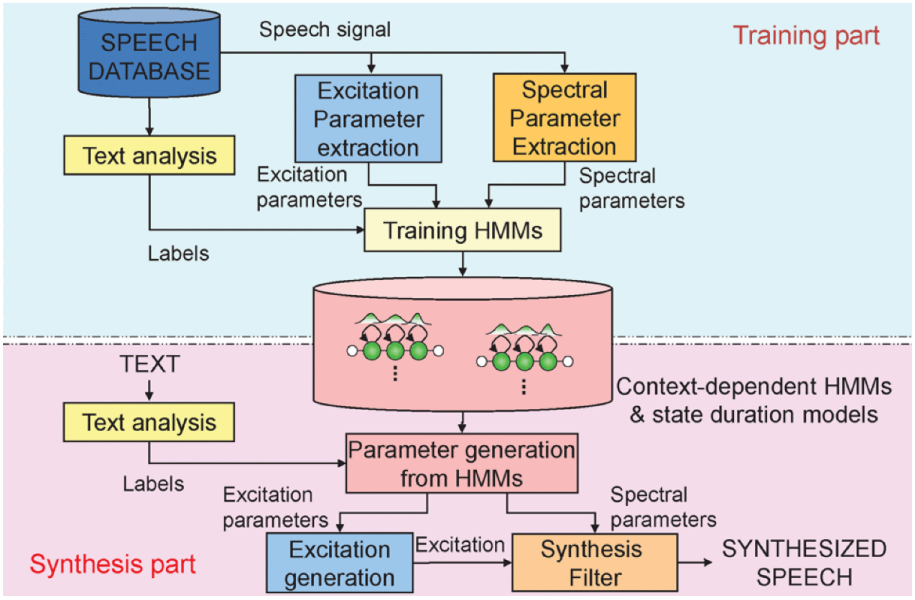


Figure 2.1: The HMM-based speech synthesis system overview [15]

2.2 Speech Morphing Techniques

As the problem mentioned in section 1.2, whether on the emotion or physics dimension, it's hard for human actors to record voice samples with a continuous gradient deliberately. Therefore, this research considers synthesizing morphing voices with gradual expression changes rather than human records. Based on TANDEM-STRAIGHT [22] [23], which a high preference speech analysis and modification framework, and morphing techniques [24], this research be able to synthesize morphing voices from pairs of typical emotion references. The morphing algorithm [25] based on TANDEM-STRAIGHT is implemented as a five-stage procedure. The first step is to extract parameters, include fundamental frequency (F0), aperiodicity spectrogram and interference-free spectrographic representation (STRAIGHT spectrogram) [22], of each reference utterance. The second step is to align parameters respecting to the time and frequency coordinates of two reference utterances. The third step is to interpolate or extrapolate parameters represented on the aligned time-frequency coordinates based on the given morphing rate(s). The fourth step is to deform the time-frequency coordinates with the given morphing rate(s). The final step is to resynthesize sound using the morphed parameters on the morphed time-frequency coordinate [26]. Listening experiments of previous researches show that naturalness of morphed voices was comparable to natural speech samples [27], which indicated that TANDEM-

STRAIGHT and morphing procedure enables stimulus continuum between different emotional expressions as a powerful tool for investigating the corresponding relationship between acoustic features and different emotions [28].

2.3 Valence - Activation Domain

In this research, a method for representing emotion with continuous degree is important. As mentioned in section 1.1, a small number of certain emotions may not provide humans with a sufficient level of discrimination. Besides, categorical emotion representation lacks the information about intensity degree [6] [7].

Except for the categorical approach, Another emotion representation is mapping emotions as points to n-dimensional space. Figure 2.3 [29] shows a common dimensional emotion representation which the three dimensions space which includes Arousal (excited-calm), Valence (positive-negative), and Dominance (powerful-weak) axes.The names of those three dimensions vary in different literature (e.g., valence, energy, and dominance; evaluation, activity, and potency; and evaluation, activation, and power). Dimensional approach was long investigated by Russel [30] [31], and it was suggested that the valence and arousal are two fundamental dimensions of emotional representation. Considering that Dominance domain is used to distinguish Fear and Anger which two kinds of emotions related to power, and this research morphs voices between Neutral, Happy, Sad and Angry, therefore, this study decided to use the Valence-Arousal space to represent emotion as shown in Fig. 2.4. On the V-A space, neutral is close to the origin; the 1st, 2nd, and 3rd quadrants of V-A space are corresponding to the happy, angry and sad emotions.

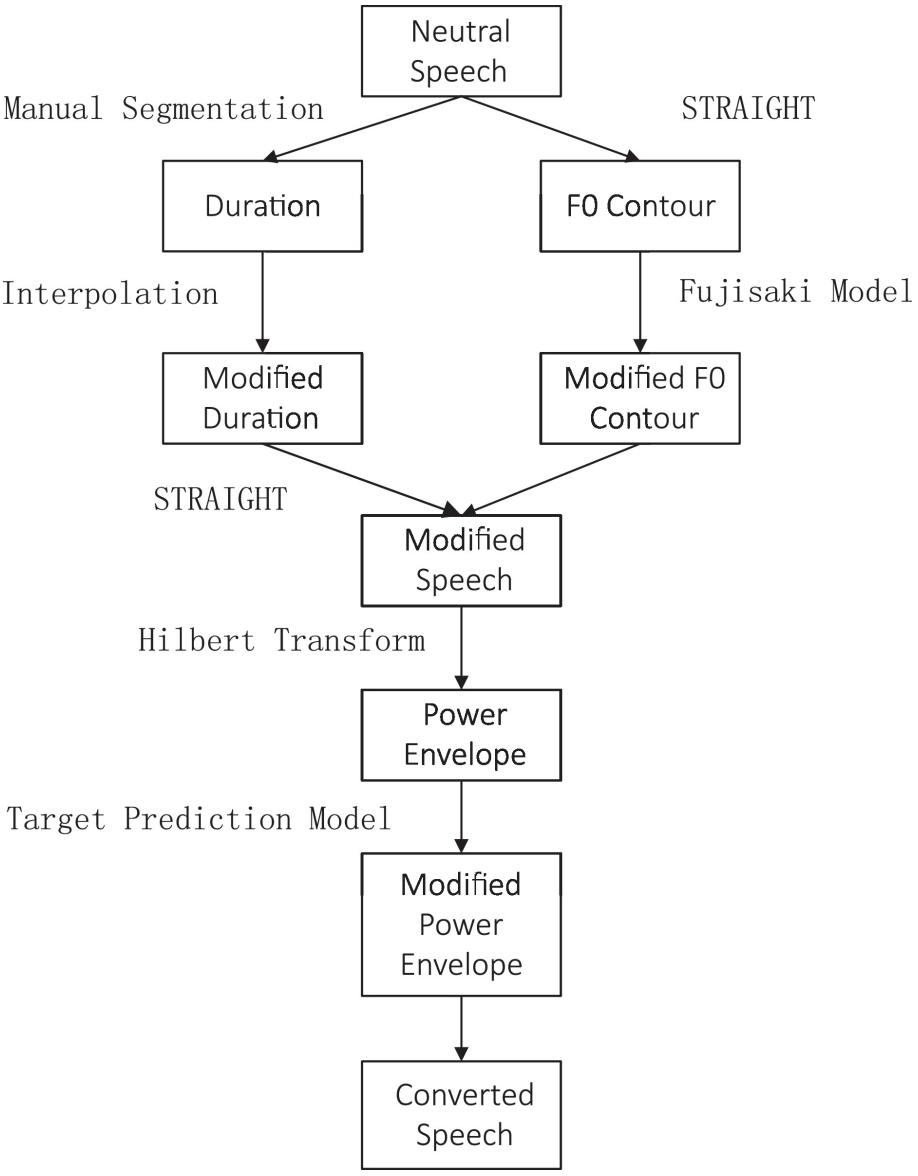


Figure 2.2: The modifying procedure of Xue's emotional voice conversion system [8]

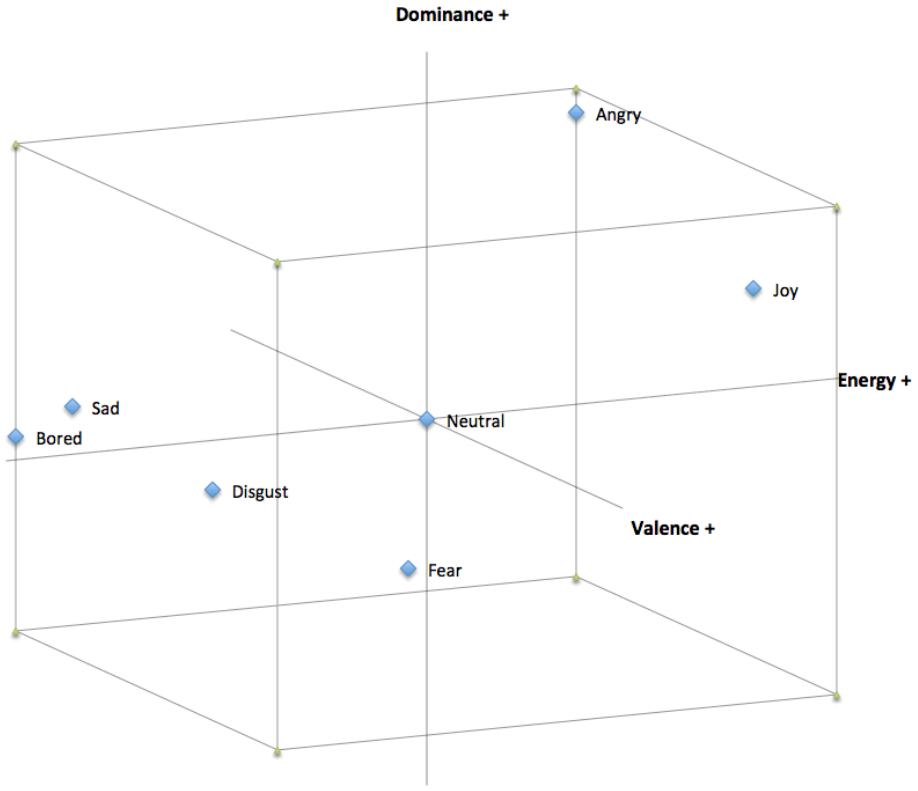


Figure 2.3: Three dimensions space for emotion representation [29]

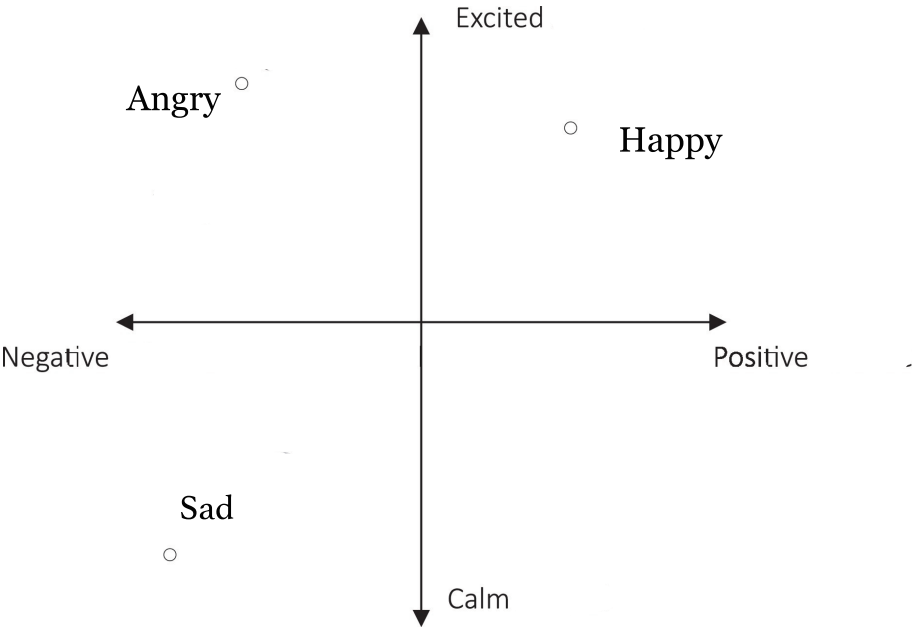


Figure 2.4: Valence-Arousal space

Chapter 3

Synthesis of Morphing Voices

This chapter describes the corpus of reference voices for the morphing process and the procedure to make morphing voices.

3.1 Corpus

The corpus for morphing voices was chosen from the Fujitsu database, which recorded by one professional female voice actress from Fujitsu Laboratory. The Fujitsu database contains 20 different Japanese sentences repeated in 5 different emotion categories: Neutral, Happy, Sad, Cold Anger, and Hot Anger. Those emotional voices are saved with sampling frequency 22050Hz and quantization of 16 bit. Considering the time cost and the number of participation of listening tests, also this study is aiming to synthesize morphed voices that are widely distributed in the V-A space as possible; This study chosen 10 sentences as the reference voices, including Neutral, Happy, Hot Angry, and Sad emotional voices which correspond to the origin, (Very Positive, Very Excited),(Very Negative, Very Excited), and (Very Negative, Very Clam) location on V-A space as Fig. 2.3 and 2.4 shown. In the following, this thesis refers to "Hot Angry " by "Angry". Table 3.1 lists the Japanese utterances and the English translations of reference voices.

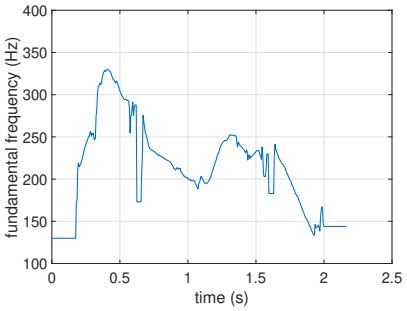
3.2 Morphing Processing

This section illustrates how this study interpolates reference voices to get morphing voices in details. Before applying morphing techniques, it is necessary to decompose each reference voice into three terms, which fundamental frequency (F0), aperiodicity spectrogram, and interference-free spectrographic representation (STRAIGHT spectrogram) based on TANDEM-

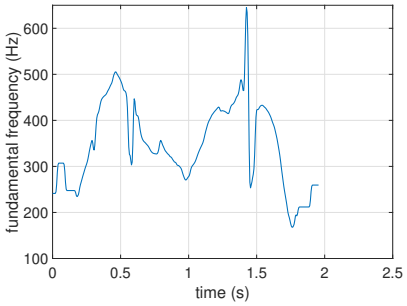
Table 3.1: Lists of sentences of reference voices from Fujitsu database, and translated version in English

	Japanese Sentence	English Translation
1	Atarashi meru ga todoite imasu.	You have got a new mail.
2	Machiawase wa Aoyama rashin desu.	I heard that we would meet in Aoyama.
3	Atarashi kuruma o kaimashita.	I bought a new car.
4	Sonna no furui meishindesu yo.	Thats an old superstition.
5	Minna kara eru ga okuraretan desu.	Many people sent me cheers.
6	Watashi no tokoro ni wa todoite imasu.	I have received it.
7	Arigato wa iimasen.	I will not say thanks.
8	Hanabi o miru noni goza ga irimasu ka.	Do we need a straw mat to watch fireworks.
9	Mo shinai to itta janaidesu ka.	I had told you don not do it again.
10	Jikandori ni konai wake o oshiete kudasai.	Tell me the reason why you don ' t come on time, please.

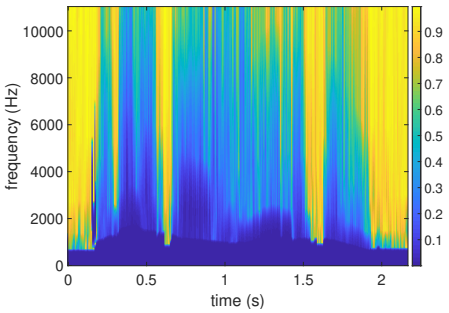
STRAIGHT system [22], [26], e.g., Fig. 3.1 shows important acoustic information be decomposed of two reference voices, which a Neutral reference and a Happy reference, calculated every 5 ms. After STRAIGHT system extracted necessary acoustic information, the temporal anchoring points of phonetic segments were manually located on the extracted STRAIGHT spectrograms as Fig. 3.2 shows. Based on those information and system setting, the morphing system is able to interpolate acoustic parameters respecting to the deformed time and frequency axes, with the given morphing rate(s), then resynthesize sounds using the morphed parameters on the reconstructed time-frequency coordinate. In this research, each morphing rate is set between 0 and 1, indicating that the morphed voices are getting closer and closer from one reference to another. Figures 3.3 and 3.4 illustrates the waveforms of a set of morphed voices, in those morphed audio waves gradually approach the happy reference from the Neutral reference. This thesis is going to discuss the acoustic features of morphed voices in details at section 4.2.



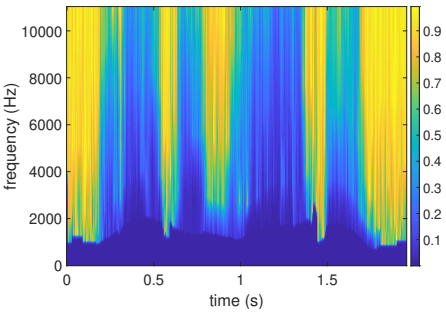
(a) F0 - Neutral Reference



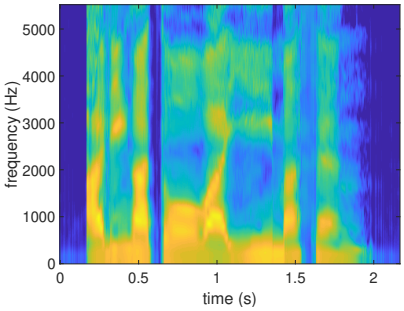
(b) F0 - Happy Reference



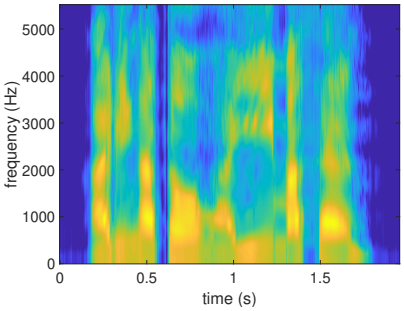
(c) Aperiodicity Spectrogram - Neutral Reference



(d) Aperiodicity Spectrogram - Happy Reference



(e) STRAIGHT Spectrogram - Neutral Reference



(f) STRAIGHT Spectrogram - Happy Reference

Figure 3.1: Extraction of acoustic features with TANDEM-STRAIGHT

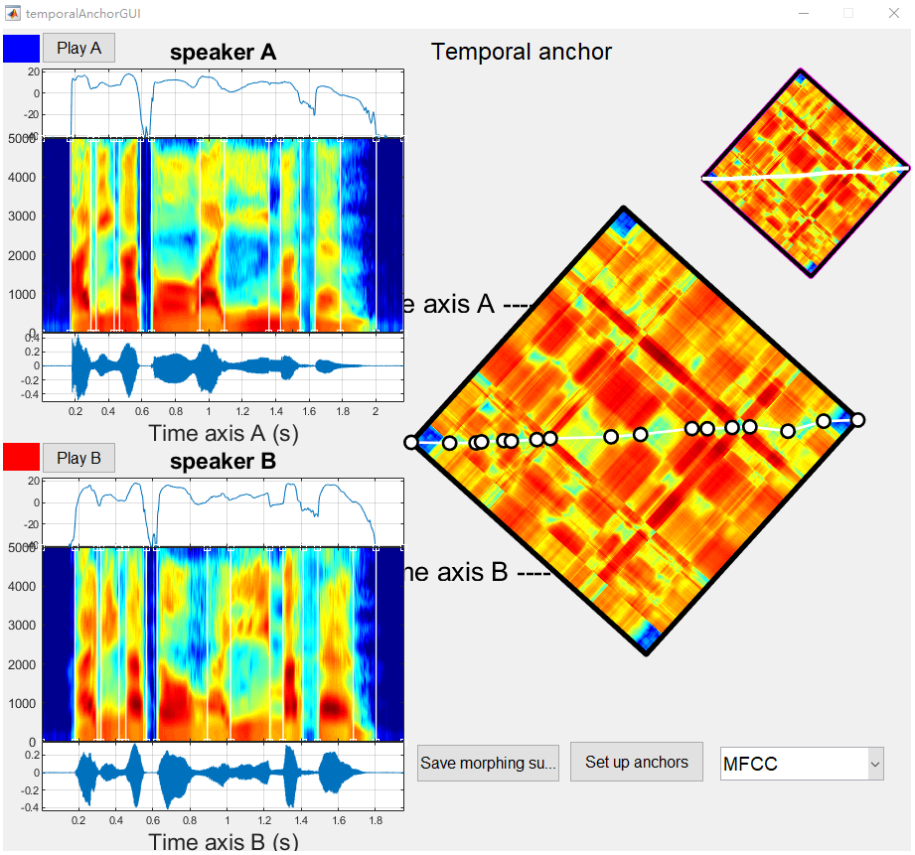
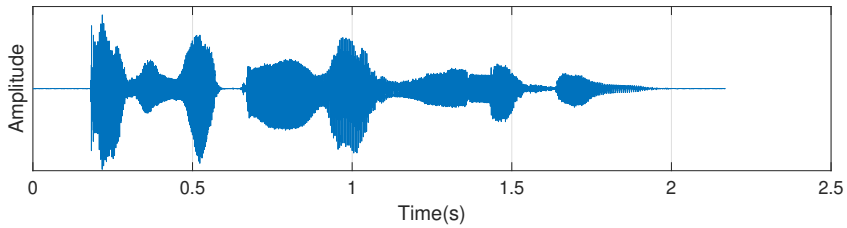
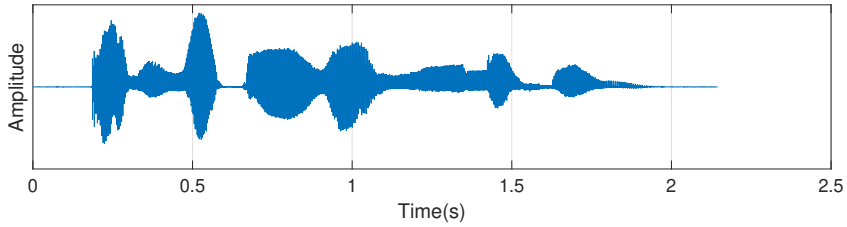


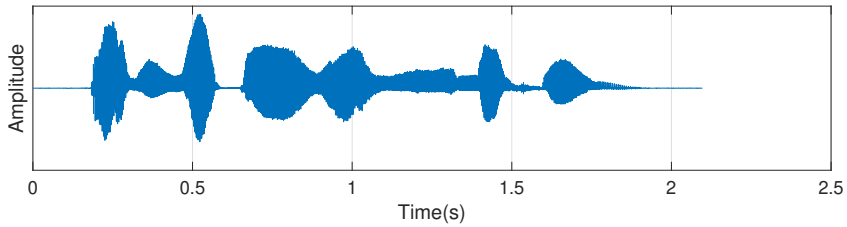
Figure 3.2: Setting temporal anchoring points



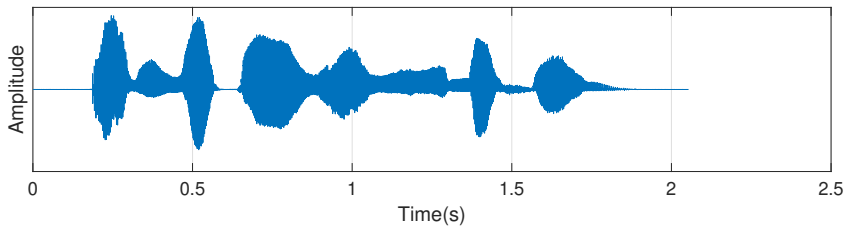
(a) Neutral Reference



(b) Morphed Voice 1

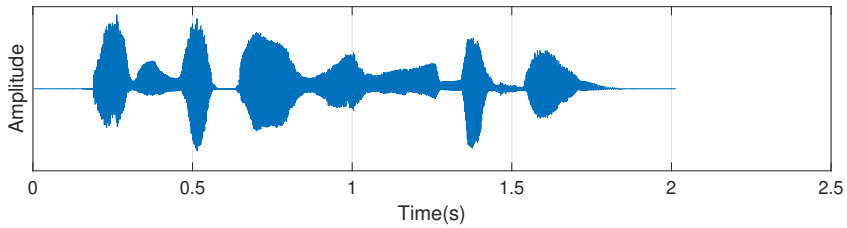


(c) Morphed Voice 2

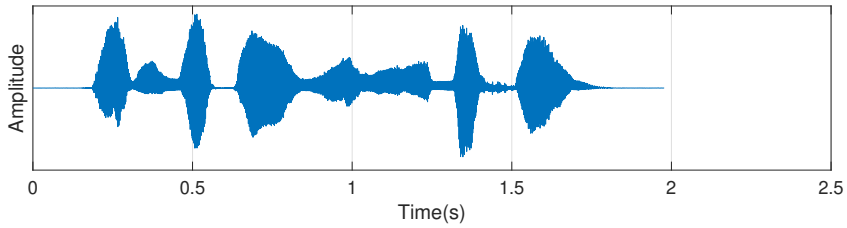


(d) Morphed Voice 3

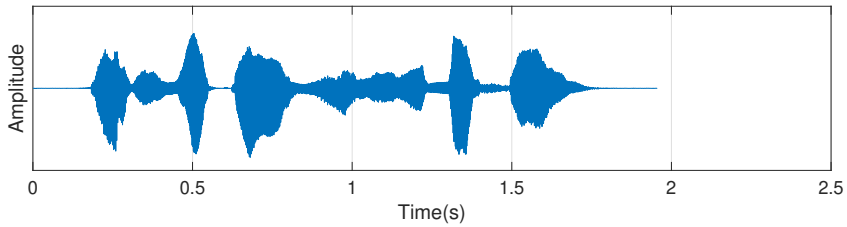
Figure 3.3: Waveform of Morphing Voice Sequences (1)



(a) Morphed Voice 4



(b) Morphed Voice 5



(c) Happy Reference

Figure 3.4: Waveform of Morphing Voice Sequences (2)

Chapter 4

Acoustic Features Extraction

In this chapter discusses two topics related to acoustic features. At first, several acoustic features are determined in section 4.1, which features may have relationships with emotion perception based on related researches. Then, illustrating how this study extracts those features in the same section. Based on extracted features, this study discusses the characteristics of the morphed voices in acoustic features, verify whether morphing processes, mentioned in section 3.2, successfully generated morphed voices with equidistantly changing acoustic features.

4.1 Acoustic Feature Extraction

It had been proved that the emotional speech perception varies respecting the acoustic features [2]. Accordance with the previous works [10], [8], F0, power, spectrum, and duration related features significantly impact on emotion perception. Also, considering that Japanese is generally regarded as a notion pitch-accent system and importance of accent components for emotion perception are emphasized by previous reports [10], [8], [1], this study separates each voice according to the criteria of accentual phrases and obverse features in detail. The accentual structure of each sentence is listed in table 4.1, which # represents the accentual boundary. Figure 4.1 shows an F0 contour and split accentual phrases of a voice sample with the sentence *'Atarashi# meru ga# todoite imasu.'*. The followings are acoustic features used in this study. Except the time of accentual phrases are marked by manual segmentation; others acoustic features are obtained by multiple estimation methods [32] [33] [34] [35]. It is worth noting that, except the length of accentual phrase and the total length of voice (the end of the last accentual phrases minus the beginning of the first accentual phrases), this study

calculates F0, power, formants, and voice activity time related features if and only if the speech frame with a voice activity probability [36] greater or equal to 99%.

- **F0 features:** F0 contour, Mean value of F0 (AP), highest F0 (HP), lowest F0 (LP), rising slope to maximum F0 (RSP), and range of F0 (RP).
- **Power features:** Mean value of intensity (AI), range of intensity (RI), minimum value of Mel log power (LMP), and range of Mel log power(RMP).
- **Formants features:** The first three formants contours (F1, F2, F3), Mean value of the first three formants (AF1, AF2, AF3), maximum value of the first three formants (HF1, HF2, HF3), and minimum value of the first three formants (LF1, LF2, LF3).
- **Duration features:** Total length (TL), voice activation length (VAL).

Further, this study normalizes features of each voice by the neutral reference which has the same content as follows:

$$F_{\text{Normalized}} = \frac{F_{\text{original}} - F_{\text{Neutral}}}{F_{\text{Neutral}}} \tag{4.1}$$

F_{Neutral} ,F_{original}, and $F_{\text{Normalized}}$ mean the feature values of neutral reference voices, the feature values of normalizing targets, and the normalized feature values. By normalization, acoustic features of emotional voices are represented as relative degrees to the Neutral reference.

Table 4.1: Accentual structure of each sentence

	Accentual structure of each sentence
1	Atarashi# meru ga# todoite imasu.
2	Machiawase wa# Aoyama# rashin desu.
3	Atarashi# kuruma o# kaimashita.
4	Sonna no# furui# meishindesu yo.
5	Minna kara# eru ga# okuraretan desu.
6	Watashi no tokoro ni wa# todoite imasu.
7	Arigato wa# iimasen.
8	Hanabi o# miru noni# goza ga# irimasu ka.
9	Mo shinai to# itta janaidesu ka.
10	Jikandori ni# konai wake o# oshiete kudasai.

4.2 Acoustic Features of Morphed Voices

This section discusses how acoustic features of STRAIGHT-based morphed voices vary based on those extracted features mentioned in section 4.1.

Figure 4.2 illustrates a set of F0 contours variations of Neutral-Happy morphed voices, which sentence is *'Arigato wa iimasen.'*. In those F0 contours, The neutral reference voice has the lowest F0 contour, while The happy reference voice has the highest F0 contour F0 counters of morphed voices increase to a higher level as the voices get closer to happy emotion. Also, as morphed voices are gradually transformed from neutral to happy, the duration of voices is gradually shortened. Furthermore, table 4.2 and 4.3 list some other features of a group of morphed voices, which stimuli include sentence *'Atarashi meru ga todoite imasu.'*. The name A001, B001, G001, and H001 are corresponding to Neutral, Happy, Sad, and Angry reference voices. In contrast, names like AB001-*number* represent the morphed voices, and the *number* increasing as morphed voices are gradually transformed from neutral to happy. These data listed in table 4.2 and 4.3 indicate that not only F0, but other features are also varying between morphed voices almost equidistantly. Through the results show that acoustic features of morphed voices do equidistantly vary between pairs of reference voices, the next step is to verify how these synthesized voices are distributed across the V-A space.

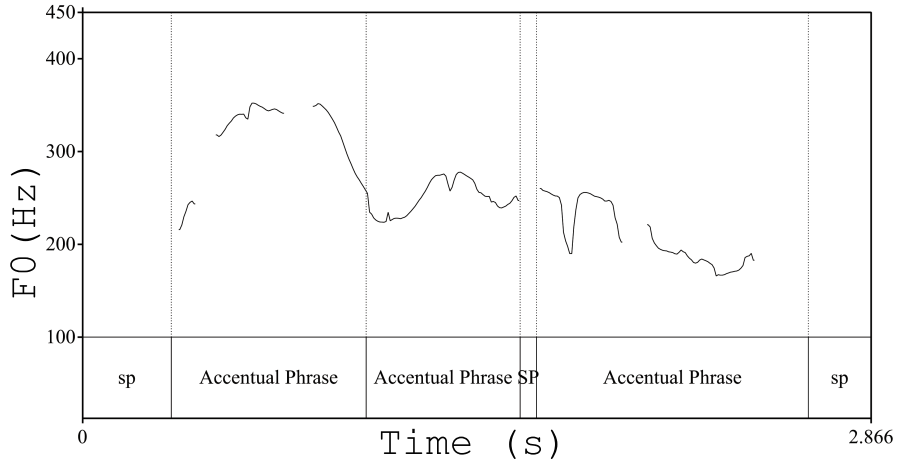


Figure 4.1: F0 contour and accentual phrases of a voice sample

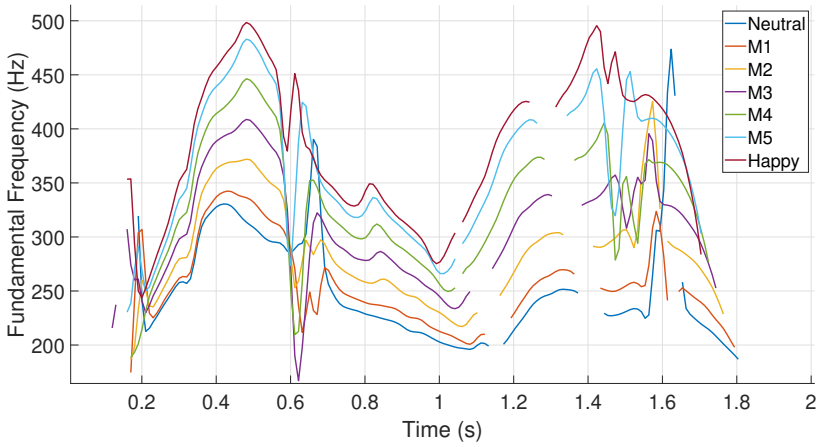


Figure 4.2: F0 contours of morphed voices (Neutral-Happy)

Table 4.2: Acoustic Features of Morphed Voices (1)

	Mean F0 (Hz)	Mean Intensity (dB)	Mean F1 (Hz)	Mean F2 (Hz)	Mean F3 (Hz)
A001	252.68	69.22	720.6	2005.25	3046.76
AB001-001	251.77	80.03	710.19	2003.97	3048.1
AB001-002	261.52	76.73	725.63	2001.84	3050.11
AB001-003	274.37	80.44	738	1993.76	3029.66
AB001-004	287.38	76.43	742.75	2007.92	3057.54
AB001-005	264.42	75.85	746.97	2012.86	3082.91
AB001-006	279.1	75.67	749.51	2012.29	3048.61
AB001-007	283.53	74.75	761.35	1973.5	3032.47
AB001-008	284.36	74.77	760.02	1984.49	3008.08
AB001-009	294.44	73.99	764.61	1976.07	2999.75
AB001-010	302.39	74	793.14	1992.91	3030.81
AB001-011	310.31	73.99	810.83	1962.47	2987.46
B001	312.2	71.18	817.5	1951.18	3011.59
G001	195.55	64.92	775.49	2060.48	3179.85
GB001-001	195.05	75	762.78	2051.14	3189.15
GB001-002	213.89	75.6	752.3	2062.38	3161.85
GB001-003	232.98	75.96	761.23	2065.94	3185.45
GB001-004	252.97	75.42	758.42	2057.98	3154.42
GB001-005	272.51	74.54	762	2069.77	3132.03
GB001-006	292.05	73.58	763.83	2072.39	3119.71
GB001-007	311.86	73.59	771.85	2037.45	3114.97
GB001-008	331.71	73.3	768.83	1994.11	3048.31
GB001-009	351.78	73.24	779.19	2022.31	3043.54
GB001-010	374.13	74.13	804.31	1967.32	3021.66

Table 4.3: Acoustic Features of Morphed Voices (2)

	Mean F0 (Hz)	Mean Intensity (dB)	Mean F1 (Hz)	Mean F2 (Hz)	Mean F3 (Hz)
GH001-001	196.2	75.03	759.3	2056.85	3188.2
GH001-002	208.11	81.34	781.76	2087.63	3202.43
GH001-003	220.95	75.87	743.04	2047.12	3179.35
GH001-004	234.05	74.52	768.25	2063.6	3195.46
GH001-005	247.14	72.55	749.01	2047.51	3191.93
GH001-006	260.24	72.19	734.45	2038.23	3183.38
GH001-007	274.1	72.07	769.81	2030.73	3167.5
GH001-008	287.07	72.49	762.42	2030.74	3160.91
GH001-009	303.09	75.7	771.04	2022.25	3163.26
GH001-010	310.57	73.44	782.63	2015.82	3162.93
H001	317.82	72.01	790.94	2022.49	3174.32
HA001-001	315.79	74.19	777.51	2019.25	3155.05
HA001-002	312.41	73.59	774.04	2025.88	3161.55
HA001-003	304.26	73.99	763.81	2033.66	3158.51
HA001-004	291.38	74.17	706.2	2022.87	3134.05
HA001-005	290.89	74.34	747.03	2013.01	3127.32
HA001-006	283.16	74.69	740.78	2027.61	3119.81
HA001-007	277.5	75.01	748.16	2034.64	3108.14
HA001-008	270.5	75.19	733.43	2029.44	3101.35
HA001-009	264.15	75.84	726.18	2031.6	3095.06
HA001-010	258.48	76.51	729.7	2044.55	3077.02
HA001-011	249.2	76.7	710.05	2008.57	3053.32
HB001-001	314.59	74.18	776.55	2013.24	3151.44
HB001-002	323.34	73.87	783.29	2023.75	3146.54
HB001-003	327.59	74.42	790.74	2014.43	3124.31
HB001-004	331.72	74.91	787.29	2012.44	3111.53
HB001-005	336.1	75.48	785.93	1997.6	3121.53
HB001-006	343.02	75.17	785.82	1997.32	3093.22
HB001-007	347.16	75.43	780.44	1992.33	3072.26
HB001-008	354.44	75.74	770.45	1963.63	3050.56
HB001-009	360.44	75.05	775.21	1976.43	3053
HB001-010	367.1	74.02	782.9	1973.72	3014.34
HB001-011	373.17	74.01	794.99	1965.45	2982.81

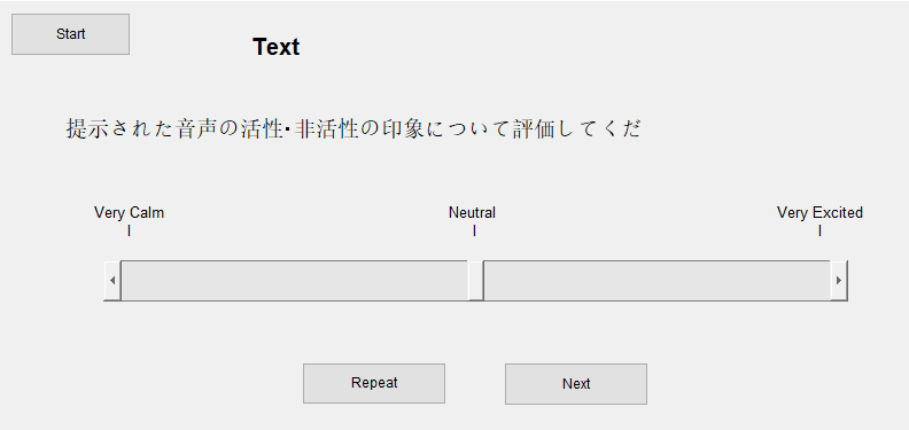
Chapter 5

Listening Tests for Emotion Evaluation Scores

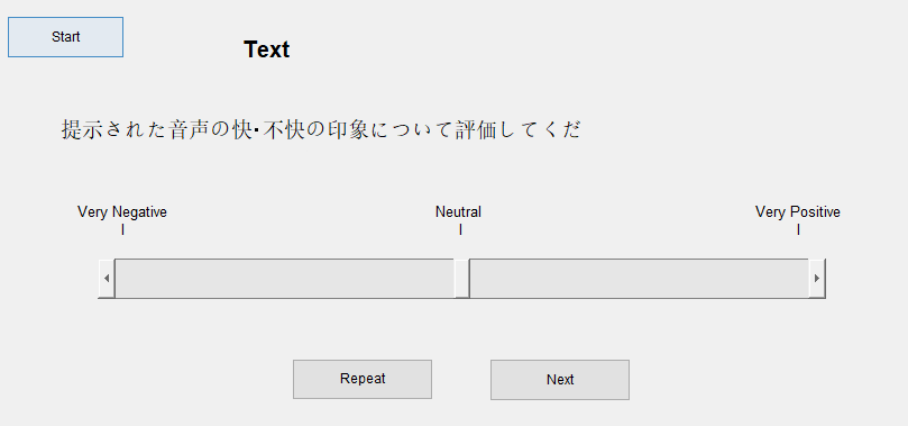
This chapter firstly discusses the listening tests carried out for collecting emotion perception of morphed voices in section 5.1. Then, section 5.2 are going to summarize the results of listening tests and discuss how morphed voices respect to acoustic features distribute on V-A space.

5.1 Listening Evaluation Tests

By finishing those morphing process mentioned in section 3.2, this research successfully collected morphed voices using interpolated acoustic features. However, it is necessary to verify whether the Valence and Arousal scores of morphed voices can be perceived in a continuous way. Therefore, this research carries a listening test out to collect evaluation results of the morphed voices with Valence and Arousal. Ten Japanese listeners with normal-hearing, aged from 22 to 27, participated in the listening test of Valence and Arousal separately for 570 stimuli, including 40 references and 530 morphed voices. The listeners were asked to evaluate Valence (very negative to very positive) and Arousal (very calm to very excite) of the stimuli by a graphic user interface (GUI), as Fig. 5.1 show, in a soundproof chamber. The audio equipment used in the experiment was calibrated to play a white noise file with the sound pressure of 64 dB. Each test plays stimuli in a random order. Although in our calculation, the *neutral*, *very positive* (*very excited*), and *very negative* (*very clam*) on the slider bar are corresponding to values 0, 2, and -2, listeners won't get any numerical tips, considering that listeners may unconsciously cluster and classify stimuli based on numerical information. In contrast, listeners can only evaluate the stimuli by the relative position of



(a) Arousal GUI



(b) Valence GUI

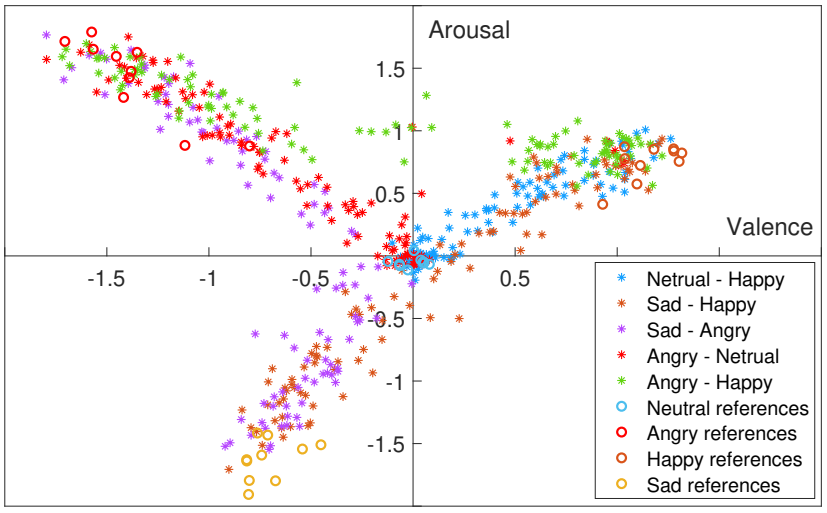
Figure 5.1: Arousal and Valence evaluation GUI

the slider on the bar and their subjective feelings.

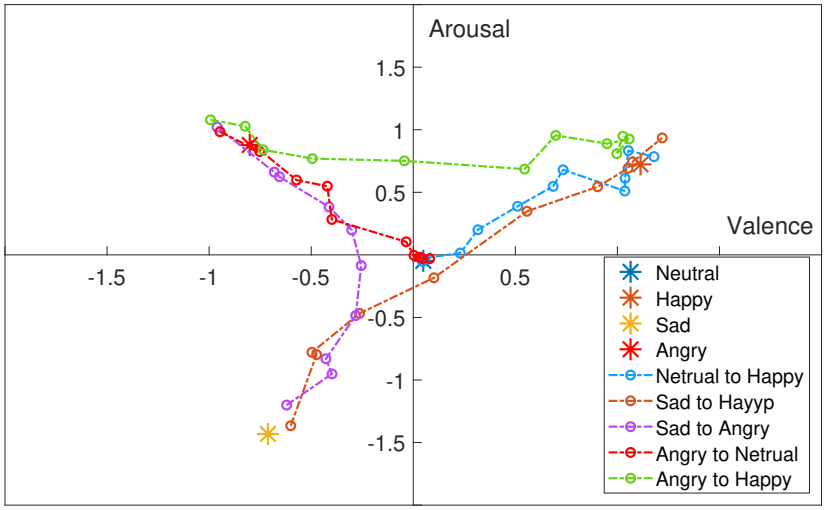
5.2 Evaluation Results

The mean values of all listeners' evaluated V-A scores are shown in Fig. 5.2a, while Fig. 5.2b illustrates scores of a set of morphed voices with the same sentence*' Atarashi meru ga todoite imasu.'*. Those two figures suggest that the morphing techniques successfully synthesize voices between neutral to other emotional categories with continuous distribution. However, the morphed voices between Angry and Happy (green symbols and lines) did not vary smoothly and suddenly changed from the second quadrant to the first quadrant. Also, some morphed voices between Sad and Angry (purple symbols

and lines) are evaluated close to neutral. It is worth noting that although the acoustic features are nearly equidistantly varying between morphed voices as mentioned in section 4.2, the evaluated V-A scores do not change as equidistant as the acoustic features. Overall, evaluation scores generally maintain a monotonous change between reference voices as Fig. 5.2b shown, but it is noteworthy that the evaluation scores of morphed voices are gathered around the reference voices, and sparsely distributed in areas between the references voices. This pattern is crucial for determining the corresponding relationships between acoustic features and evaluation scores and is going to be discussed in details at chapter 6. Those results of listening tests presented in this section show that the synthesized morphed voices do monotonously and continuously vary between the positions of reference voices on the V-A space, although the direction and magnitude of the emotional variations are not as uniform as the acoustic features variations as discussed in section 4.2.



(a) Evaluation scores of all stimuli



(b) Evaluation scores of one group

Figure 5.2: Evaluation scores

Chapter 6

Analyses of Relations between Acoustic Features and Emotion Impression

Based on the extracted acoustic features and the collected emotional evaluation scores, which were discussed in chapters 4 and 5, this study discusses the analysis results of how acoustic features relate to those emotion evaluations in this chapter. Those analyses are aiming to examine what and how acoustic features influence on perceptions of emotional speech. Firstly, section 6.1 is going to discuss what acoustic features significantly influence Arousal perception, then section 6.2 changes the topic to Valence perception. Finally, section 6.3 summarize all those analysis results and discuss how acoustic features influence emotion perceptions based on those results.

6.1 Arousal

This research found that the fundamental frequency (F0) related features are most significantly influence Arousal perception and can fit Arousal evaluation scores well, regardless of morphing references. Figure 6.1 shows that Arousal scores stalely increase (excited direction) as the mean value of F0 (AP) increases, which the most significant feature for Arousal. The fitting curve in Fig. 6.1 is a 3-degree polynomial. In most instances, this research fits features and evaluation scores with the 3-degree polynomial. The reason for using 3-degree polynomial is because the emotional scores remain stable near the stationary points of the cubic fitting function, where reference voices located, and changes rapidly between the stationary points as mentioned in section 5.2, comparing that acoustic features are changing equidistantly as

mentioned in section 4.2. Those corresponding relationships are identical to the shape characteristics of cubic functions and can get a relatively low root-mean-square deviation (RMSD) for regression. Not only AP, the max value of F0 (HP), as Fig. 6.2 shows, the max value of F0 in the first accentual phrase (HP_FAP), or the F0 range can also fit arousal scores very well. Further analyses show that the range of Mel log-power (RMP) and the range of intensity features are significant for sad-related morphed voices. Figure 6.3 illustrate that the span of RMP in Sad-Happy and Sad-Angry morphing groups are relatively larger, and Arousal scores of those two groups increase with increasing RMP. However, the Arousal scores did not significantly correspond to power related features in other morphing groups. This phenomenon shows that on the power-related features, sad is significantly lower than other categories, but there is no obvious difference in others stimuli other than sad. Since the Angry-Happy voices have almost the same scores on Arousal axis, which those green symbols and lines in Fig. 5.2a and 5.2b, so this group cannot be well fitted.

6.2 Valence

However, relations between acoustic features and Valence perception are much more complicated, which corresponding relationships vary in different morphing groups. Figure 6.4 shows how AP feature relates to Valence perception. For Neutral-Happy and Sad-Happy voices (Fig. 6.4a and Fig 6.4c), increasing F0 makes stimuli sound positive. In contrast, increasing F0 gives stimuli sound negative for Neutral-Angry (Fig. 6.4b). A remarkable phenomenon is for Sad-Angry voices (Fig. 6.4d), a certain level of F0 gives stimuli a neutral feeling, but F0 above or below this level makes stimuli sound negative, this result explains why some morphed sounds between Sad-Happy are evaluated as neutral, which the pattern had been mentioned in section 5.2.

Besides F0, more analyses illustrate that Valence perception is at least affected by F0 and formants features simultaneously, and not a single formant component can stably describe Valence perception. Figure 6.5 indicates that the mean value of the first formant (AF1) can fit Valence scores well for Neural-Happy group and Neutral-Angry group (Fig. 6.5a and 6.5b), but, feature AF1 has a weaker interpretation of Sad-Happy group(Fig. 6.5c). In contrast, the interpretation ability of the mean value of the third formant (AF3) are weaker than AF1 for Neutral-Happy group and Neutral-Angry group (Fig. 6.6a and 6.6b), but stronger for Sad-Happy group (Fig. 6.6c). Section 5.2 mentioned that the morphed voices between Angry and Happy

(those green symbols and lines in Fig 5.2a and 5.2b) did not vary smoothly but suddenly changed from the second quadrant to the first quadrant. For this phenomenon, Figure 6.7a illustrates that both positive (plus Valence score) and negative (minus Valence score) evaluated stimuli in Angry-Happy group has a relatively high level of F0 (1.2 times or more of Neutral reference). However, formant related features, especially feature AF3 can significantly fit Valence scores for Angry-Happy voices, as Fig 6.7b shows.

6.3 Discussion

As this thesis mentioned in section 1.3, the second goal of this study is going to discusses which acoustic features are important to emotional impressions and how those features relate to emotion perception. Based on the analysis results discussed in section 6.1, this study ascertains that Arousal perception can be stably described by merely using F0 related features. Power related features have a significant influence on Arousal perception while limited on sad-related voices. Also, on Arousal axis, acoustic features and Arousal scores present a common, monotonous corresponding relationship.

Comparing to Arousal, analysis results discussed in section 6.2 indicates that the corresponding relationships between acoustic features and Valence perception are more complicated. On Valence axis, acoustic features and Valence scores show a non-uniform corresponding relationship. Although F0 still has significant effects on Valence perception, the Valence perception should be simultaneously influenced by F0 and different formant components; Also, which formant components are important and how the acoustic features correspond to emotion perception are dependent on the different areas of V-A space (the different morphing references).

In section 4.1, it has been mentioned that this research extracted the acoustic features in specific accentual phrases. Table 6.1 and 6.2 list the RMSD values of global and accentual AP, AF1,and AF3 features. The variable names of the suffix 'FAP ' represent the features of the first accentual phrases while the variable names of the suffix 'EAP ' represent the features of the last accentual phrases. As can be seen from the table, global features and accentual features have the same or slightly higher level of error for fitting Valence and Arousal scores. Also, fitting results of global or accentual features are similar as Fig. 6.8 shows.

Based on the results listed above, the second sub-goal of this study had been achieved. This study observed the correspondence between acoustic features and emotion perception in more detail comparing to the previous study. In particular, formants related features, which features that have

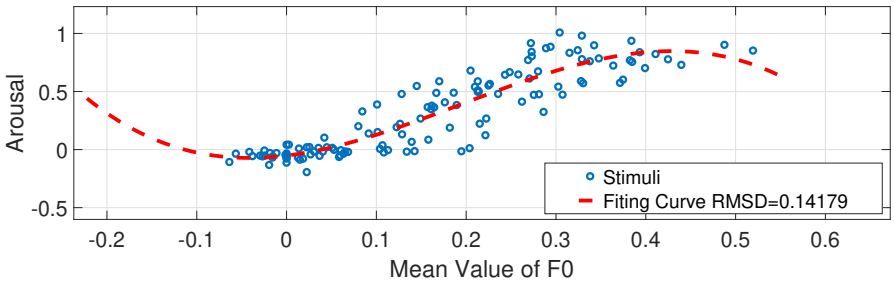
not been modified by the previous system, have the complex and significant influences on Valence perception.

Table 6.1: Fitting Error of Global and Accentual Features (Arousal)

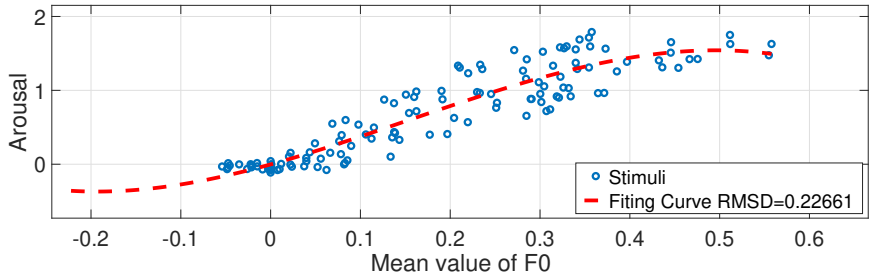
	Neutral-Happy	Neutral-Angry	Sad-Happy	Sad-Angry	Angry-Happy
AP	0.142	0.227	0.264	0.36	0.307
AP_FAP	0.2	0.361	0.375	0.477	0.32
AP_EAP	0.165	0.267	0.268	0.401	0.32
AF1	0.261	0.352	0.854	1.03	0.271
AF1_FAP	0.3	0.34	0.834	0.859	0.275
AF1_EAP	0.278	0.507	0.844	1.094	0.292
AF3	0.283	0.495	0.539	1.110	0.241
AF3_FAP	0.315	0.505	0.666	1.128	0.278
AF3_EAP	0.248	0.504	0.486	1.078	0.24

Table 6.2: Fitting Error of Global and Accentual Features (Valneve)

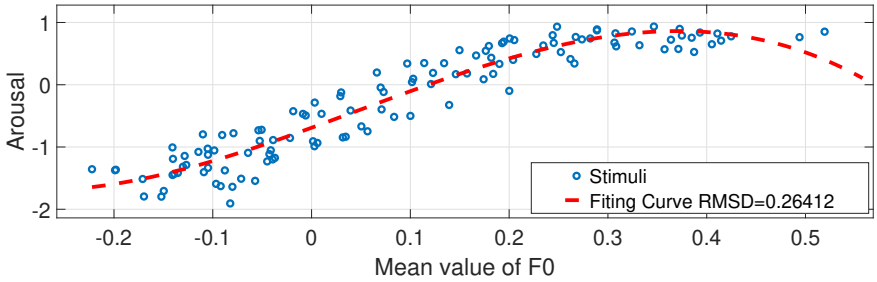
	Neutral-Happy	Neutral-Angry	Sad-Happy	Sad-Angry	Angry-Happy
AP	0.178	0.473	0.195	0.219	1.015
AP_FAP	0.248	0.547	0.318	0.295	1.048
AP_EAP	0.235	0.332	0.244	0.294	1.047
AF1	0.34	0.434	0.692	0.349	0.904
AF1_FAP	0.375	0.41	0.677	0.31	0.923
AF1_EAP	0.367	0.539	0.672	0.377	0.973
AF3	0.374	0.487	0.455	0.392	0.729
AF3_FAP	0.395	0.507	0.556	0.407	0.86
AF3_EAP	0.336	0.504	0.407	0.385	0.707



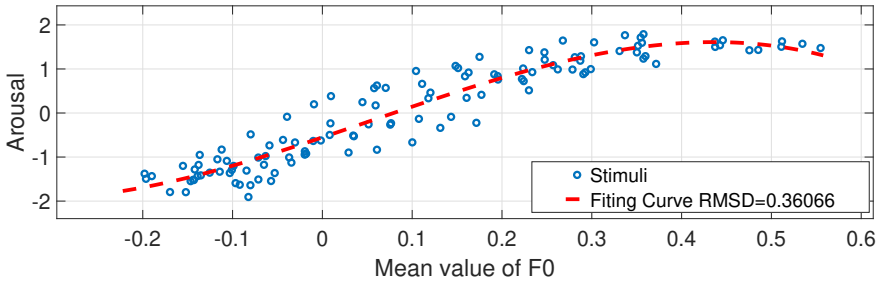
(a) Neutral-Happy voices



(b) Neutral-Angry voices

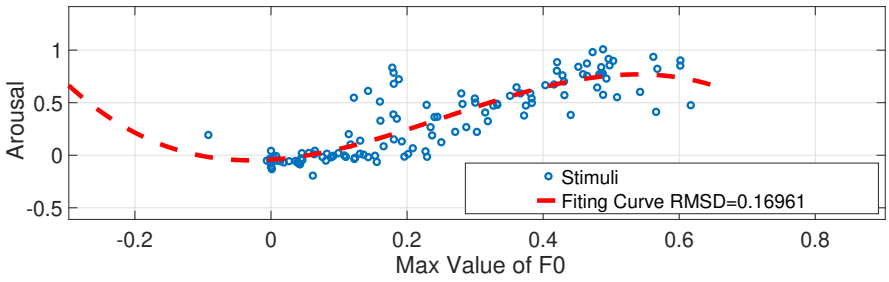


(c) Sad-Happy Voices

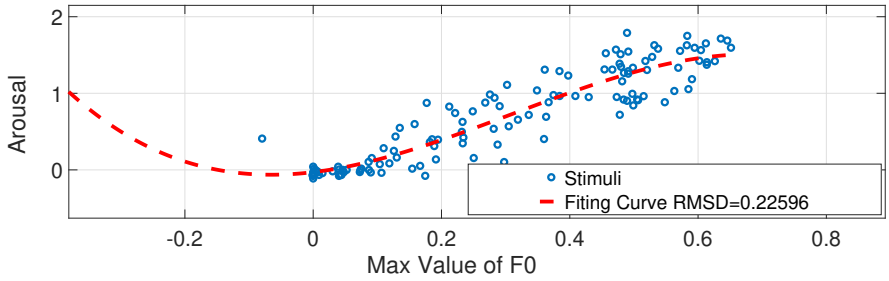


(d) Sad-Angry voices

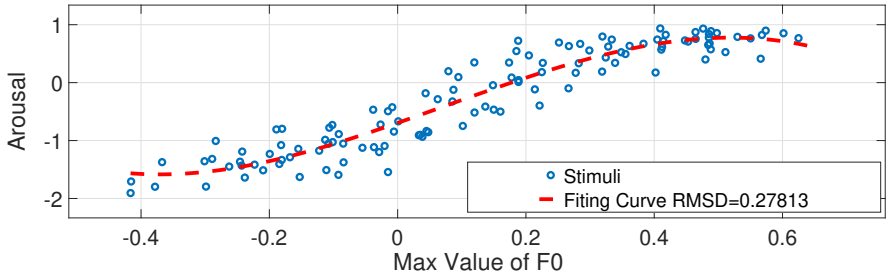
Figure 6.1: Fitting Arousal using AP feature



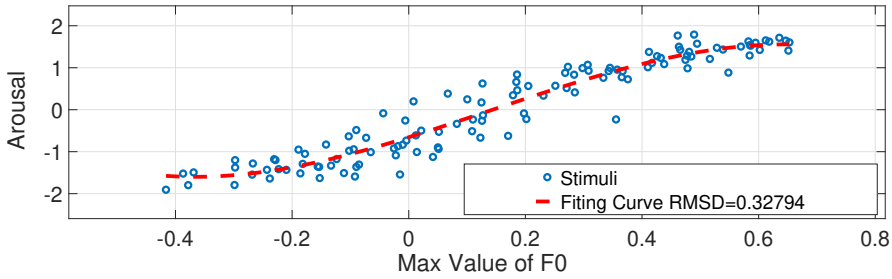
(a) Neutral-Happy voices



(b) Neutral-Angry voices

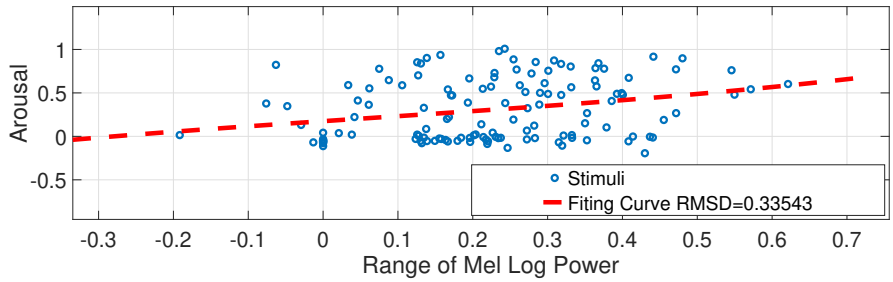


(c) Sad-Happy Voices

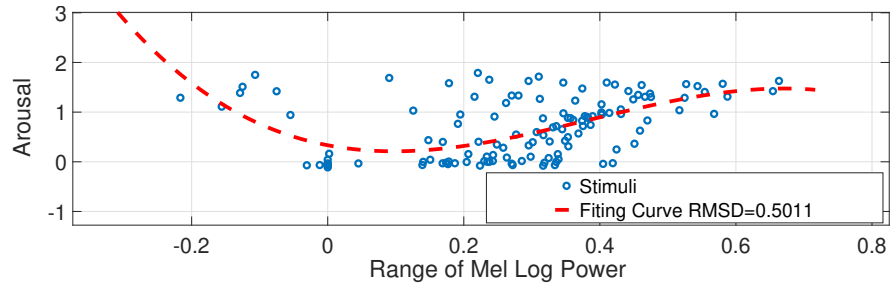


(d) Sad-Angry voices

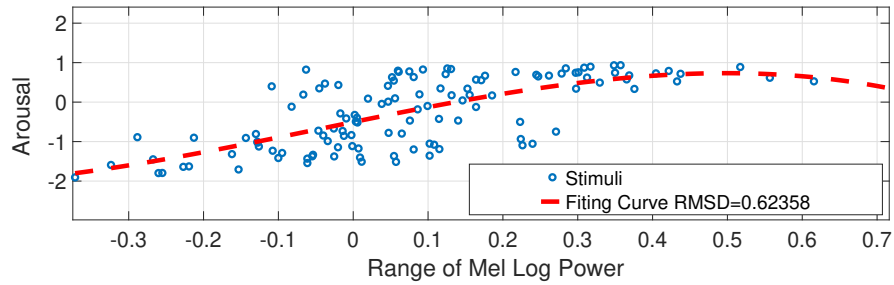
Figure 6.2: Fitting Arousal using HP feature



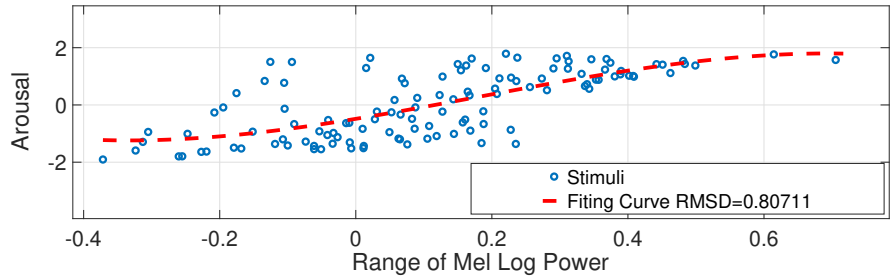
(a) Neutral-Happy voices



(b) Neutral-Angry voices

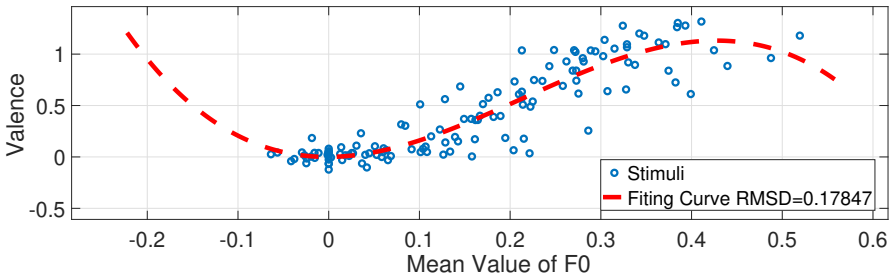


(c) Sad-Happy Voices

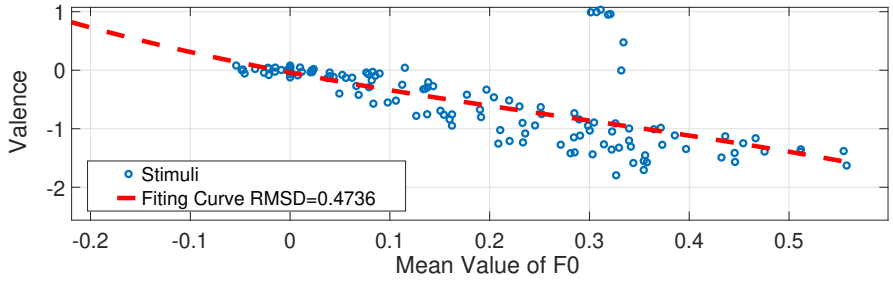


(d) Sad-Angry voices

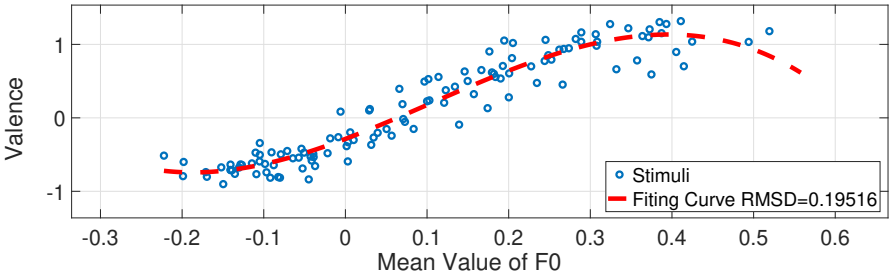
Figure 6.3: Fitting Arousal using RMP feature



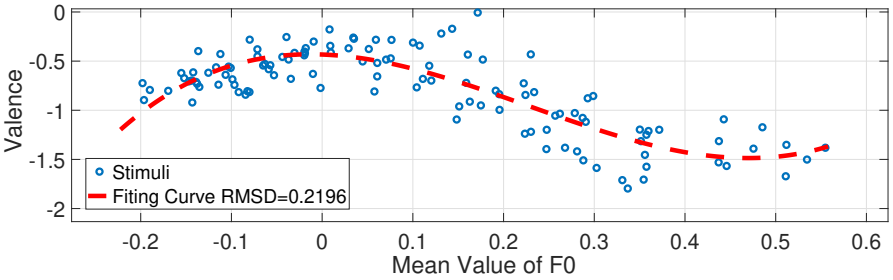
(a) Neutral-Happy voices



(b) Neutral-Angry voices

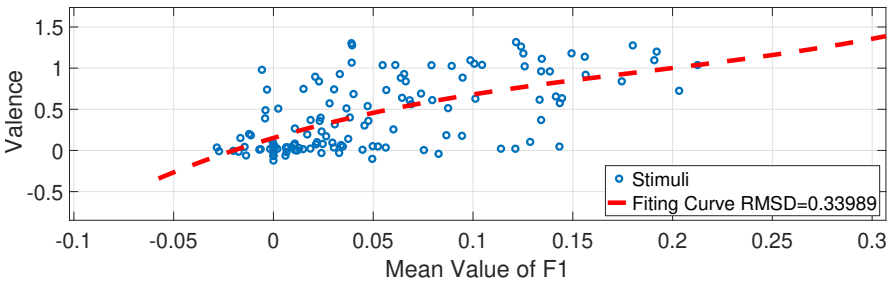


(c) Sad-Happy Voices

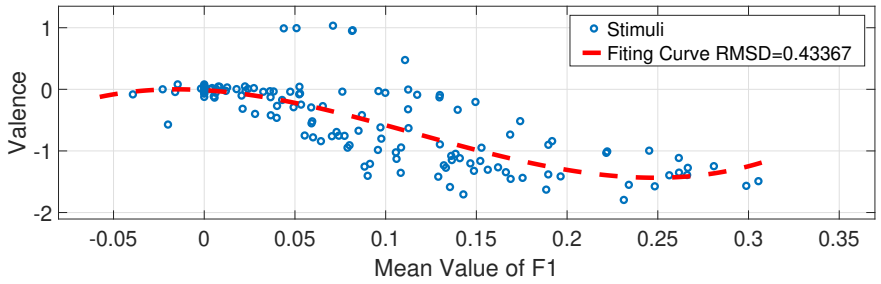


(d) Sad-Angry voices

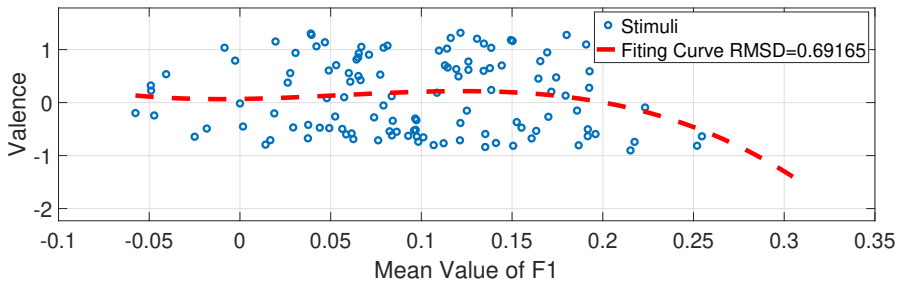
Figure 6.4: Fitting Valence using AP feature



(a) Neutral-Happy voices

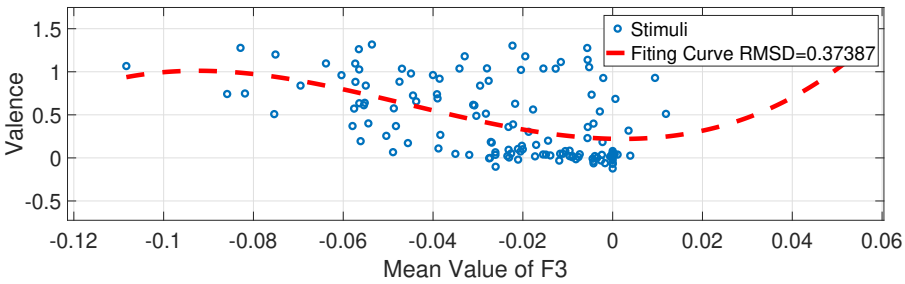


(b) Neutral-Angry voices

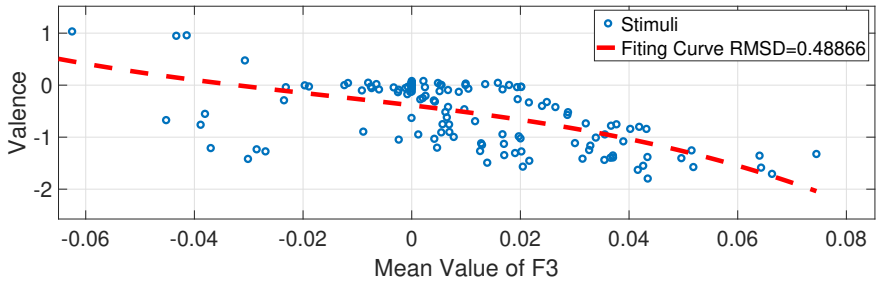


(c) Sad-Happy Voices

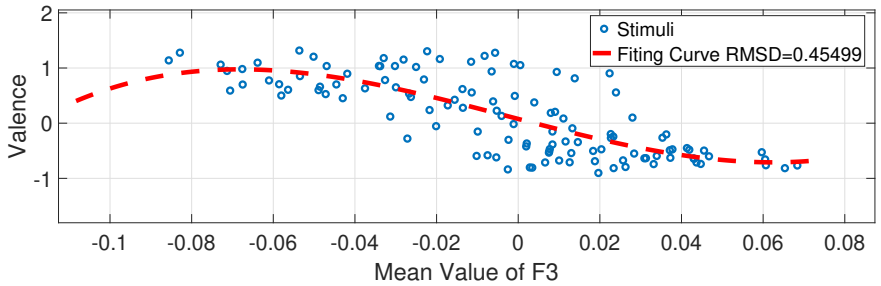
Figure 6.5: Fitting Valence using AF1



(a) Neutral-Happy voices

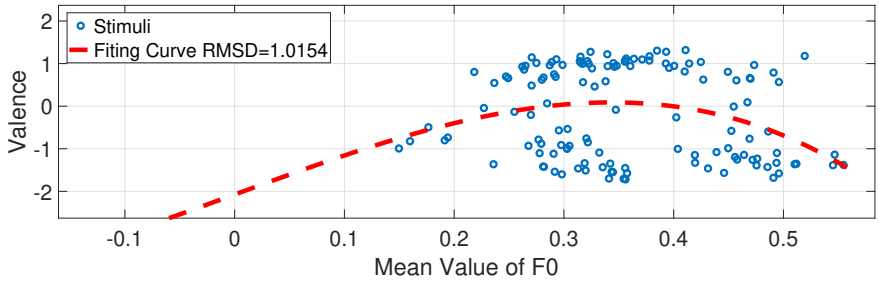


(b) Neutral-Angry voices

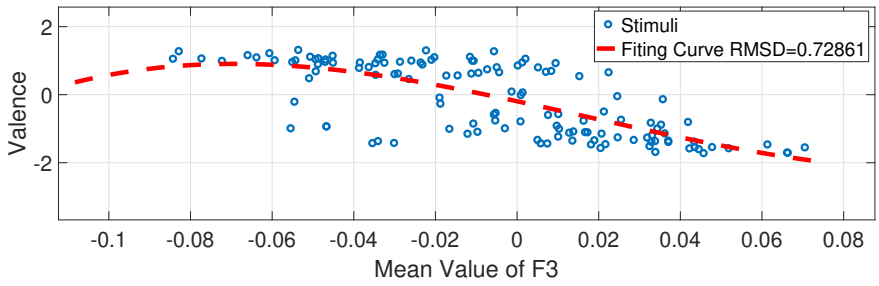


(c) Sad-Happy Voices

Figure 6.6: Fitting Valence using AF3

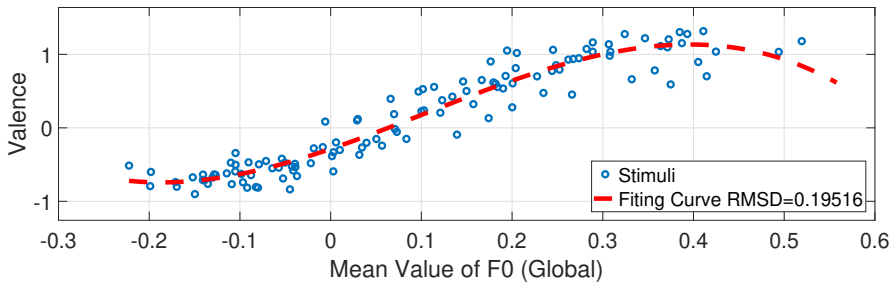


(a) Feature Mean Value of F0

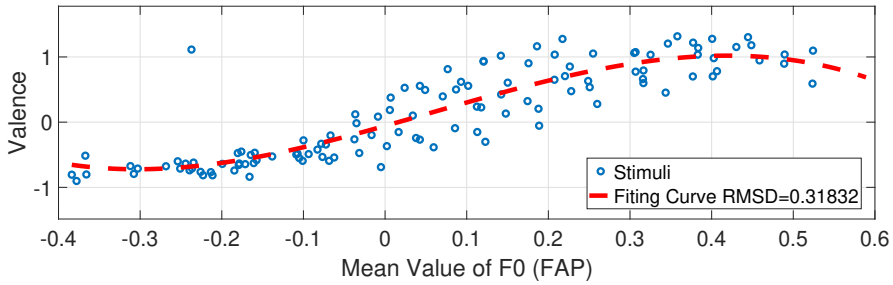


(b) Feature Mean Value of the Third Formant

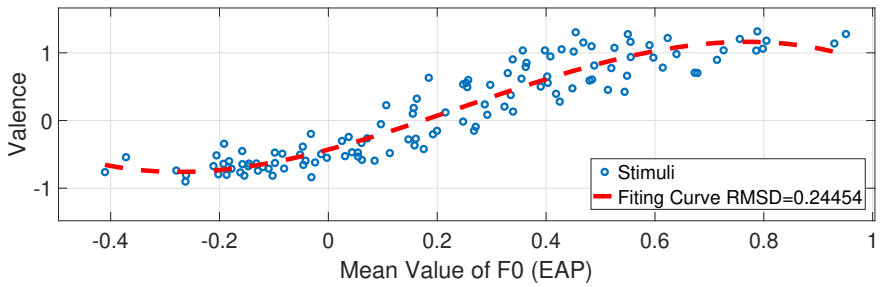
Figure 6.7: Fitting Valence Scores of Angry-Happy Group



(a) Mean Value of F0



(b) Mean Value of F0 in The First Accentual Phrases



(c) Mean Value of F0 in The Last Accentual Phrases

Figure 6.8: Fitting Valence Scores using Global and Accentual Features

Chapter 7

Conclusion

7.1 Summary

As mentioned in section 1.3, this research has two goals. Firstly, this study is going to obtain emotional speech samples with continuous distribution on the V-A space using morphing techniques; Secondly, this study discusses what acoustic features are important to emotional impressions and how those features relate to emotion perception by achieving the first goal. Section 4.2 analyzed the acoustic features of morphed voices, supports that this research had successfully synthesized morphed voices between pairs of reference voices by interpolating acoustic features equidistantly. Next, evaluation scores from listening tests mentioned in section 5.2 verified that synthesized morphed voices are continuously distributed on the V-A space. According to those results, this study confirms that the first sub-goal had been achieved.

Based on acoustic features extracted and emotional evaluation scores collected, this research investigates how acoustic features relate to emotion perception and discusses those results in chapter 6. Base on analyses results, this research ascertains that Arousal perception can be stably described by merely using F0 related features; Power related features significantly influence the Arousal perception, however, limited in sad-related morphed voices. In contrast, Valence perception is at least simultaneously affected by F0 and formants features with significant. Notably, this research discussed how F0 and formant related features influence Valence perception based on Angry-Happy voices mentioned in section 6.2. Those analyses can explain why angry voices synthesized by Xue *et al.*'s system, without formats features modification, are evaluated as happy emotion.

Also, considering the significances and corresponding relationship of different acoustic features varying in different morphing groups as mentioned

in section 6.3, this study has hypothesized that acoustic features, like F0 or different formant components, impact Valence perception together in a non-uniform way on different areas of V-A space. Those analysis results indicate that in order to get an emotional sound synthesis system with better Valence control, it is necessary to propose modification rules for different formant components separately depends on different areas of V-A space. According to the analyses mentioned above, the second sub-goal of this study had been achieved.

7.2 Contribution

A sophisticated rule-based emotional voices conversion system with continuous emotion representation should be able to convert standard input voice to any location on the emotion space through modification rules. In order to achieve this goal, it is necessary to analyze the emotional voices continuously distributed on the emotion space to obtain the relationships between acoustic features and emotional perceptions, and propose corresponding modification rules.

Under the ultimate goal mentioned above, this study has two contributions. Firstly, this study has successfully synthesized morphed voices that continuously changed in physical acoustic features and emotional impression, which voices samples could not be obtained through the traditional recording process. All synthesized morphed voices, the extracted acoustic features, and the emotional evaluation scores obtained from the listening tests are not only meaningful material for this study but also can be used as research materials in subsequent researches.

Secondly, based on synthetic morphed voices, this study analyzes the relationship between various acoustic features and emotion evaluation scores. Those fitting rules obtained in this study can not only be used to adjust the modification rules of the existing system, but also help the subsequent researches to propose a new modification model, focusing on those features that have a significant impact on the emotional perceptions but have not been adjusted by the existing systems, such as formants modification model for Xue's system.

7.3 Remained Works

The modification rules for adjusting acoustic features of stand input voices to emotional voices are essential for emotional voices conversion as mentioned

in section 2.1 and 7.2. Since the results indicate that formant components have the significant influence on valence impression and an emotional speech conversion system without formants control has a defect on Valence adjusting, it is necessary to propose a suitable method to modify the formants information for Valence control. Besides, considering that multiple acoustic features impact Valence perception together in a non-uniform way on different areas of V-A space, a highly sophisticated system may require different modification rules for different target synthetic voices at different locations in the V-A space.

Bibliography

[1] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004, International Conference*, 2004.

[2] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective computing and intelligent interaction*. Springer, 2005, pp. 981–995.

[3] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[4] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with mcc and f0 features," in *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*. IEEE, 2016, pp. 1–5.

[5] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[6] M. Schroder, "Expressing degree of activation in synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1128–1136, 2006.

[7] J. Dang, A. Li, D. Erickson, A. Suemitsu, M. Akagi, K. Sakuraba, N. Minematsu, and K. Hirose, "Comparison of emotion perception among different cultures," *Acoustical Science and Technology*, vol. 31, no. 6, pp. 394–402, 2010.

[8] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[9] X. Li and M. Akagi, "Multilingual speech emotion recognition system based on a three-layer model." 2016.

[10] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[11] M. Schröder, "Emotional speech synthesis: A review," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[12] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech communication*, vol. 52, no. 5, pp. 394–404, 2010.

[13] A. W. Black, "Unit selection and emotional speech," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[14] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[15] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

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

[17] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards emotional speech synthesis: A rule based approach," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[18] J. M. Montero, J. M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: From speech database to tts," in *Fifth International Conference on Spoken Language Processing*, 1998.

[19] 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 sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[20] M. Akagi and Y. Tohkura, "Spectrum target prediction model and its application to speech recognition," *Computer Speech & Language*, vol. 4, no. 4, pp. 325–344, 1990.

[21] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion to emotional speech based on three-layered model in dimensional approach and parameterization of dynamic features in prosody," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–6.

[22] H. Kawahara and M. Morise, "Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, 2011.

[23] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3933–3936.

[24] H. Kawahara, T. Takahashi, M. Morise, and H. Banno, "Development of exploratory research tools based on tandem-straight," in *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal and Information Processing Association, 2009 Annual ⋯, 2009, pp. 111–120.

[25] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–I.

[26] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3905–3908.

[27] H. Matsui and H. Kawahara, "Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[28] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[29] S. Lui, "A preliminary analysis of the continuous axis value of the three-dimensional pad speech emotional state model," in *The 16th International Conference on Digital Audio Effects (DAFx)*, 2013.

[30] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[31] ——, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[32] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program] version 6.0.46," 2019. [Online]. Available: http://www.praat.org

[33] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[34] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[35] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson Upper Saddle River, NJ, 2011, vol. 64.

[36] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

Publications

1. Zi Wang, Maori Kobayashi and Masato Akagi, "**Study on Relations between Emotion Perception and Acoustic Features using Speech Morphing Techniques**"(*The 2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'19)*, Hawaii, US)