A Study on a Three-Layer Model for the Perception of Expressive Speech

by

Chun-Fang Huang

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology

Dec. 2008

Abstract

The goal of this research work is to find the answer to the question about what role non-linguistic information plays in the perception of expressive speech categories, and specifically, in the case with people who are from different culture/native-language backgrounds. A multi-layered model is proposed, which is based on the fact that there is a vagueness nature in human cognition. To achieve the research goal, this model will be built by perceptual experiments, verified by rule-based speech morphing, and applied to the analysis of non-linguistic verbal information.

The resulting model by the building process suggests that before listeners decide to which expressive speech category a speech sound belongs, they qualify a voice according to different descriptors, where each descriptor is an adjective for voice description. Different combinations of such descriptors by which a listener qualifies a voice can lead to the decision about to which different expressive speech categories that voice belongs.

To verify the resulting model, a rule-based speech morphing approach is adopted. The built model is transformed to rules, which are used to create morphed utterances that are perceived as different semantic primitives and expressive speech categories with different intensity levels. The verification process goes a step further to show the validity of the finding in the building process.

Finally, the same model is applied with the same stimuli but different listeners with different culture/native-language background to show that the role non-linguistic information plays in perception of expressive speech categories is common to people who are from different culture/native-language background.

Acknowledgements

It is not possible to complete this research work without numerous supports from many people. I cannot hope to be exhaustive in my thanks to all the people who contributed in some way to my research work, but there are certain individuals whom I wish to single out for recognition.

First of all, I would like to express my deepest gratitude and appreciation to my supervision Professor Masato Akagi for supervising my research with great care and advices. I am very lucky to have him to be the supervisor of my master and PhD study. His constant encouragement and many perceptive comments do not only help me to clarify my thinking when I was in a mess but also inspire me to find my own direction of research. He teaches me how to do a research meanwhile provides all supports I need, as well as demonstrate how to be a good researcher. I wish I could be a good researcher like him some day. My research work would never be successful without him.

My sincere appreciation goes to Professor Donna Erickson (from Faculty of Music, Showa University of Music). As my advisor of sub-theme, she does not only provide invaluable suggestions on my research, but also grateful encouragement and mental support. I am most grateful for the effort she devoted to reading my many drafts and making constructive suggestions. She is a great supervisor and a very good friend.

I would like also to express my sincere thanks to all member of the jury, Professor Donna Erickson (from Faculty of Music, Showa University of Music), Professor Jianwu Dang, Associate Professor Isao Tokuda and Associate Professor Masashi Unoki of the School of Information Science of JAIST. It is my honor and I would like to thank you all for serving on my thesis committee. I appreciate their efforts and the time they have spent examining this thesis.

I am very grateful to my principal advisor Professor Jianwu Dang for invaluable comments and encouragement during my work.

I wish also to extend my great thanks to Associate Professor Masashi Unoki for not only his great help on my research work but also for his scrupulosity in many business trips and my life in JAIST.

I would like also to thank Associate Professor Isao Tokuda for his kind supervision during regular meetings.

I wish to express my gratitude to Fujitsu Lab for the permission to use the emotional speech database.

My special thanks go to Dr. Kazuhito Ito for helping me to organize an experiment at start and provided many valuable discussions.

I would like to thank Dr. Yuichi Ishimoto, Dr. Hironori Nishimoto, Dr. Takeshi Saitou, Mr. Atsushi Haniu and all other member of the acoustic information science Lab. in JAIST. They have been very friendly and given me a helping hand which made my study life in JAIST been easier and more joyful.

I am deeply indebted to the COE offered by JAIST, for three years financial support in research and business trips.

I am forever grateful and in depth to my mother, in memory of my father and my family for their endless loving support throughout my life.

Finally, this dissertation is dedicated to my husband Brian, my sons Wei-Wei and Wei-Chi. They are the most important to me. My life is nothing without them.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

Communication is one of the most important activities of a human being. Humans use different kinds of communication tools, such as gestures, writing, music or speech to interact with each other. The information transferred by different communication tools is encoded through various types of media and recognized by a variety of receptor receivers. For example, music not only gives listeners beautiful sounds with melody to hear, but also communicates the emotion of the composers and performers [80]. Gestures, another example, communicate both information and emotion [40]. In speech communication, it is clearly not only language that is decoded, but also other types of information, such as age, gender, or even emotional categories, such as joy, sadness, anger etc. are decoded [22, 48, 49], and may not necessarily be all that apparent. It is interesting to know the kind of information in speech that can lead to the judgment of emotional categories. This question does not only apply to the situation that a listener is acquainted with the language they heard, but also to those that are not acquainted with that language. Although previous work had shown a statistical relationship between the physical characteristics of voice and the judgment of emotional state, these results are still far from satisfactory. The knowledge of how acoustic characteristics in voice are related to the judgment of expressive speech perception gained from previous work does not provide the necessary information to understand how humans' interpretation of acoustic characteristics in voice involves the human vagueness nature.

The answer to this question is not apparent but is possible to be approached from different aspects, such as psychology, physiology, or engineering. For the application of expressive speech, such as the automatic emotional category detection or expressive speech synthesis, it is not only necessary to consider this question by understanding how acoustic characteristics of voices affect the judgment of emotional categories, but also by understanding how the vagueness nature of humans affects this judgment. The acquirement of this understanding may the key point to a better application of expressive speech detection and synthesis.

This vagueness nature in the context of expressive speech can be understood better by the following observation. When listening to the voice of a speaker, one hears that a voice "sounds bright and slightly fast", which in turn one can interpret that "the speaker is happy". Nevertheless, we do not say, "This voice sounds as if its fundamental frequency is 300 Hz", and thereby proceed to perceive what emotion the speaker is expressing, simply by identifying the absolute pitch of the voice. At the same time, the subtle changes in the descriptors (i.e. adjectives) of the sounds, e.g., “bright-sounding” or “fast-sounding”, can also lead to different judgments of emotional categories. For example, two voice sounds *slightly* *bright* and *bright* respectively can lead to different judgment of *slightly* *joyful* and *joyful*.

Therefore, to find the answer about what role non-linguistic information plays in perception of expressive speech categories and to apply this finding to the development of expressive speech detection and synthesis, a systematic understanding of the relationship between acoustic characteristics in voice and the vagueness nature of human perception is necessary. Daily observation shows the marvelous ability of humans to decode emotional categories from speech even without the understanding of language; a systematic understanding of this relationship can clarify this “gift” of humans. However, this relationship is not that apparent. There are various types of acoustic characteristics in a voice. On one hand, a subtle change of them can lead to different perceptions of voice description and emotional categories. At the same time, it is also possible that different people may use different descriptors for describing voices. On the other hand, people from different culture/native-language backgrounds also show a certain degree of similarity in the judgment of expressive speech perception. Even the type of voice data collection for studying expressive speech may change these results. Besides, the usage of adjectives for voice description that leads to the judgment of emotional categories is fuzzy. All this complexity and the fuzzy relationship suggest the necessity of constructing a model of expressive speech perception for revealing the relations between various types of elements involved in expressive speech perception by a “non-traditional” method for representing the vagueness nature of human perception.

This dissertation introduces the construction and the application of this perceptual model, which exhibits the following three characteristics:

(1) It takes a human’s vagueness nature into consideration

(2) It uses descriptors (i.e. adjectives) of the sounds to represent the vagueness nature of humans. The usage of different descriptors may be the result from the change of acoustic characteristics in the voice.

(3) It considers the judgment of expressive speech categories as resulting from the combination of different voice descriptors.

The goal of this research work therefore is to use this model for finding the answer to the question about what role non-linguistic information plays in the perception of expressive speech categories, especially in cases where such information may be common to people from different culture/native-language background.

1.1 Research background

In this research work, the construction of the model for expressive speech perception is based on two assumptions.

- The first assumption is deduced from the observation described above, that is, before listeners can decide to which expressive speech category a speech sound belongs, they will qualify a voice according to different descriptors, where each descriptor is an adjective for voice description. Different combinations of such descriptors by which a listener qualifies a voice can lead to the decision about to which different expressive speech categories that voice belongs. Such qualifications and decisions are vague rather than crisp.

- The second assumption, which is derived from the first and is also based on daily observation, is that people who are from different cultures/native-language backgrounds have common characteristics for perception of expressive speech categories.

The involvement of adjectives is not new to the study of acoustics. However, previous work usually focused on the selection of adjectives for voice description. Even though they showed that some adjectives are regularly used for voice description, they still failed to provide insight into how this usage is related to the changes of acoustic characteristics and how the usage of adjectives can be employed to understand the perception judgments. In a similar way, although previous work has already confirmed the common characteristics between people with different native-language/cultures background, we still need an understanding of how these common acoustic characteristics are interpreted by the human vagueness nature involvement of adjectives in this study may remedy this problem. All in all, different from previous work, descriptors play the central role to study expressive speech perception.

An important implication of these two assumptions is that there are actually two relationships that should be considered in this research. One is that the judgment of expressive speech categories is based on the descriptors qualified for voices. The other is that acoustic cues affect the choice of voice descriptors.

To find support for the two assumptions, the focal point for achieving the research goal becomes the building of the relationships in the model. Since the work is limited to the context of non-linguistic verbal information [25, 62], the elements about discrete symbolic information that are connected to syntax and phonetic information of one language are excluded from the model. Instead, acoustic cues, e.g., pitch, duration, loudness, etc., voice descriptors, e.g., “bright-sounding” or “fast-sounding”, and emotional categories, e.g., joy, sadness, anger, etc. are included. From the fact that these two assumptions are new to this research topic, it is necessary to carefully consider different factors involved in this research, which include data collection, the types of acoustic cues measured, the mathematical tool used to build the relationships between the involved elements, and the supplementary support of using voice descriptors in studying human perception. The novelty of this model also implies that there will be no literature to evaluate the effectiveness of the constructed model. Therefore an additional evaluation of the model is also necessary.

- Data collection is essential because data collected by using different methods may exhibit different characteristics. For research involved with perceptual experiments, the different characteristics exhibited in the collected data essentially have an effect on experimental results. Inappropriate choice of data collection type may result in undesired results. In Section 1.1.1, the types of data collection and the characteristics of them are discussed.

- An ideal approach for acoustic cue measurements is to measure as many types as possible in order not to overlook any important hint. However, due to time and resource constraints, a more realistic approach is to only consider those that have shown significance in expressive speech perception. Previous work on this topic provides a good source of evaluation of the different choices. Section 1.1.2 reviews literature for this evaluation.

- Mathematic tools directly influence the relationship of the resulting model. One popular tool is statistics. Statistics provide an in depth understanding about how two elements are related. The basic assumption of this tool is that the characteristics of both elements can be exactly described by number values. However, for the vagueness nature of human perception, we need a different mathematical tool for them. This evaluation is described in Section 1.1.3.

- Human perception involving voice descriptors not only can be observed from our daily speech communication but also from other communication scenes, such as music. In Section 1.1.4, a review of literature in music provides a supplementary support for applying this idea in this research work.

- Since the model is about human perception, it should be evaluated by using a human-oriented approach. An important issue in this process is to find a simple form for representing the relations between elements involved in the model. These include the relations between different types of acoustic characteristics and the choice of voice descriptors for the voice that exhibit these characteristics, or the relations between different types of voice descriptors and judgment that results from a voice with these voice descriptors. To this end, a rule system will be developed, which contains the combination of elements and the quantity of the elements to be used in the evaluation process.

As pointed out by Erickson in [19], the method of data collection adopted should be decided by the goal of the research work. There are two major types of data which are mainly different in the methods used for data collection. One is a spontaneous database, and the other is an actor-portrayed database [16]. Previous work that examines expressive speech from the relationship between acoustic cues and expressive speech categories have shown a general agreement that the most influential factors belong to the prosodic cues, including voice quality cues. To correctly express the relationships between the different types of elements within the proposed model, a mathematical tool should be appropriately chosen. This mathematical tool should satisfy the characteristics of elements that are involved in a relationship.

1.1.1 Data collection

Many researchers [15, 27, 58, 64] have found that who acts the utterances and how the utterances are recorded can affect the results of the study. A spontaneous database records naturally speaking utterances from real-life speech. An actor-portrayed database records utterances produced by professional actors/actresses. Some examples of both types are described in this section. A spontaneous database is ideal for research because it is closer to the nature of emotional speech perception. Conversely, an actor-portrayed database usually results in a rather explicit detection of emotional categories. However, due to the emphasis of the stereotypical performance by the actors/actresses, it may suffer the problem that some subtle factors are hidden [64].

For the spontaneous database examples, to study the communication between human and automatic dialogue system, Batliner, Fischer, Huber, Spilker, and Nöth [6] use a Wizard-of-Oz (WOZ) technique to acquire natural responses from human subjects. This technique hides the fact to subjects that their communication with a computer is actually operated by humans. Therefore, it is expected that natural spontaneous responses from subjects can be obtained, rather their opinions about what single emotional category a speech sound belongs to. Devillers, Vidrascu, and Lamel [15] use two real-life databases from call centers. Their intention is to show that the emotion categories detected in real-life communication are a mixture, rather than a single category. To explore the psychological aspect of automatic emotion detection, Grimm, Kroschel, Mower, and Grimm, Mower, Kroschel, and Narayana [27] use two databases, EMA Corpus and VAM Corpus. EMA Corpus contains acted utterances. On the other hand, VAM contains spontaneous utterances recorded from TV-shows. They used the later in their research and compared the results generated from these two sets of data. They showed a consistent result between these two types of database when applying their approach for emotional detection.

For the actor-portrayed database example, Williams and Stevens [82] used utterances that imitated the crash of “Hindenburg”. To exclude the problems of data-dependent results, Banse and Scherer [4] used data with 1344 voice samples which portrayed 14 emotion categories. Hashizawa, Takeda, Hamzah, and Ohyama [29] used voices recorded by four professional actors, which included two males and two females, to portray utterances (words) with three types of emotional categories, which consists of four different degrees. They studied the intensity levels of three expressive categories, anger, sadness, and joy. They found the relationship between these categories and F0 and speech rate.

Previous work suggests that each method has its distinct characteristics. There is no perfect choice in this. For our research goal that is about non-linguistic verbal information, it is better to have data that clearly reveal the important acoustic characteristics in the voice

1.1.2 Acoustic cues measurements

The review of literature has shown a general agreement that the most influential factors belong prosody which also includes voice quality. Since a sentence can be uttered with different prosody depending on the emotional states of the speaker, the change of prosody of an utterance greatly influences the perception of expressive speech. Therefore, for studying expressive speech perception, prosody is an important factor needed to be investigated. Those acoustic cues considered significant for prosody largely are extracted from fundamental frequency, intensity, and duration. In addition, voice quality is another major focus that researchers have paid much attention to. One definition of voice quality (which also can be referred to as “timbre”) is “the quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar” (ANSI, 1973. *Psychoacoustical terminology. Technical Report*, S.3.30, American National Standard Report.) Although there is still a lack of a consistent definition of voice quality, most voice quality measurements are extracted from the acoustic spectrum.

Prosody

Prosody mainly conveys the perception of pitch, loudness, and rate of speaking. The variations of intonation, pause, stress pattern, speed, belong to what we call the prosody of a sentence. Since a sentence can be uttered with different prosodic characteristics depending on the emotional states of speaker, the prosodic change in an utterance greatly influences the perception of expressive speech. Therefore, for studying expressive speech perception, prosody is an important factor needed to be investigated.

- Fundamental frequency (F0)

From the aspect of its physical meaning, F0 reflects the pitch that is perceived by a listener. The F0 contour represents the change of F0 in the time domain that provides information about the accent and intonation of a sentence. Since such information deeply affects the perception of expressive speech, F0 is the acoustic cue which has been studied from the earliest time and most frequently in the field of expressive speech perception. In [19], Erickson summarized a list of previous research on the types of acoustic cues that are related to the perception of expressive speech categories. No matter which method is used for data collection (acted speech vs. spontaneous) and languages used, most work has found that F0 is important to the perception of expressive speech categories. It is known that voice pitch is strongly determined by F0 [64]. For instance, Paeschke [58], who measured global F0 trend of over 500 German utterances, found that the final F0 movement is different between boredom, sadness, anxiety, and disgust, but almost meaningless for the emotions of happiness and anger. Ishii and Campbell [33] have found that in Japanese the phrase final section of duration and F0 are related to phonation types of expressive speech, e.g., modal, creaky, whispery.

- Power

From the aspect of its physical meaning, power primarily reflects the loudness that is perceived by a listener. It is primarily determined by the volume of air flow of breath sent out by the lungs. When a speaker is in different emotional states, the variation of power could be very wide. Power has also been reported with regard to its relationship to expressive speech (e.g. [32, 44, 65, and 73]. It is also suggested to be influential on voice quality, as suggest by Ishii and Campbell [34]. In their study, power is measured by RMS.

- Duration

From the aspect of its physical meaning, duration primarily reflects how long the listener perceives the sound to be. The same word or same sentence produced with different lengths can be produced differently. Generally, people in a good mood such as happy or cheerful speak faster. In the study of expressive speech perception, for example, [29, 30, 32, 45, 60, 65, 73] have show an effect of duration or speech rates on expressive speech perception in different languages, i.e., German, Japanese, American English, Korean, and Canadian English.

Voice quality

- Spectrum

Voice quality is largely extracted from the spectrum. The spectrum is the speech signal represented in the time-frequency domain, resulting from mathematic transforms of the speech signal over a series of time, such as the Fourier transform,. It is a fact that the spectrum conveys meaningful information for speech perception. For example, the first three formants (F1, F2 and F3) predict the perception of the vowel. In addition to these, the high formants in the spectrum are related to the vocal tract length, and acoustic cues extracted from the spectrum are considered an important indication of voice quality [33, 34]. Although the spectrum is a not-as-well investigated feature compared to F0 movement, it has been found to have a relationship with expressive speech perception (e.g. [18, 44, 45, 65]). Scherer, Banse, Wallbott, and Goldbeck [65] measured low-frequency energy (below 635Hz) and spectral slope of spectral energy distribution and found a relationship to Fear, Joy, Sadness, and Anger. Ehrette, Chateau, d’Alessandro, and Maffiolo [18], who used 100 French utterances recorded by 20 professional actresses, found a relationship between 20 perceptual attributes, (adjectives for voice description freely chosen by subjects), and voice quality attributes (spectral centroide). In other languages, such as Japanese, Maekawa [45], measured six types of paralinguistic information in speech, and found the relationship between these six types and the vowel spectrum.

From the above review, it is clear that fundamental frequency, power, duration, and spectrum are related to expressive speech perception. Although it is not clear how these acoustic cues are related to non-linguistic verbal information, e.g. voice descriptors, they are the candidates for examining for comparison with previous studies.

1.1.3 Mathematical tool

To correctly express the relationships between different types of elements within the proposed model, a mathematical tool should be appropriately chosen because it directly influences the resulting relationship. This mathematical tool should satisfy the characteristics of elements that are involved in the relationship. The characteristics include:

- Linearity or nonlinearity of relationship. A linear relationship means that the cause and effect is proportionate. Conversely, a nonlinear relationship means that the cause and effect is disproportionate. The assumption of the role voice descriptors play in expressive speech perception implies that a human's judgment of expressive speech categories is a quality-issue rather than a quantity-issue. Therefore, the perceptual relationship between voice descriptors and expressive speech categories is possibly non-linear rather than linear.

- Precision and vagueness of relationship. A precise relationship which includes an understanding of the cause and effect and can be modeled by precise number values and expressed by exact equations is necessary. Conversely, a vague relationship is one that the cause and effect can only be understood by imprecise information, and can only be modeled by a rather ambiguous way. A precise relationship can be frequently seen in our natural world; nevertheless, a vagueness relationship is abundant in our human interactions—such as the assumption that we made in this research that voice descriptors are important in human perception. Therefore, it is necessary to take such vagueness into account when choosing a mathematical tool to deal with human perception.

- Black box and white box of relationship [71]. A black-box relationship means that there is no previous knowledge but only measured or observed data are known about the relationship. Conversely, a white-box relationship means that there is structural knowledge known about the relationship. In this research work, the structural knowledge of using voice descriptors in expressive speech perception can be obtained in advance.

For one part of the first assumption made in this research--that the judgment of expressive speech categories is based on voice descriptors--it is necessary to develop a suitable mathematical tool to account for this fuzzy relationship. At the same time, for the other part of this assumption--how acoustic cues affect the choice of voice descriptors—it is necessary to have a mathematical tool that is linear and precise.

The choices of tool we have are statistics, neural network, or fuzzy logic. Statistics is a general tool for describing the characteristics of a relationship. By statistical methodology, the most significant features can be selected for the classification or identification of emotions (e.g., [5, 9, 42, 52, 53, 65, 66, 67, 74, 79, 81, and 82]). However, statistical methodology mainly manages linear and precise relationships but not fuzzy relationships.

Neural network (NN) can be used to model the relationship between two types of elements, which are regarded as input and output, respectively, where structural knowledge is not known. Recently-emerging in the field of the perception of expressive speech categories is the hybrid Hidden Markov Model – Neural Network (HMM/NN). This approach overcomes the problem of classical Continuous Density Hidden Markov Models (CDHMM) [13]. For example, Athanaselis, Bakamidis, Dologlou, Cowie, Douglas-Cowie, and Cox [2] use this approach to detect emotion from speech. One unique characteristic of this is that the relationship which maps to expressive speech categories includes both language content and prosodic information. Besides HMM/NN, there are many other types of neural networks. In the context of this research topic, another popular type is Support Vector Machine (SVM) and Neural Network (NN). They have been used for studying the perception of expressive speech categories in different languages, such as Japanese [56], Swedish [57], Dutch [51], and German [46]. Neural Network can deal with various relationships probably including nonlinear and vague relationships; however, it makes a black box type of relationship since it cannot provide structural knowledge of a relationship.

Fuzzy logic is another emerging tool, which requires known structural knowledge. For example, Lee and Narayanan [11] built the relationship between vocal cues of input utterances and outputs of emotions described by a set of rules. These rules are used in fuzzy inference systems (FIS) for recognizing two groups of emotion labels, negative and non-negative emotions. In [27], Grimm, Mower, Kroschel, and Narayana built a fuzzy inference system for estimating continuous-values of the mapping from emotion primitives of valence, activation, and dominance to emotional categories. The FIS is used as the estimator for emotion detection. In [3], there is an interesting implementation of a robot for emotion detection from the prosody of natural speech by using fuzzy logic. Fuzzy logic provides the capability to deal with the characteristics of nonlinearity, vagueness and white box of a relationship.

From the above review; it is clear that two different methods are necessary for the two different parts of the first assumption. First, the fuzzy relationship between voice descriptors and expressive speech categories requires fuzzy logic because it is non-linear, vague, and white-box. Second, the crisp relationship between voice descriptors and acoustic cues requires statistics because it is linear and precise.

1.1.4 Voice descriptors for music perception

Using qualitative descriptors to describe expressive utterances is also found with other types of communication. In music, the use of adjectives for describing sound has been conducted by many researchers. For example, qualitative descriptions of loud, fast, and staccato music performance give listeners a perception of anger [36]. The perception of these three qualitative descriptions is actually related to the acoustic features of tempo, sound level, and articulation. In music the word “timbre” is also used to describe the difference between two instruments, such as a violin and trumpet, playing the same sound (e.g., having the same F0, loudness and duration). Ueda [77] discussed the possibility of a hierarchical structure of qualitative adjectives for timbre, and suggested that the psychological perception of sound can be characterized by acoustic features. Ueda and Akagi [78] examined how the perception of sharpness and brightness is affected by amplitude envelope shapes, sound-pressure level, and duration of broadband noise. It is also widely used for music assessment. Darke [14] asked musicians to use 12 verbal adjectives for assessing a number of music sounds, and found there was agreement among the adjectives listeners used to describe their perception of timbre. Traube, Depalle, and Wanderley [75] studied the relationship between acoustic features, gesture parameters of guitar playing, and adjective descriptions of guitar timbre. They confirmed the relationship between perceived brightness and gesture parameters of playing a guitar. Coincidently, the use of fuzzy logic to build a perception model can also be seen in music. Friberg [23] used fuzzy logic to map measured parameters, such as sound level or tempo, to emotional perception of sound receivers. They also developed a rule system for modeling the music performance [24]. They used the measured parameters for modeling different styles of performance. Due to the abundant and complex parameters of rules, they simplified the rules selection by relating rules to semantic descriptions, such as lento, andante, vivace, etc., which are qualitative terms for describing music performance. They applied the rule system to synthesize music for different emotional expressions.

1.2 Research approach

To achieve the research goal, that is using a model for finding the answer to the question about what role non-linguistic information plays in perception of expressive speech categories, where such information is common to people who are from different culture/native-language background, two assumptions were made. Since these two assumptions imply two relationships, the building of these two relationships becomes the focal point in the research work. However, since much previous related work mainly focused on the statistical relationship between expressive speech and acoustic cues, there are no previous studies that are useful for evaluating the resulting relationships that are going to be built. Therefore, another accompanying focal point is to evaluate the relationships that will be built.

To clarify this building and evaluation process, a multi-layered approach is proposed. The two relationships form a three-layered model, which exhibit the three characteristics described in the beginning of this chapter by including “acoustic features”, e.g., pitch, duration, loudness, etc., as “semantic primitives”, e.g. voice descriptors of bright, fast, slow, etc., and “expressive speech categories”, e.g. joy, sadness, anger, etc. The construction of this three-layered model becomes a two-step process.

In the first step, in order to find those descriptors (i.e., adjectives) that can be used to describe the perception of expressive vocalizations, a thorough selection of semantic primitives should be conducted. After the semantic primitives are selected, analysis of a large number of acoustic features will be necessary to support the relationship between semantic primitives and acoustic features. Moreover, in order to understand the fuzzy relationship between the linguistic description of acoustic perception and expressive speech, a fuzzy inference system will be built. After the two relationships are built, the effectiveness of relationships should be evaluated.

In the second step, to verify the effectiveness of relationships built in step 1, we take a more basic, human-oriented approach instead of an engineering application, as in building a system for automatic expressive speech classification [27]. We use the semantic primitive specifications with corresponding acoustic features of the first step to resynthesiz the expressive speech utterances, and ask human subjects to evaluate the effectiveness of the resynthesized utterances. This approach includes three techniques: (1) a voice-morphed technique for rule verification, (2) expression of the relationships built by step 1 as rules, and (3) examination of various types of human perception of emotion as well as the intensity of the perceptions.

Voice morphing, which is also referred to as voice transformation and voice conversion, is a technique to modify the identification of a source speaker to a different target speaker [88, 89, 90, 91]. The conversion mapping from source to target is learned and aligned by different methods, such as vector quantized codebooks [92], hidden Markov models (HMM), or Gaussian mixture models (GMM) [93, 94]. However, as indicated by Pfitzinger [95], voice morphing and voice conversion are different. In voice morphing, new voices that are learned from the relationship between source and target voices are interpolated. Conversely, in voice conversion, unseen voices of the source speaker are extrapolated toward the target speaker.

Different from the modification technique described above, in this research work rules are created with parameters that change the acoustic characteristics of a neutral utterance from a specific speaker to those that are perceived as having certain semantic primitives or belonging to certain expressive speech categories. These parameters come from the fuzzy inference system and the correlation of these two relationships. Changing speech in this way is a technique explicated by Scherer [64], and technically developed by Kawahara et al. [37].

This rule-based approach implements and controls the variations of the perceptual model. It helps to verify the model we build in the first step. It takes a neutral voice as input and generates a different voice as output by systematically manipulating the acoustic features of the input voice. The parameters will then be modified to create variations of the rules that should give different intensity-levels of perception. The verification process is simplified by the combination of this type of speech-morphing and base-rule development. The relationships revealed in the building process then can be verified by changing a neutral utterance to other utterances which result in different types and intensities of perception. This approach is also called copy synthesis, resynthesis or a type of voice-morphing [96, 97, 98].

After the construction of this three-layered model, it will be applied to provide support for the second assumption to show that people who are from different cultures/native-language background have common characteristics for perception of expressive speech categories.

1.3 Dissertation organization

This paper is organized as follows:

In Chapter 1, the orientation of this research topic is clarified. The description in this chapter makes it clear what different factors should be considered before processing the research work. The factors considered here include data collection, acoustic cues measurement, and mathematical tools. Previous work related to each factor is reviewed as well. The information summarized provides a clear outline for the approach. Moreover, a brief introduction to the proposal of the multi-layer approach for expressive speech perception is described.

In Chapter 2, a more detail explanation about the three-layer model proposed in this research work is given. The hypothesis we attempt to prove is that people perceive expressive speech not directly from a change of acoustic features, but rather from a composite of different types of "smaller" perceptions that are expressed by semantic primitives. A conceptual diagram is used to describe the constituents of the model, which consist of three layers, expressive speech, semantic primitives and acoustic features, and the relationships between the layers.

Chapter 3 starts with the detailed description of how the model is built, specifically, the construction of the two types of relationships-- the relationship between expressive speech and semantic primitives and the relationship between semantic primitives and acoustic features. The first relationship, between categories of emotion and semantic primitives, was built by conducting three experiments and applying fuzzy inference systems. The second relationship, between semantic primitives and acoustic features, was built by analyzing acoustic features in speech signals. Finally, the resulting model is presented.

The essence of how to verify the resulting model is explained in Chapter 4t. It is followed by the description of the manner and results of verification for the model. The resulting model was verified by a rule-based speech-morphing approach. The rules are created from the analytic results. These rules are used to morph the acoustic characteristics of a neutral utterance to the perception of certain semantic primitives or expressive speech categories. Different from other research studies, the various types of human perception of emotion as well as the intensity of the perceptions from the resulting models are also verified.

The actual application of the resulting models is reported in Chapter 5. The main objective of the application is to pursue the fundamental nature of the role that non-linguistic information plays in speech communication. That is, do we humans have something in common especially concerning the perception of expressive speech? The manner in which this application works is also discussed. Moreover, one feature of the application which is highlighted is that common perceptual features are linked to semantic features. By comparing the perception model built from Japanese and Taiwanese subjects, we see that certain common as well as different features of expressive speech perception exist among subjects from different linguistic backgrounds.

Chapter 6 provides an overall conclusion of the thesis. It includes a brief summary, as well as a discussion of the possible significance and applications of this research. A number of possible directions extending from this work are pointed out and put forth.

Chapter 2. Overview of the Model

2.1 Conceptual diagram of the model

A diagram shown in Figure 2-1 conceptually illustrates the proposed multi-layered approach of this research work. This diagram visualizes the two assumptions we made in expressive speech perception as a layered-structure model. This model illustrates that people perceive expressive speech categories (i.e. the topmost layer) not directly from a change of acoustic features (i.e. the bottommost layer), but rather from a composite of different types of "smaller" perceptions that are expressed by semantic primitives (i.e. the middle layer). It also shows that the significance of this research work is that the model approximates a human's behavior. Humans' perceptions always have a quality of vagueness; for example, humans use vague linguistic forms which have not precise values, such as "slow" or "slightly slow". Moreover, this model does not only apply to people who are acquainted with the language of the voice they heard, but also to those who are not acquainted with that language.

A close look at the model shows that it consists of three layers: expressive speech, semantic primitives, and acoustic features. Five categories of expressive speech are studied, *Neutral*, *Joy*, *Cold Anger*, *Sadness*, and *Hot Anger. Anger*, *sadness*, and *Joy* are the most basic and well-investigated emotional categories. In addition to these three categories of emotion, *Neutral* and different types of anger, *Hot Anger* and *Cold Anger are investigated* in order to provide a comparison. The semantic primitives are considered to be a set of adjectives often used by listeners to describe utterance characteristics, for example, high, low, bright, quiet, etc. The acoustic features are the measurable acoustic characteristics in the speech signal, such as fundamental frequency, power envelope, spectrum, etc.

These three layers form two relationships that help to understand expressive speech perception better. The first relationship is that between semantic primitives and expressive speech categories. From this relationship, one can know how the perception of voice affects the judgment of expressive speech categories. The second relationship is that between acoustic features and semantic primitives. From this relationship, one can know how the acoustic characteristics in voice affect the perception of voice. These two relationships also apply to the communication scene when people are not acquainted with the language of the voice they heard.

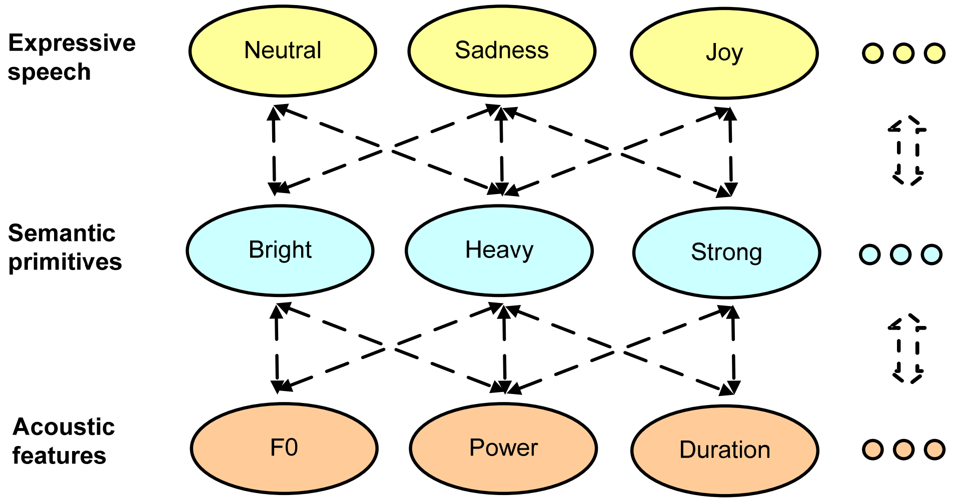


Figure 2-1. Conceptual diagram of the perceptual model.

2.2 Construction process

Figure 2-1 not only shows the static structure of expressive speech perception, but also implies the necessary process to construct it. As described in Chapter 1, the focal point of this research work becomes the building of the two relationships shown in Figure 2-1. Another accompanying focal point is the evaluation of the model. After finishing the building and evaluation of the model, the model is applied to find common features of people who are or are not acquainted with the language of the voice they heard to see if they can still perceive expressive speech. Figure 2-2 illustrates the overall process. Figures 2-3 to 2-5 illustrate the building, evaluation, and application processes along with the corresponding sections in this dissertation.

The following chapters describe the building of the model by perceptual experiments, the evaluation of the model by speech morphing, and the application of the model to the analysis of non-linguistic verbal information.

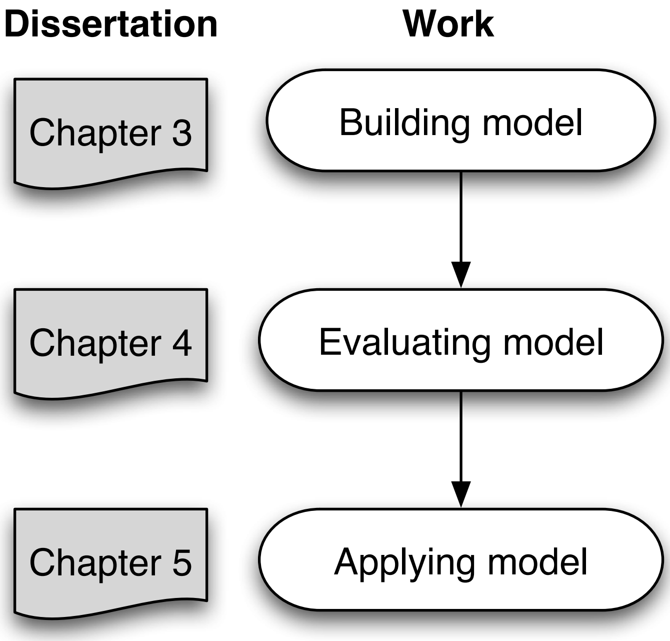


Figure 2-2. Overall process of model building, evaluation, and application

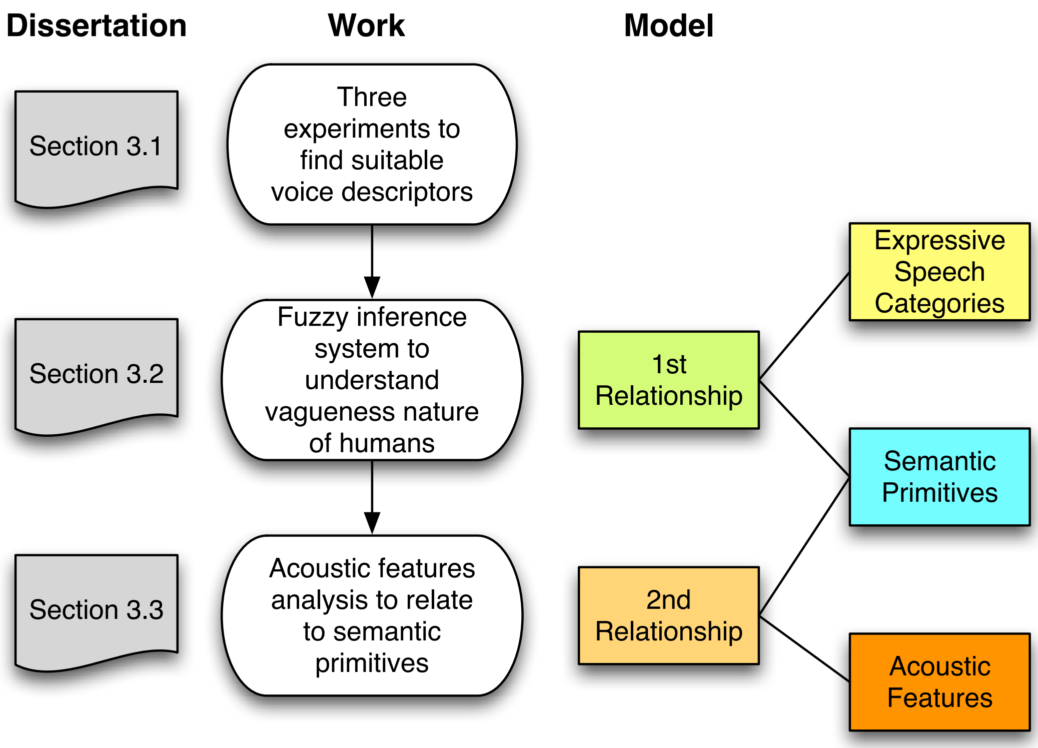

Dissertation
Work
Model
Section 3.1
Three experiments to find suitable voice descriptors
Section 3.2
Fuzzy inference system to understand vagueness nature of humans
Section 3.3
Acoustic features analysis to relate to semantic primitives
1st Relationship
2nd Relationship
Expressive Speech Categories
Semantic Primitives
Acoustic Features

Figure 2-3. Building process, and the corresponding sections in the dissertation and in the model.

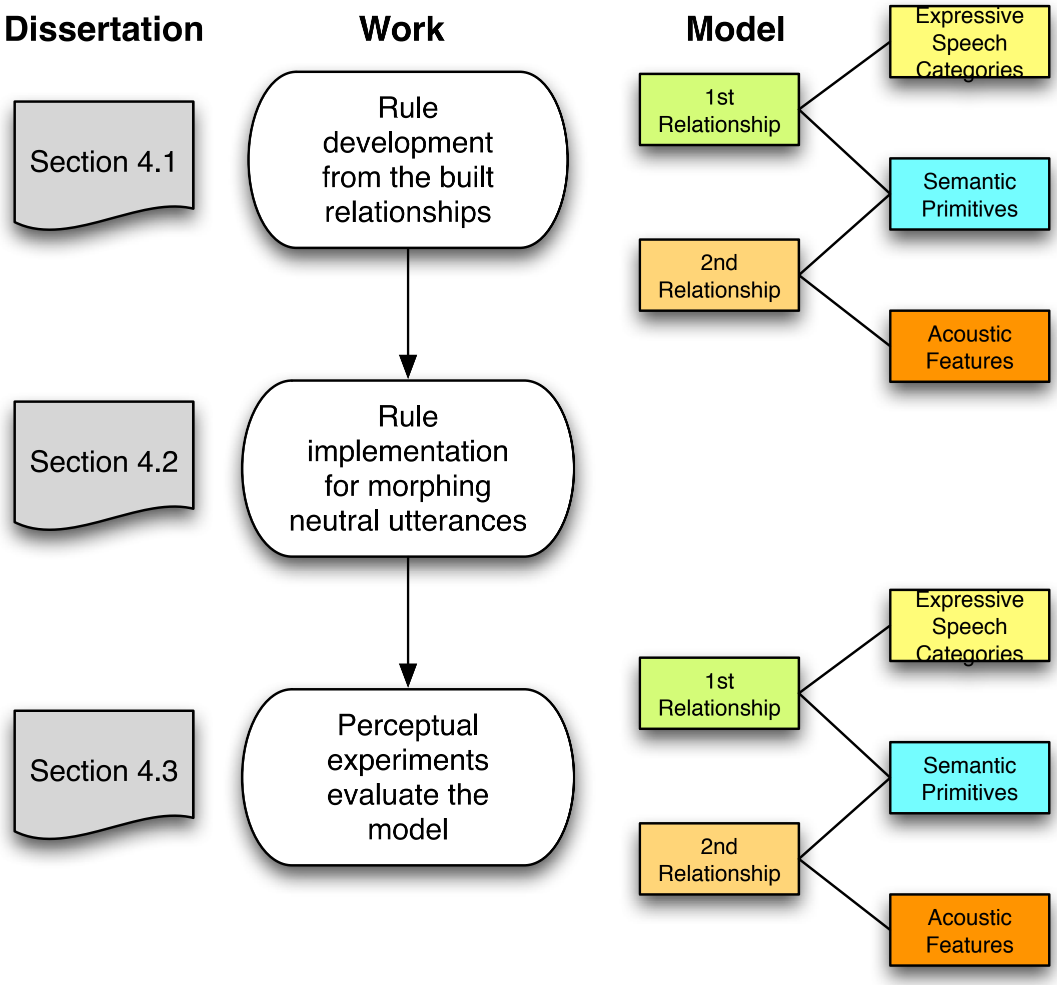

Dissertation
Work
Model
Section 4.1
Rule development from the built relationships
1st Relationship
2nd Relationship
Expressive Speech Categories
Semantic Primitives
Acoustic Features
Section 4.2
Rule implementation for morphing neutral utterances
Section 4.3
Perceptual experiments evaluate the model
1st Relationship
2nd Relationship
Expressive Speech Categories
Semantic Primitives
Acoustic Features

Figure 2-4. Evaluation process and the corresponding sections in the dissertation and in the model.

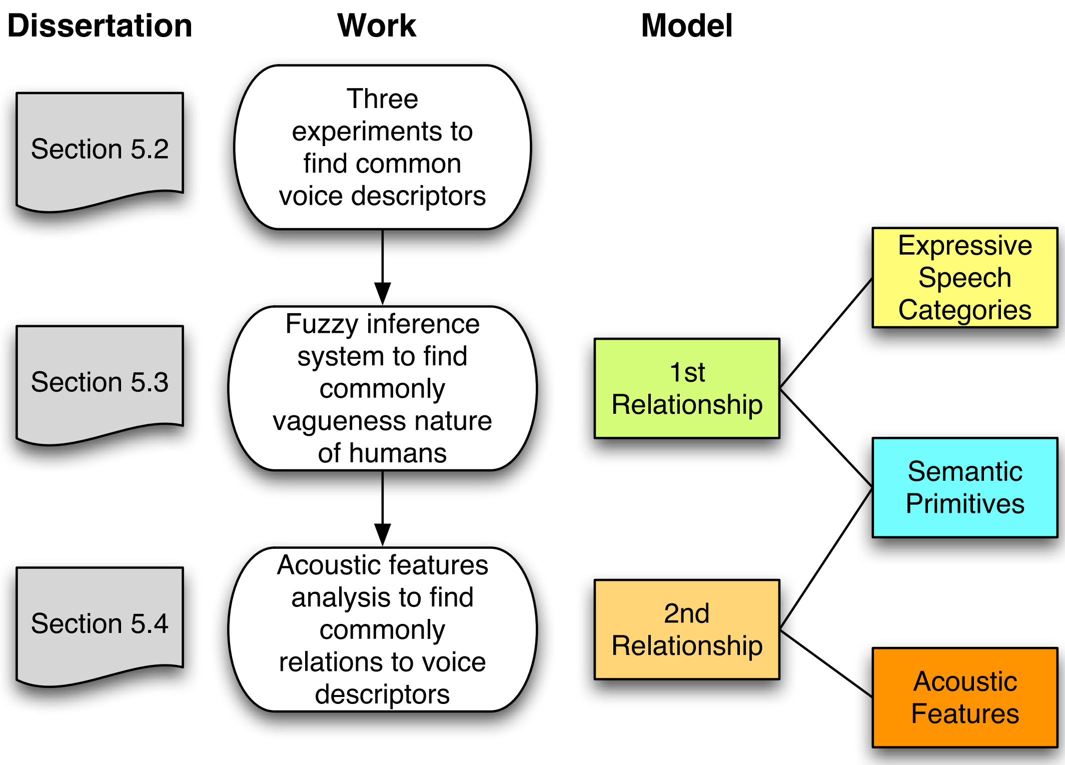

Dissertation
Work
Model
Section 5.2
Three experiments to find common voice descriptors
Section 5.3
Fuzzy inference system to find commonly vagueness nature of humans
Section 5.4
Acoustic features analysis to find commonly relations to voice descriptors
1st Relationship
2nd Relationship
Expressive Speech Categories
Semantic Primitives
Acoustic Features

Figure 2-5. Application process and the corresponding sections in the dissertation and in the model.

Chapter 3. The Building of the Three Layered Model

This chapter describes the first-step of the model construction. In this step, the suitable semantic primitives for voice description will be chosen. From here, the relationship between semantic primitives and acoustic features, and the relationship between semantic primitives and expressive speech categories will be clarified.

This research firstly uses a top-down method to build the model and then uses a button-up method to verify the model. The process is designed by this way because it is necessary to discovery related adjectives as semantic primitive for the perceptual model. And, as the definition of semantic primitive in this study, it should be suitable for describing expressive speech that is related to the perception of expressive speech. Therefore, it is reasonable to investigate the relationship between expressive speech and semantic primitive first and then the relationship between semantic primitive and acoustic features. When verifying the built model, because the method is rule-based speech morphing that controls acoustic features and resynthesized voice, conversely, the relationship between semantic primitive and acoustic features is verified first and then the relationship between expressive speech and semantic primitive.

The work consists of the following three sub tasks.

- In order to find those descriptors (e.g., adjectives) that can be used to describe the perception of expressive vocalizations, a thorough selection of semantic primitives will be necessary. Three experiments are carried out in order to determine which adjectives are suitable for describing expressive speech.

- To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech, a fuzzy inference system will be built. In order to build FIS with highly reliable probability, another experiment, Experiment 4, is conducted for collecting more utterances as input data.

- In order to support the relationship between semantic primitives and acoustic features, analysis of a large number of acoustic features will be necessary.

This order is important because without the finding of suitable semantic primitives for expressive category judgments, it is not possible to conduct analysis of acoustic features to semantic primitives.

However, another possibility is firstly a bottom-up approach to relate every possible semantic primitive to acoustic features, and then the selection of semantic primitives that are suitable for expressive speech categories description. This approach may sound *objective* at first because we consider every semantic primitive that many be used in voice description. However, some concerns listed below are considered:

1. Top-down approach corresponds to effective humans processing [86]. Our brain tries to process information effectively. Therefore, it is not likely that we humans enumerate different semantic primitives for every occurrence of expressive speech judgment and then select semantic primitives that are suitable for this judgment.

2. Top-down approach corresponds to humans' behavior. Humans' brains process information following a certain path [87]. Therefore, we humans use certain types of semantic primitives to describe expressive speech categories.

For the first task, detailed descriptions and discussions of those three experiments conducted for selecting appropriate semantic primitives are provided in Section 3.1. Those three experiments are (1) examination of listeners' perception of expressive speech, (2) construction of a psychological distance model, and (3) selection of semantic primitives.

For the second task, Section 3.2 starts with an explanation of the reasons of fuzzy logic application for building the relationship between expressive speech categories and semantic primitives. It is followed by an introduction to an experiment which is conducted to collect more data in order to build the fuzzy inference system. After the explanation of the fuzzy inference system (FIS) construction, those semantic primitives that are most related to the five categories (*Neutral*, *Joy*, *Cold Anger*, *Sadness*, and *Hot Anger*) of expressive speech will be clarified.

For the third task, Section 3.3 first provides a discussion with a brief literature review regarding acoustic cues related to expressive speech. The acoustic cues involved with prosody or pitch related parameters are mainly extracted from F0 contour, power envelop and duration. Conversely, voice quality related parameters are mainly extracted from the spectrum. Consequently, acoustic feature analysis done in this study were in terms of F0 contour, power envelop and duration, and spectrum. Detail introduction to the process of acoustic feature analysis is given in this section as well. The correlation coefficients between semantic primitives and acoustic features were then calculated for depicting the relationship.

After the completion of these three tasks, for each category, the two relationships are combined and are visualized by a layered structure conceptual diagram. From these five diagrams, the non-linguistic verbal information can be characterized by acoustic features of utterances and adjectives of voice description. Finally, the five diagrams are discussed.

3.1 Three experiments for semantic-primitive selection

To find semantic primitives that can be used to describe the perception of expressive vocalizations, three experiments are conducted. The goal of Experiment 1 is to examine subjects' perception of expressive speech utterances. The goal of Experiment 2 is to show the reliability of the voice database used by the subjects. The results of this experiment should show that subjects can clearly distinguish the expressive speech categories of each utterance they heard. The goal of Experiment 3 is to determine suitable semantic primitives for the perceptual model. To clarify which adjective is more appropriate for describing expressive speech and how each adjective was related to each category of expressive speech, the selected semantic primitives are superimposed into the perceptual space built in Experiment 2 by applying a multiple regression analysis.

To achieve each of these goals, in Experiment 1, subjects are asked to evaluate the perceived categories of utterances in terms of each of the five categories of expressive speech looked at in this study, e.g., *Neutral*, *Joy*, *Cold Anger*, *Sadness*, and *Hot Anger*. In Experiment 2, subjects are asked to evaluate utterances selected according to the results of Experiment 1. The results are used to construct a perceptual space of categories of expressive speech by using a multidimensional scaling (MDS) technique. (See also Maekawa and Kitagawa [47] for a similar approach.). Examination of the perceptual space assisted in selecting suitable adjectives as semantic primitives, which were then fine-tuned in Experiment 3 to arrive at a set of semantic primitives used in the model. In Experiment 3, subjects are asked to evaluate utterances by the semantic primitives they perceive. Finally, from the experimental results, fuzzy inference systems are built for representing the relationship between expressive speech and semantic primitives.

3.1.1 Experiment 1: Examination of Listeners' Perception of Expressive Speech Utterances

To examine subjects' perception of expressive speech utterances, Experiment 1 was conducted. This experiment looks at how listeners perceive all utterances in the voice corpus in terms of the five categories of expressive speech. According to subjects' ratings, the results show that most utterances can be easily perceived as belonging to their intended categories. The rating results were utilized in choosing appropriate utterances from a voice corpus concerning expressive speech categories in the subsequent experiments.

Stimuli

The stimuli are Japanese utterances produced by a professional actress. Stimuli were selected from the database produced and recorded by Fujitsu Laboratory. A professional actress was asked to produce utterances using five expressive speech categories, i.e., *Neutral*, *Joy*, *Cold Anger*, *Sadness*, and *Hot Anger*. In the database, there are 19 different Japanese sentences. Each sentence has one utterance in Neutral and two utterances in each of the other categories. Thus, for each sentence, there are 9 utterances and for all 19 sentences, there are 171 utterances. The detailed information of the corpus is shown in Table 3-1, Table 3-2, and Table 3-3.

Table 3-1. Specifications of Voice Data

Item	Value
Sampling frequency	22050 Hz
Quantization	16bit
Sentences	19

Table 3-2. 19 Sentences Used in Experiment 1. This table lists all sentences used in Experiment 1. First column shows the id numbers of the sentences. Second column shows the sentences in Japanese. Id 14 was not in the database.

Id	Sentence
1	新しいメールが届いています
2	頭にくることなんてありません
3	待ち合わせは青山らしいんです
4	新しい車を買いました
5	いらないメールがあったら捨てて下さい
6	そんなの古い迷信です。
7	みんなからエールが送られたんです。
8	手紙が届いたはずです。
9	ずっとみています。
10	私のところには届いています。
11	ありがとうございました
12	申し訳ございません
13	ありがとうは言いません
15	気が遠くなりそうでした。
16	こちらの手違いもございました。
17	花火を見るのにゴザがいりますか。
18	もうしないと言ったじゃないですか。
19	時間通りに来ない訳を教えてください。
20	サービスエリアで合流しましょう。

Table 3-3. 171 Utterances Used in Experiment 1. First column shows the utterances id numbers. Second column shows the intended emotion category. The UID is composed of a letter and a numerical code. The letter represented what emotion it is (a: Neutral, b, c: Joy, d, e: Cold Anger, f, g: Sadness, h, i: Hot Anger) and the numerical code presented what sentence it was in a014, b014, c014, d014, e014, F014, g014, h014, and i014 were not used.

UID	Expressive Speech Category
a001 ~ a020	Neutral
b001 ~ b020	Joy
c001 ~ c020	
d001 ~ d020	Cold Anger
e001 ~ e020	
F001 ~ F020	Sadness
g001 ~ g020	
h001 ~ h020	Hot Anger
i001 ~ i020	

Subjects

12 graduate students, native male Japanese speakers, average age 25 years, with normal hearing participated in the experiment.

Environment and equipment:

This experiment was conducted in a sound proof room. The detailed information of the other equipment is listed in Table 3-4 and the configuration of the equipment is depicted in Figure 3-1.

Table 3-4. Equipment of Experiment 1

Equipment	Specification
Server for sound presenting	DAT + LINK & Work Station (Linux)
Headphone	STAX SR-404
Headphone amp	STAX SRM-1/MK-2
D/A converter	STAX DAC-TALENT BD.

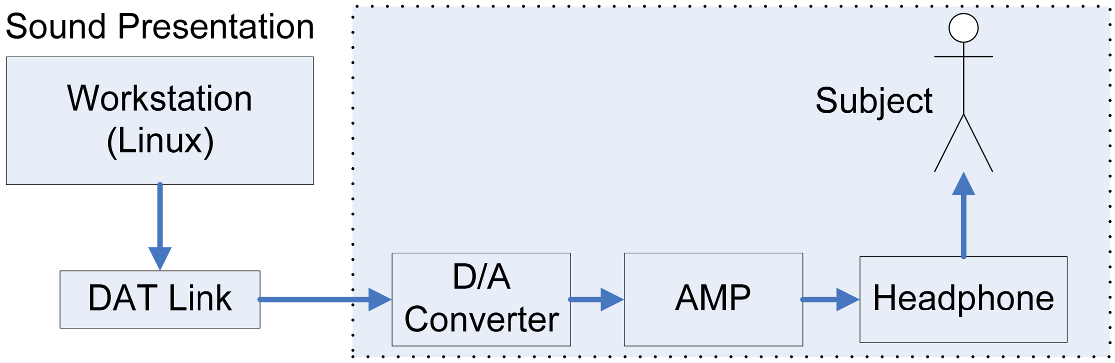


Figure 3-1. Configuration of the Equipment Used in All Experiments of this Research Work

Method

Subjects were asked to rate the 171 utterances according to the perceived degree of each of the 5 expressive speech categories. An example of the questionnaire is shown in Table 3-5. There was a total of 5 points for one utterance. That is, if a subject perceived that an utterance belonged to one expressive speech category without any doubt, then the subject gave it 5 points. Conversely, if the subject was confused within two or even more expressive speech categories, then the subject divided the 5 points among these expressive speech categories according to what seemed appropriate. The stimuli were randomly presented to each subject through binaural headphones at a comfortable sound pressure level in a soundproof room. Each utterance was presented twice followed by a pause of 2 seconds.

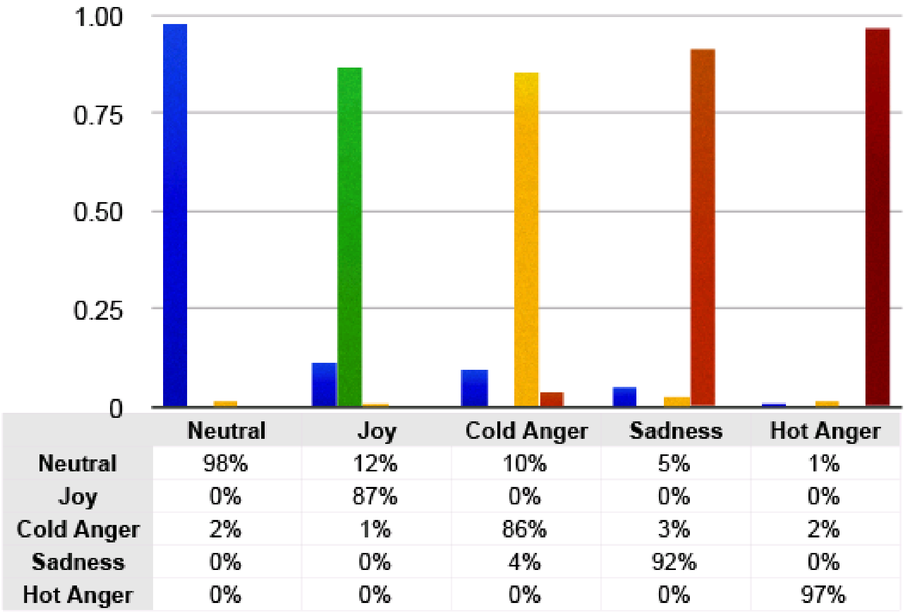
Table 3-5. Evaluation form used in Experiment 1.

<table>
<tr><td colspan="5">Sentence 1
新しいメールが届いています
(A new e-mail has arrived)</td></tr>
<tr><td>N</td><td>J</td><td>CA</td><td>S</td><td>HA</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
</table>

Results and discussion

Table 3-6 shows the confusion matrix of each intended category. The results show that most utterances can be easily perceived as belonging to their intended categories. However, *Cold Anger* (CA) had the lowest percentage and was easily confused with *Neutral* (N). The one most confused with *Neutral* (N) was *Joy* (J). The rating result will be used when considering which utterances should be selected as stimuli in the following experiments.

Table 3-6. Percentage of ratings of the 5 intended categories. The columns are the intended categories and the rows, the categories rated by subjects.



	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	98%	12%	10%	5%	1%
Joy	0%	87%	0%	0%	0%
Cold Anger	2%	1%	86%	3%	2%
Sadness	0%	0%	4%	92%	0%
Hot Anger	0%	0%	0%	0%	97%

3.1.2 Experiment 2: Construction of a Psychological Distance Modal

To find what adjectives should be semantic primitives in the model, Experiments 2 and 3 were conducted. Experiment 2 is conducted to construct a psychological distance model of utterances in the different expressive speech categories by applying a multidimensional scaling technique. A psychological distance model is considered as a perceptual space that illustrated the similarity among stimuli. Experiment 3 is to select suitable semantic primitives for the proposed model from the resulting perceptual space built in Experiment 2. It is described in Section 3.1.3.

Stimuli

15 utterances were chosen according to the ratings in Experiment 1. In order to expand the perceptual space of expressive speech, for each of the five categories, three utterances were selected: (1) one that was least confused, (2) one that was most confused, (3) one that fell in the middle.

Subjects

They are identical to those in Experiment 1.

Environment and equipment setting

They are identical to those in Experiment 1.

Method

Experiment 2 applied Scheffe's method of paired comparison. Subjects were asked to rate each of the utterance pairs on a 5-point Liker-type scale (from -2 to 2, including 0, -2 = totally different, 2 = extremely similar) according to how similar they perceived them to be. The pair-wise stimuli were randomly presented to each subject through binaural headphones at a comfortable sound pressure level in a soundproof room. Each utterance pair was presented followed by a gap of 2 sec. The SPSS 11.0J for Windows MDS ALSCAL procedure, non-metric model of Shepard and Kruskal, using the symmetric, matrix conditional, and ordinal options, was applied to the ratings.

Results and discussion:

Figure 3-2 shows the distribution of utterances in the resulting 3-dimensional perceptual space (STRESS value was 7%). In the figure, one circle represents one utterance and plot symbols like 'J' indicate utterances of *Joy*, etc. The number after each symbol corresponds to the select utterances (1), (2), and (3) explained in Stimuli above. As the distribution shows, all categories of expressive speech are separated clearly; moreover, utterances of the same category are close to each other. The distribution in perceptual space indicates which expressive speech utterances are similar, and exactly how similar they are. This information can be used to determine the semantic primitives suitable for Experiment 3.

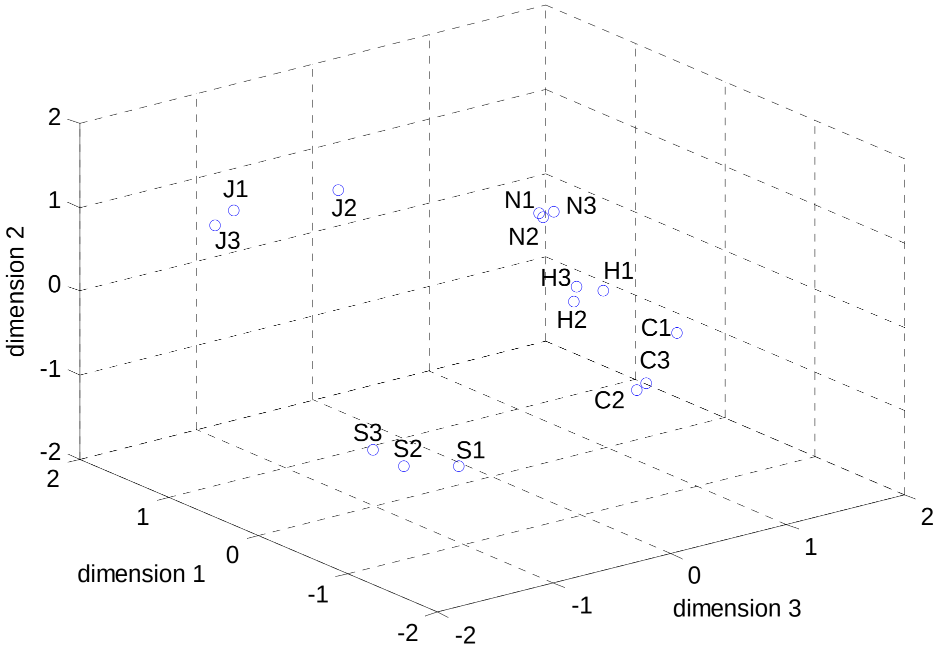


Figure 3-2. The resulting perceptual space of utterances in different categories of expressive speech. One circle represents one utterance. The number after each symbol represents the selected utterances.

3.1.3 Experiment 3: Semantic Primitives Selection

Experiment 3 is to select suitable semantic primitives for the proposed model from the resulting perceptual space built in Experiment 2.

Pre-experiments

As described above, semantic primitives are defined as adjectives appropriate for describing speech. In order to determine adjectives related to expressive speech from a large number of possible adjectives applicable to sound, tone, or voice, we carried out this pre-experiment.

Sixty adjectives were selected as candidates for semantic primitives. 46 of these were from the work by Ueda [76] (with English glosses), in which he asked 166 listeners to choose adjectives which are often used to describe voice timbre. However, there were 4 adjectives removed from his original list that we considered less-related to expressive speech perception. Since the original adjectives are for music description, an extra 14 adjectives considered relative to expressive speech were added based on informal perception tests. In the pre-experiment, subjects listened to 25 utterances (five for each category of expressive speech which were randomly chosen from the expressive speech database) and were asked to circle which adjectives seemed most appropriate to describe each utterance they heard. The 34 adjectives listed in Table 3-7 were the ones most frequently circled, and were thus chosen for Experiment 3.

Table 3-7. 34 Adjectives Chosen from the Pre-Experiment. Experiment 3 selects the 17 adjectives with character shading as semantic primitives for the perceptual model. The third column shows the correlation coefficients that are calculated for selecting suitable semantic primitives.

ID	Adjective (Japanese)	Adjective (English)	Correlation Coefficient
1	明るい	bright	0.979
2	暗い	dark	0.968
3	声の高い	high	0.95
4	声の低い	low	0.897
5	強い	strong	0.989
6	弱い	weak	0.936
7	太い	thick	0.927
8	細い	thin	0.872
9	堅い	hard	0.977
10	柔らかい	soft	0.966
11	重い	heavy	0.95
12	軽い	light	0.961
13	鋭い	sharp	0.962
14	鈍い	dull	0.93
15	耳障りな	rough	0.906
16	流暢な	fluent	0.866
17	荒っぽい	violent	0.944
18	滑らかな	smooth	0.885
19	うるさい	noisy	0.895
20	静かな	quiet	0.984
21	ざわついた	raucous	0.881
22	落ち着いた	calm	0.981
23	落ち着きのない	unstable	0.89
24	きれいな	clean	0.911
25	汚い	dirty	0.91
26	濁った	muddy	0.89
27	明らかな	clear	0.975
28	あいまいな	vague	0.931
29	明瞭な	plain	0.952
30	かすれた	husky	0.898
31	抑揚のある	well-modulated	0.956
32	単調な	monotonous	0.971
33	早い	fast	0.904
34	ゆっくり	slow	0.605

Stimuli

They were the same as in Experiment 2.

Subjects

They were the same as in Experiment 1.

Environment and equipment

They were the same as in Experiment 1.

Method

The stimuli were randomly presented to each subject through binaural headphones at a comfortable sound pressure level in a sound proof room. Subjects were asked to rate each of the 34 adjectives on a 4-point scale (0: very appropriate, 3: not very appropriate) when they heard each utterance, and to indicate how appropriate the adjective was for describing the utterance they heard. In order to clarify which adjectives were more appropriate for describing expressive speech and how each adjective was related to each category of expressive speech, the 34 adjectives were superimposed by the application of a multiple regression analysis into the perceptual space built in Experiment 2. Equation (1) is the regression equation.

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 \quad (1)$$

, where x_1, x_2 and x_3 are the positions (x_1, x_2, x_3) of one utterance in the 3 -dimensional perceptual space, and y is the rating of an adjective for a particular utterance. Regression coefficients a_1, a_2 and a_3 were calculated by performing a least squares fit. In addition, the multiple correlation coefficient of each adjective was computed.

Result and discussion

Figure 3-3 is the diagram that presents the 34 adjectives plotted in *dimension-1* against *dimension-2*, *dimension-1* against *dimension-3*, and *dimension-3* against *dimension-2* of the 3D perceptual space. The utterances are represented by the same IDs as in Figure 3-2 and a line in the plot indicates an adjective by marking its ID number which is listed in Table 3. The direction of the arrowhead of each line indicates that the adjective was increasingly related to the utterances. For example, the adjective clean (ID: 24) was more related to the utterance J2 (*Joy*) than to the utterance C2 (*Cold*

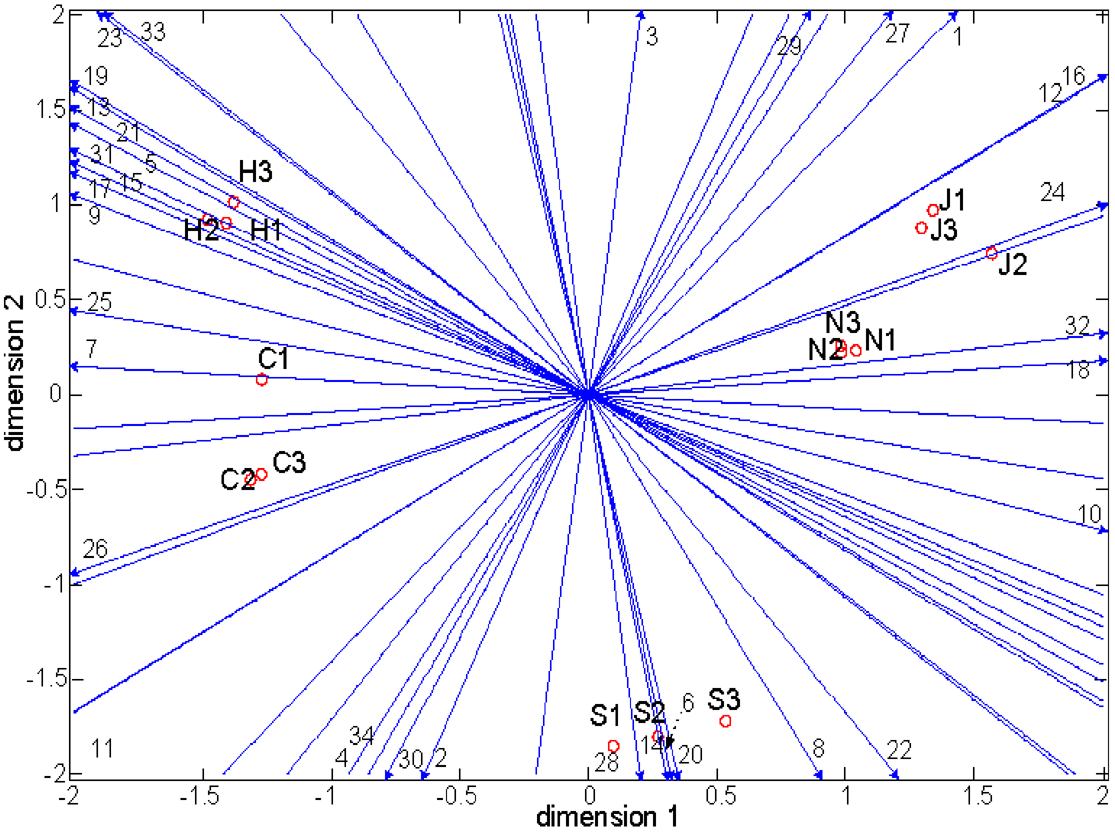
Anger). In this way, it was possible to find to which category each adjective was related. Semantic primitives were selected according to the following three criteria:

(1) The direction of each adjective in the perceptual space: This indicates which category the adjective is most related to. For example, the adjective *Monotonous* (32) is most closely related to the category *Neutral*.

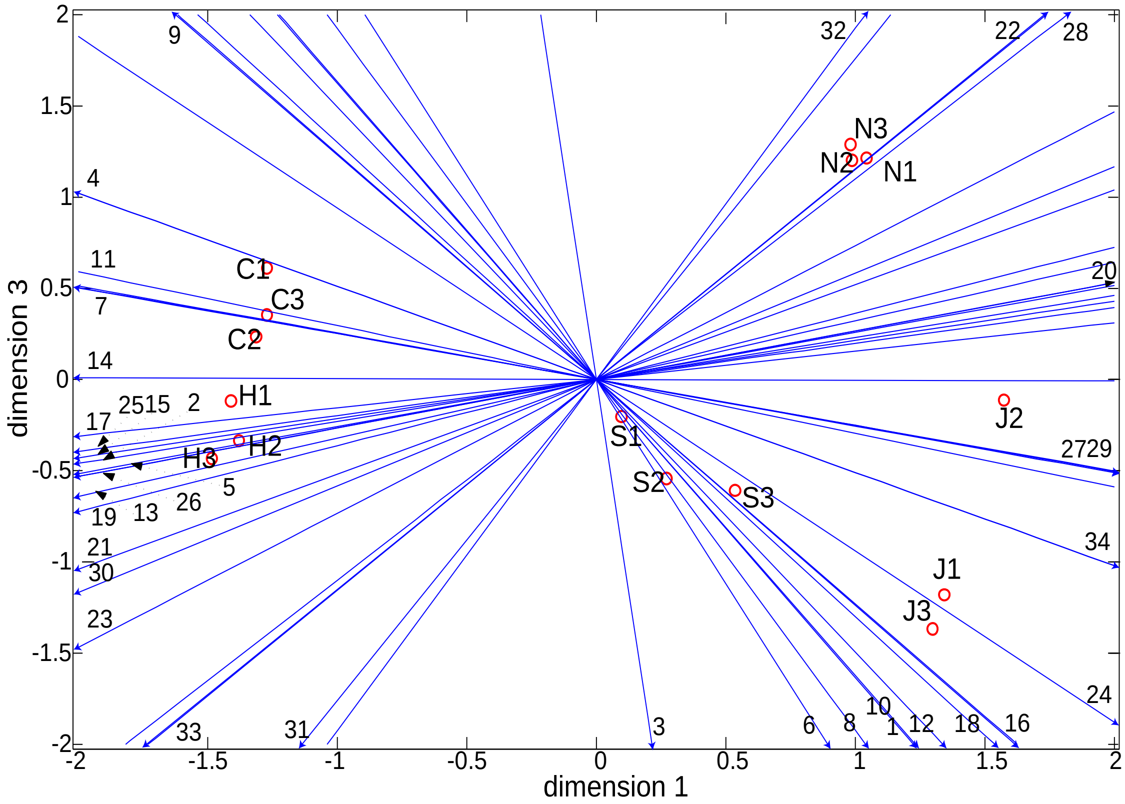
(2) The angle between each pair of adjectives: The smaller the angle is, the more similar the two adjectives.

(3) The multiple correlation coefficient of each adjective: When the multiple correlation coefficient of one adjective is higher, it means the adjective is more appropriate for describing expressive speech.

Candidate adjectives were chosen according to criteria (1) and (2). Criterion (3) was used to decide the final list of 17 adjectives which were chosen as semantic primitives: *bright*, *dark*, *high*, *low*, *strong*, *weak*, *calm*, *unstable*, *well-modulated*, *monotonous*, *heavy*, *clear*, *noisy*, *quiet*, *sharp*, *fast* and *slow* (see Table 3-6). These semantic primitives reflect a balanced selection of widely-used adjectives that describe expressive speech.

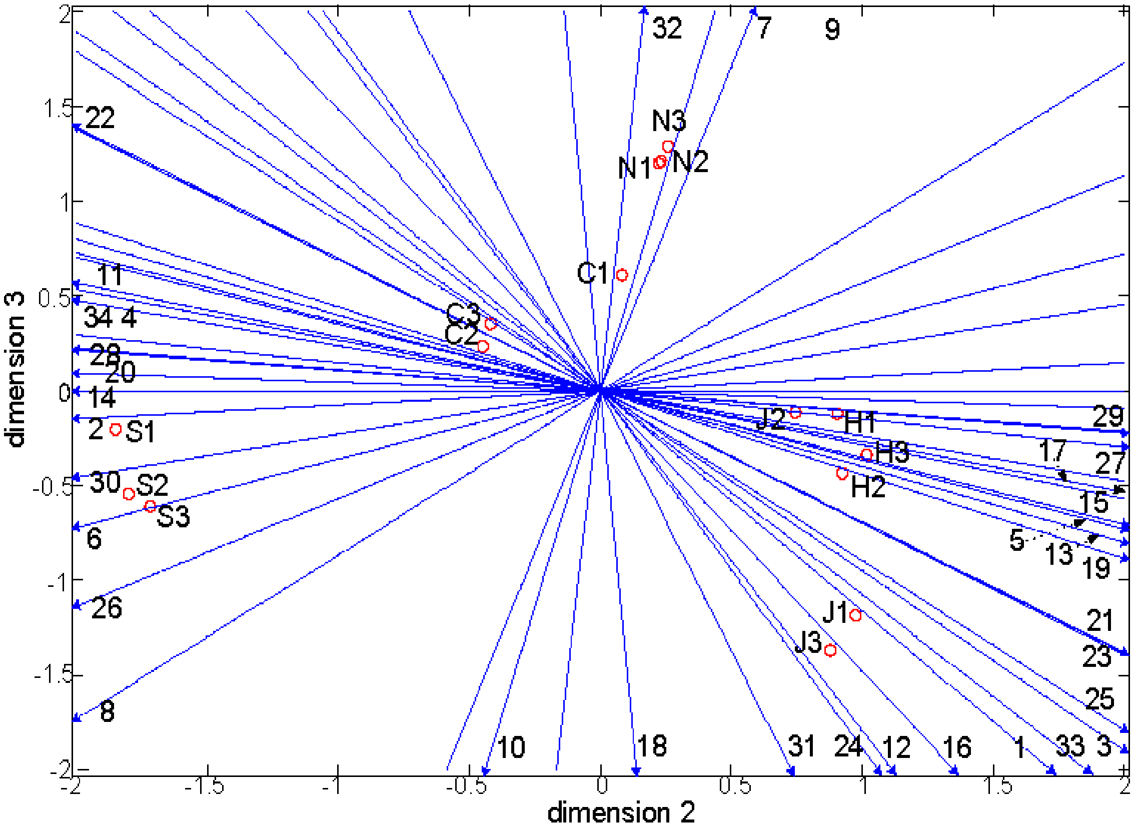

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
dimension 2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
dimension 1
23
33
3
29
27
1
19
13
21
31
5
17
15
9
H3
H2
H1
16
12
24
J1
J3
J2
N3
N2
N1
25
7
C1
32
18
C3
C2
10
26
11
34
4
30
2
S1
S2
6
S3
28
14
20
8
22

(a)



dimension 3
dimension 1
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
9
32
22
28
N3
N2
N1
4
11
C1
C3
7
C2
20
14
H1
2515
2
17
J2
H2
H3
2729
S1
S2
S3
5
19
13
26
34
21
30
J1
J3
23
24
33
31
3
6
8
10
1
12
18
16

(b)



(c)

Figure 3-3. Direction of adjectives in perceptual space. The figure was plotted with arrow-headed lines in (a) dimension-1 against dimension-2, (b) dimension-1 against dimension-3, and (c) dimension-2 against dimension-3 of the perceptual space in Figure 3-2.

3.2 Fuzzy inference system

To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech, a fuzzy inference system (FIS) will be built. In order to build FIS with highly reliable probability, another experiment, Experiment 4, was conducted for collecting more utterances as input data. Before the description of FIS building, the reason of using fuzzy logic is explained first.

3.2.1 Why apply fuzzy logic?

This research suggests that the relationship between expressive speech category and semantic primitive represents the way humans use linguistic forms (e.g., words) to describe what they perceive when they hear expressive speech. That expression of the perception is vague, not precise. Since statistical methodology mainly manages linear and precise relationships but not fuzzy relationships, fuzzy logic is a more appropriate mathematical tool for describing this non-linear, vague, and white-box relationship.

The reasons are the following:

- Fuzzy logic embeds existing structured human knowledge (experience, expertise, heuristics) into workable mathematics [39], and this is exactly what the model proposes to do in dealing with the perception of expressive speech.
- Fuzzy logic is based on natural language [35], and the natural language used in our model is in the form of semantic primitives.
- Fuzzy logic models nonlinear functions of arbitrary complexity [83], and the relationship between expressive speech categories and semantic primitives is certainly complex and nonlinear.

Thus, fuzzy logic would be able to address what is the relationship between, for example, the perception of “slightly slow“ /“slow” /“very slow” and a sad vocal expression.

3.2.2 Experiment 4: Data collection for FIS building

In this experiment, more utterances are collected as input data for FIS building. In order to create a set of well-balanced data so that the resulting relationship not only expresses the most-well-perceived categories but also provides the different intensity levels of categories, seven utterances were selected for each of the five categories from the 171 utterances evaluated in Experiment 1: two of them had the highest rating of “5”; three of them had the middle rating of “3”, and two of them had the lowest rating of “1”, which totaled to 35 utterances as the stimuli for this experiment.

The subjects and environment setting were identical to those in the previous experiments. Subjects were asked to rate each of the 17 adjectives, which were listed in Table 3-7, on a 4-point scale (0: very appropriate, 3: not very appropriate) when they heard each utterance, and to indicate how appropriate the adjective was for describing the utterance they heard.

3.2.3 Construction of FIS

By applying adaptive neuro-fuzzy [10], a technique within the Fuzzy Logic Toolbox of MATLAB, we built a fuzzy inference system (FIS) for each category of expressive speech. As was expected by FIS, input data were the perceptible degrees of semantic primitives and output data were the perceptible degrees of expressive speech categories. 50 utterances are used as the input data. 15 out of 50 utterances were collected from Experiment 1 and Experiment 3, as mentioned in Sections 3.1.1 and 3.1.3 and another 35 utterances are from Experiment 4.

The following steps were used to analyze the experimental data in order to construct the FIS.

- Step 1: To construct an initial FIS model.

  The purpose of this step was to construct a raw model by applying the method of subtractive clustering to analyze the experimental results. This decided the number of membership functions for each input variable and rules for FIS, so that it could be trained and adjusted. The training data used in this step are 40 of the 50 utterances.

  An initial FIS model, a first-order Sugeno type model [72] with 3 membership functions for each variable was constructed. This corresponds to a 3-level-rating system (low, mid, and high) which appears to be sufficient for describing intensity of a semantic primitive.

- Step 2: To train the initial FIS by adaptive neuro-fuzzy methodology.

  To improve capability of the model by adjusting membership function parameters, the adaptive neuro-fuzzy technique was used here. This is a hybrid method consisting of back-propagation for the parameters associated with the input membership functions, and least squares estimation for the parameters associated with the output membership functions. This step generated a trained FIS model with 3 membership functions for each variable, with a 3-rule structure, and still it was a first-order Sugeno type model.

- Step 3: To train the resultant FIS of step 3 by adaptive neuro-fuzzy methodology.

  The purpose of the step was to make the refined model fit the target system but not overfit the training data. The checking data used in this step are another 10 utterances of the 50 utterances. Eventually, for each expressive speech category, a final FIS to describe the relation between the expressive speech category and the 17 semantic primitives was completed.

3.2.4 Evaluation

In order to evaluate the relationship built by FIS, we calculated regression lines that describe the relationship between input (perceptible degrees of semantic primitives) and output (perceptible degrees of expressive speech) of each FIS. Therefore, the slope of the regression line explains the relationship between expressive speech and semantic primitive. The absolute value of the coefficient of the regression line indicates how much the semantic primitives affect the categories of expressive speech. A positive value of slope indicates that the relationship has a positive correlation, and vice versa.

For example, Figure 3-4 shows one of the results. The solid line is the FIS output and the dotted line is the regression line of the output. The figure depicts a non-linear relationship between *Neutral* and the semantic primitive *Monotonous* on the left and another non-linear relationship between *Hot Anger* and *Monotonous* on the right.

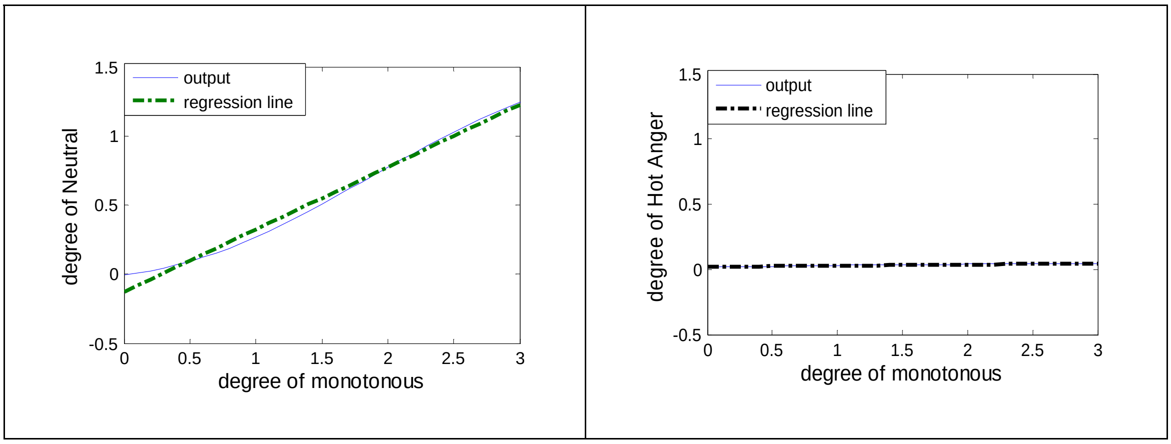


Figure 3-4. Slope of regression line. Left graph describes the relationship between *monotonous* and *Neutral*, right graph describes the relationship between *monotonous* and *Hot Anger*.

Table 3-8. 5 Semantic primitives that are most related to each category of expressive speech. SP column lists the semantic primitives and S column lists the slope of the regression line as described in Figure 3-4.

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
SP	S	SP	S	SP	S	SP	S	SP	S
monotonous	0.270	bright	0.101	heavy	0.197	heavy	0.074	w ell-modulated	0.124
clear	0.127	unstable	0.063	w ell-modulated	0.091	w eak	0.065	unstable	0.120
calm	0.103	clear	0.034	low	0.090	quiet	0.057	sharp	0.103
heavy	-0.329	quiet	-0.039	slow	-0.231	strong	-0.049	calm	-0.063
w eak	-0.181	w eak	-0.036	clear	-0.062	sharp	-0.079	quiet	-0.047

3.2.5 Converging of FIS

Figure 3-5 and Figure 3-6 plot the root mean squared error (RMSE) of the training data set at each epoch and RMSE of the checking data set at each epoch when building FIS of Neutral. Figure 3-7 and Figure 3-8 plot the information for FIS of Joy. Figure 3-9 and Figure 3-10 plot the information for FIS of Cold Anger. Figure 3-11 and Figure 3-12 plot the information for FIS of Sadness. Figure 3-13 and Figure 3-14 plot the information for Hot Anger.

For FIS of *Natural*, *Joy*, *Sadness*, and *Hot Anger* (see Figures 3-5, 3-6, 3-7, 3-8, 3-11, 3-12, 3-13 and 3-14), both training error and checking error can be reduced to the lowest at 200 epochs and converged to a settled state finally. That suggests that the generated FIS is reliable.

For FIS of *Cold Anger* (see Figures 3-9 and 3-10), the smallest value of the checking data error occurs at the 115th epoch, after which it increases slightly. The minimal error against the training data locates at the 115th epoch, after which it increases slightly and then converges to a settled state finally. This result suggests perhaps there is an overfit of the system to the training data. Overfitting occurs when the fuzzy system is trained to fit the training data so well that it no longer does a very good job of fitting the checking data. To avoid this defeat, when building the FIS of Cold Anger, the epoch is fixed to 115 which makes the training error and the checking error smallest.

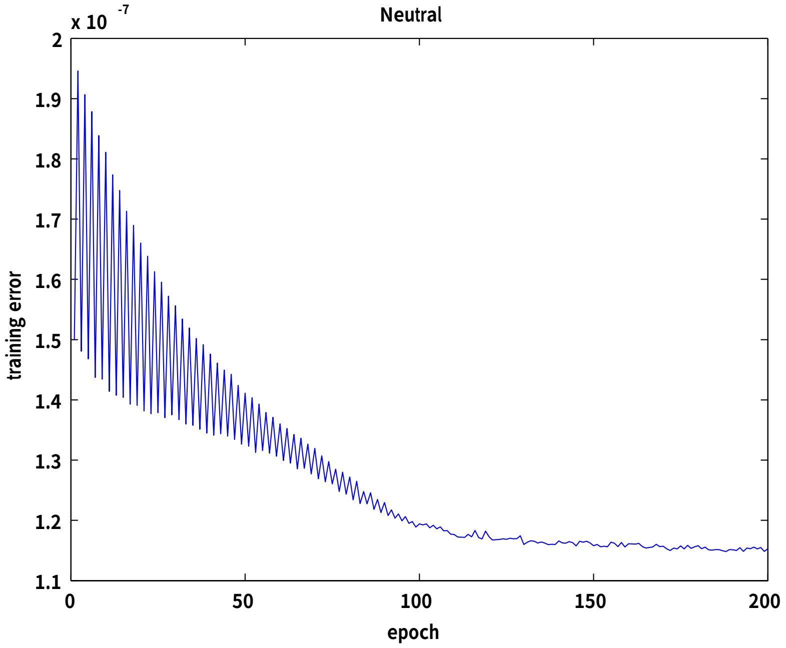


Figure 3-5. Root mean squared error (RMSE) of the training data set of Neutral

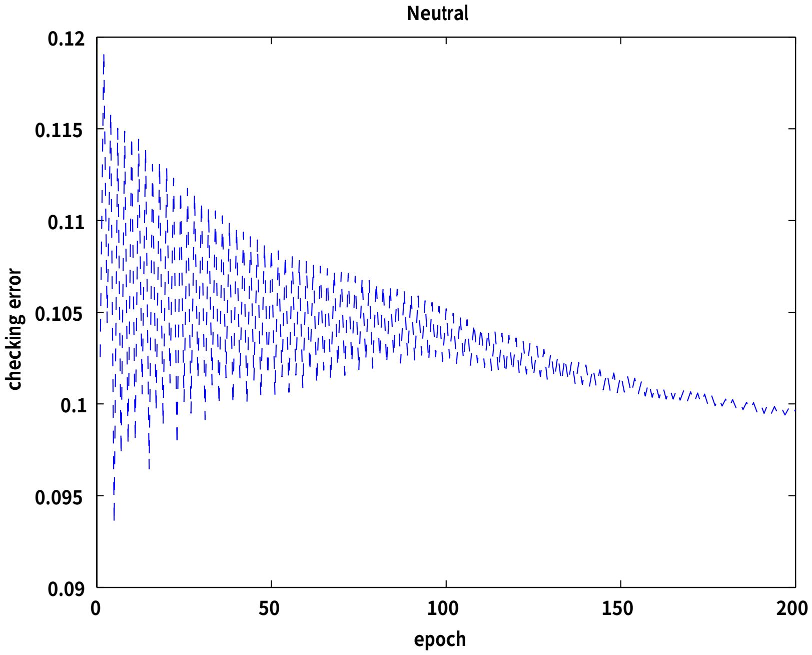


Figure 3-6. Root mean squared error (RMSE) of the checking data set of Neutral

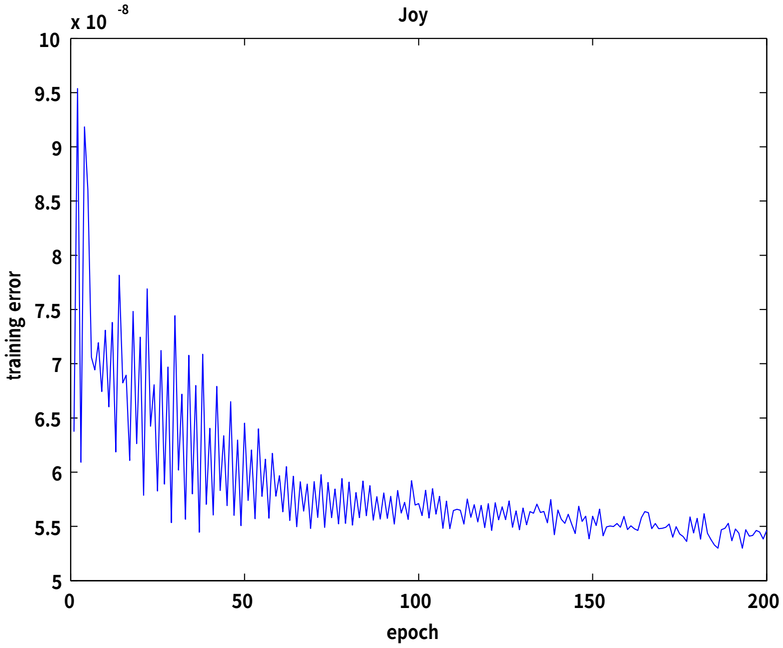


Figure 3-7. Root mean squared error (RMSE) of the training data set of Joy

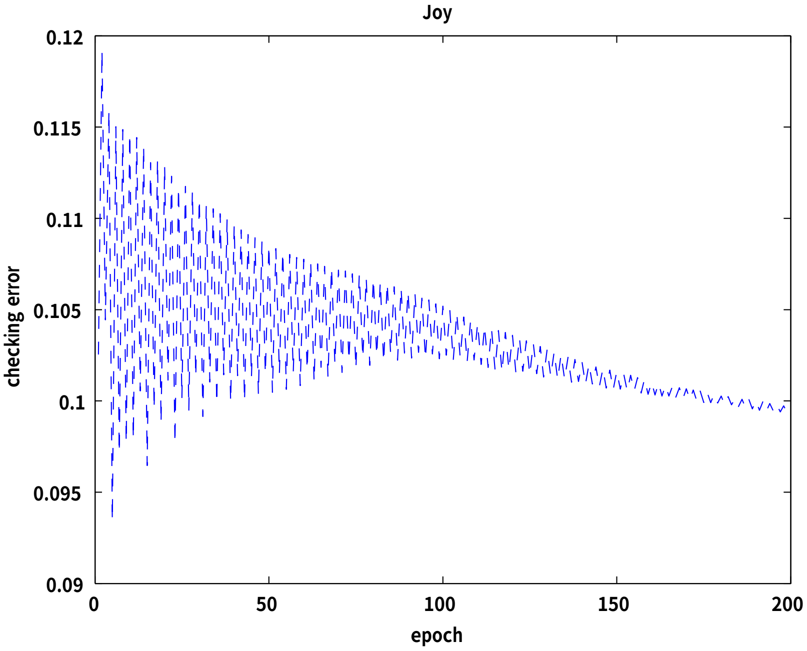


Figure 3-8. Root mean squared error (RMSE) of the checking data set of Joy

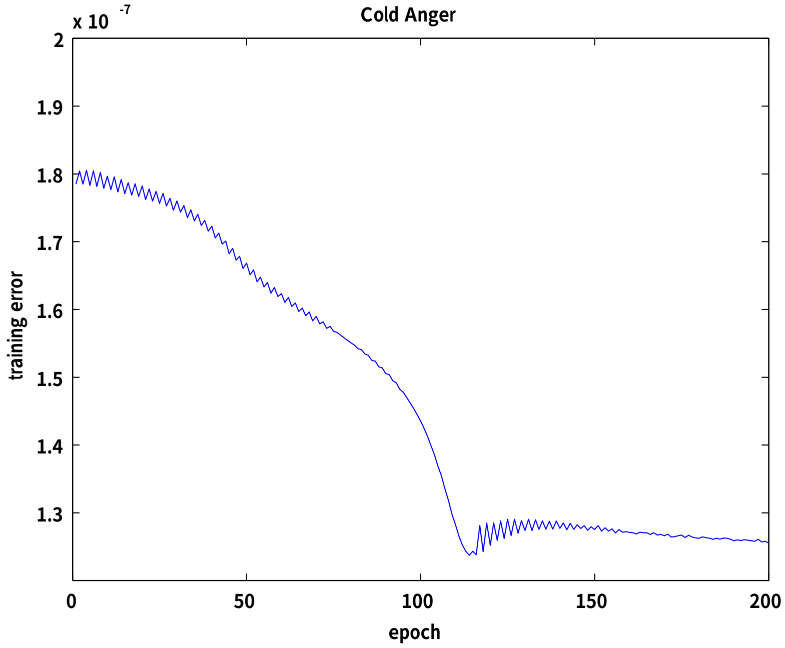


Figure 3-9. Root mean squared error (RMSE) of the training data set of Cold Anger

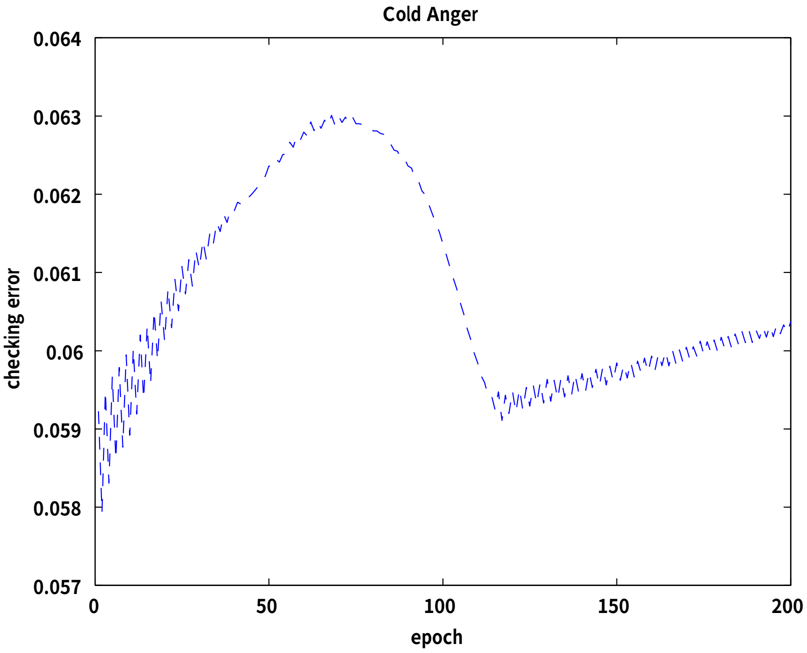


Figure 3-10. Root mean squared error (RMSE) of the checking data set of Cold Anger

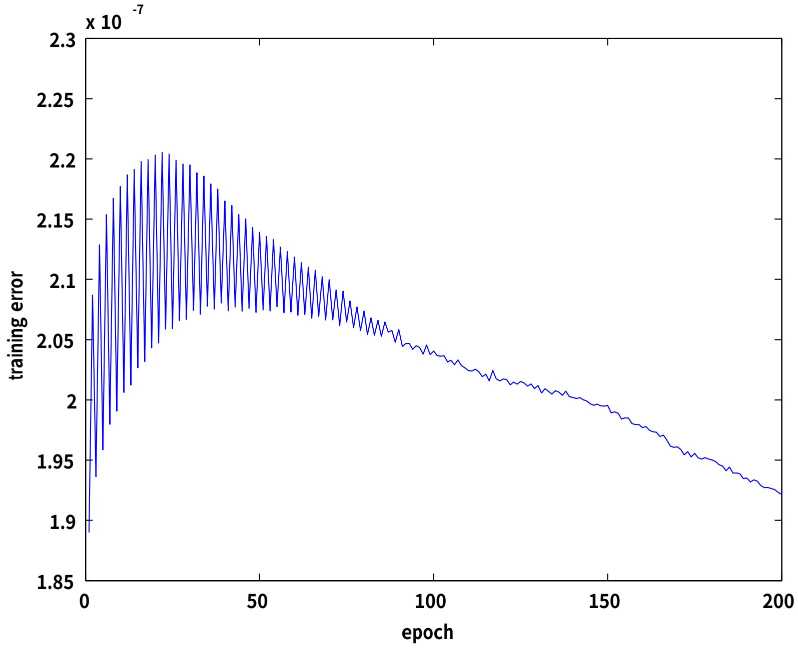


Figure 3-11. Root mean squared error (RMSE) of the training data set of Sadness

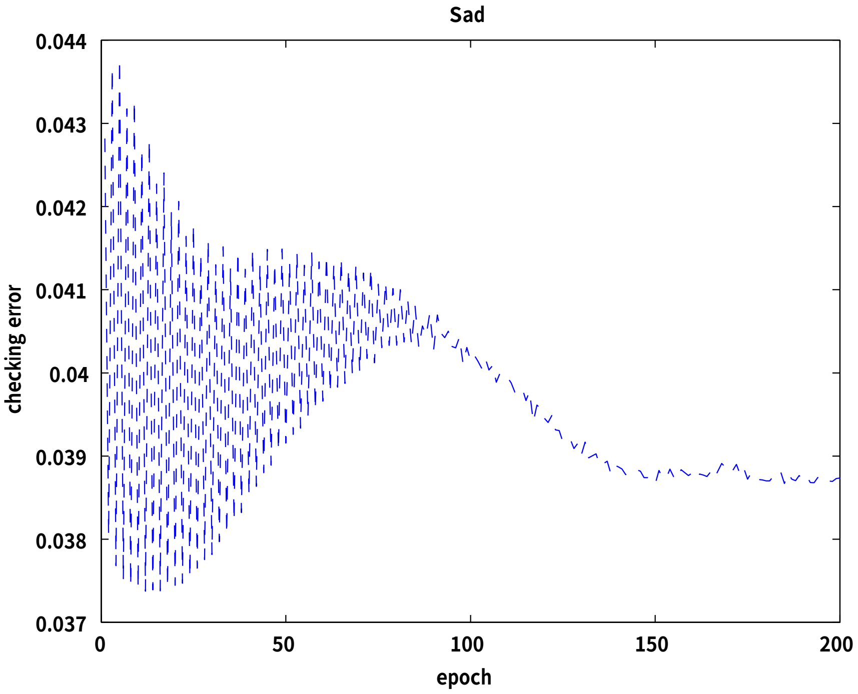


Figure 3-12. Root mean squared error (RMSE) of the checking data set of Sadness

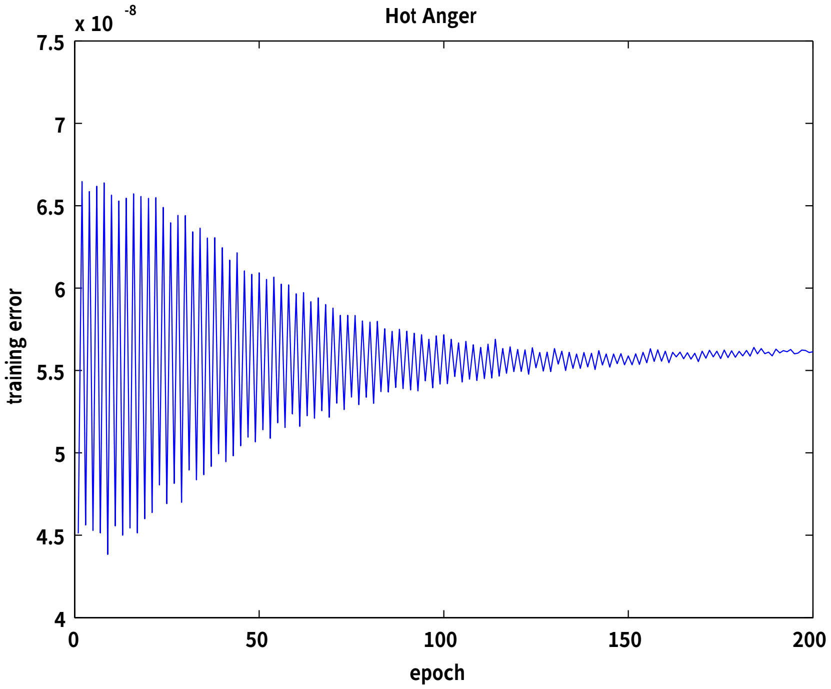


Figure 3-13. Root mean squared error (RMSE) of the training data set of Hot Anger

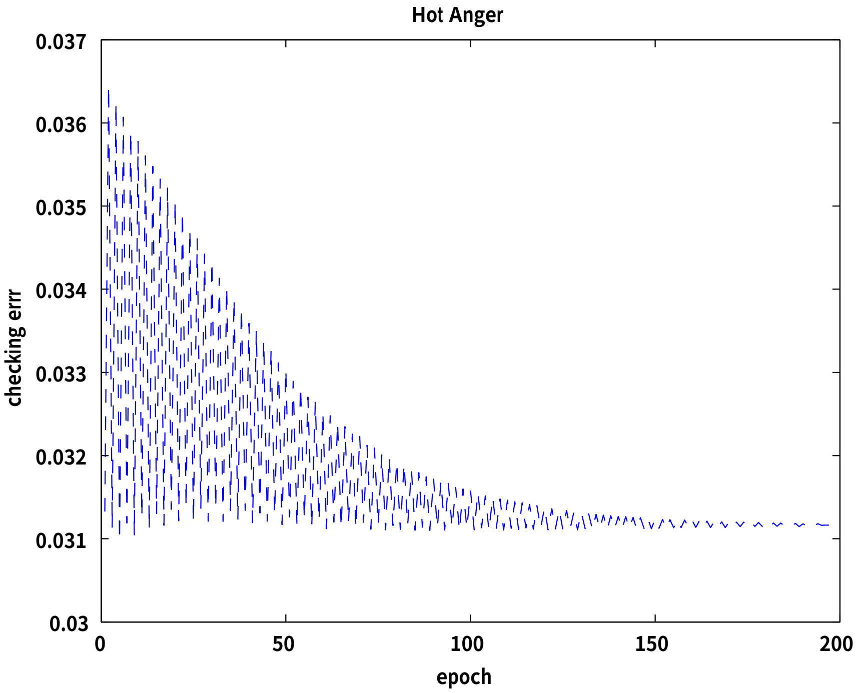


Figure 3-14. Root mean squared error (RMSE) of the checking data set of Hot Anger

Figures 3-15 to 3-19 show the root mean squared error (RMSE) of the 50 sentences that were used to train and check FIS. The percentage of error smaller than 0.3 is 70% for Neutral, 68% for Joy, 68% for Cold Anger, 92% for Sad, and 64% for Hot Anger. They also suggest reliability of the built FIS.

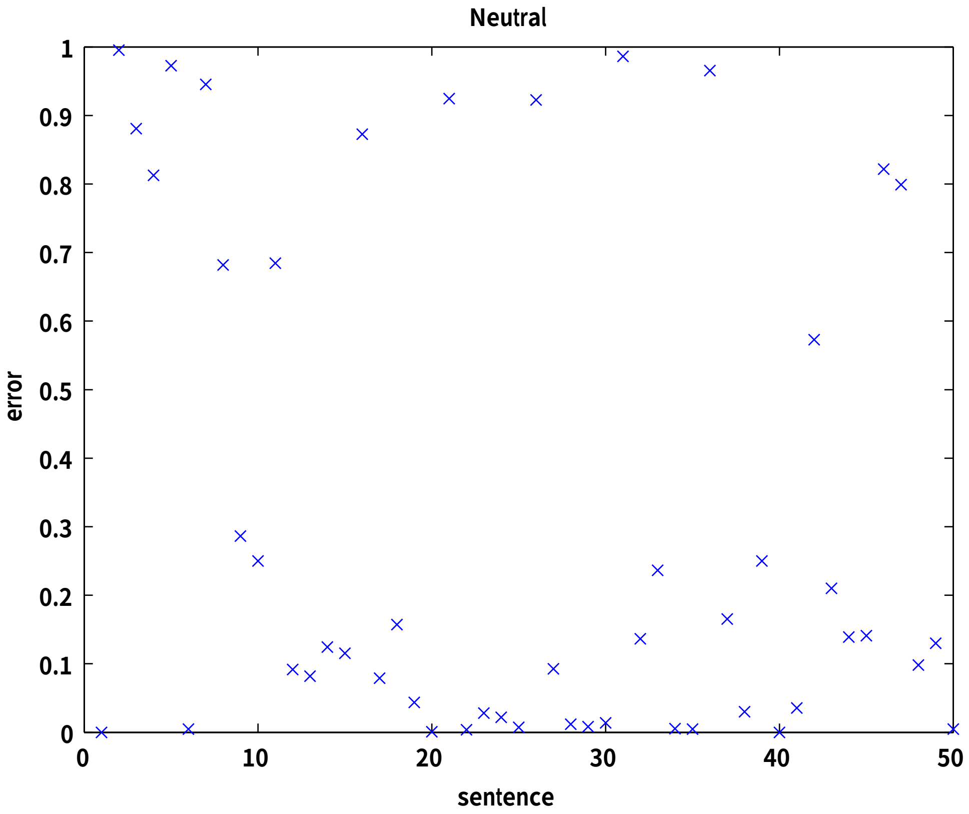


Figure 3-15. Root mean squared error (RMSE) of the 50 sentences of Neutral that used to train and check FIS

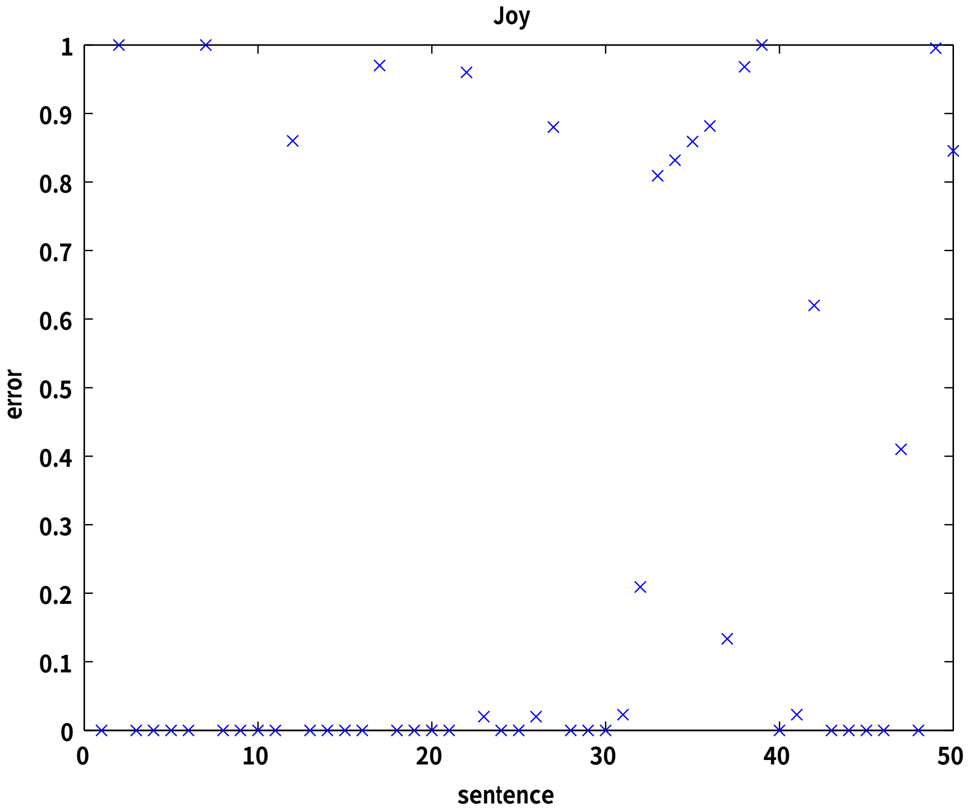


Figure 3-16. Root mean squared error (RMSE) of the 50 sentences of Joy that used to train and check FIS

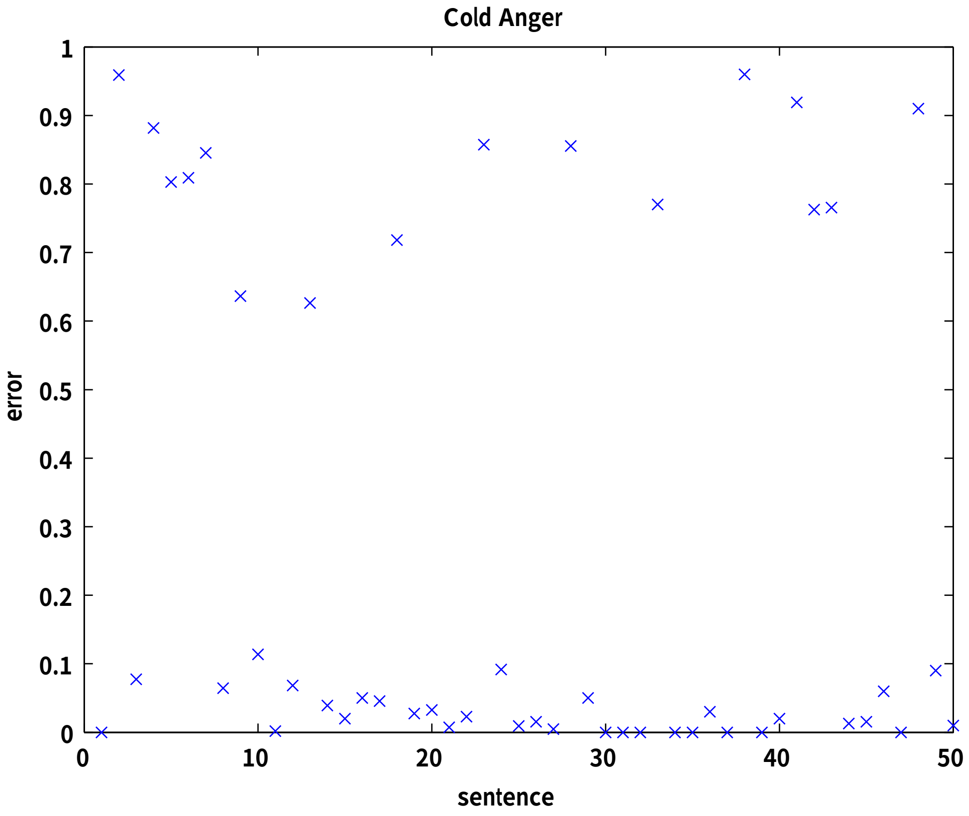


Figure 3-17. Root mean squared error (RMSE) of the 50 sentences of Cold Anger that used to train and check FIS

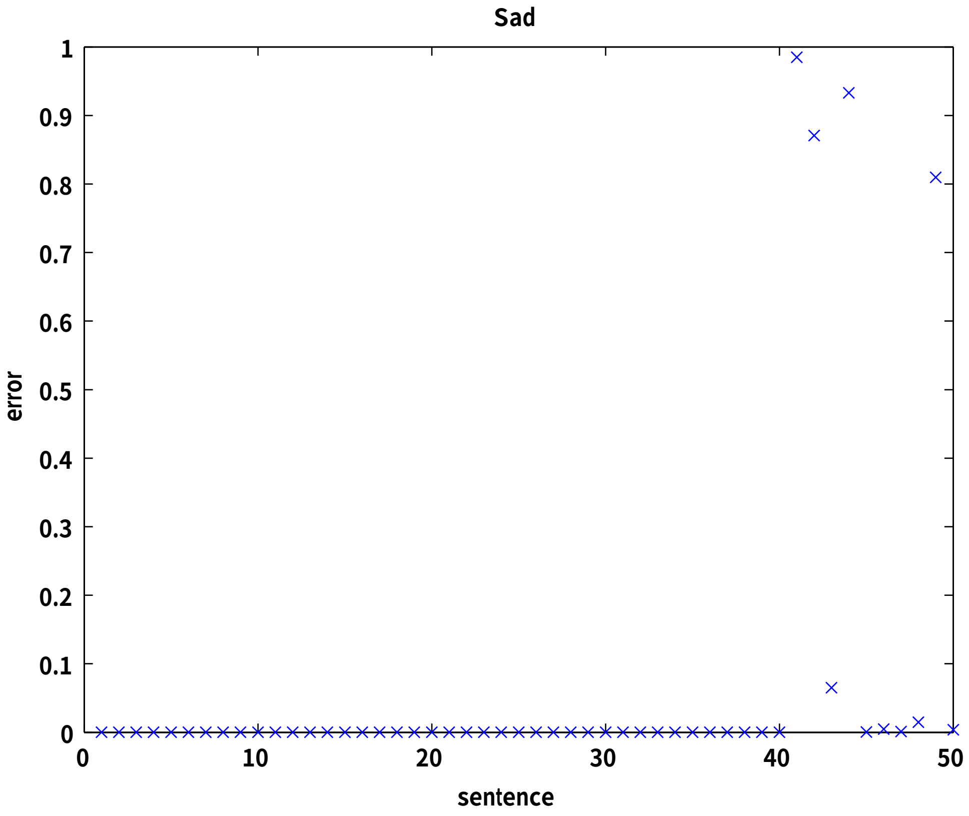


Figure 3-18. Root mean squared error (RMSE) of the 50 sentences of Sad that used to train and check FIS

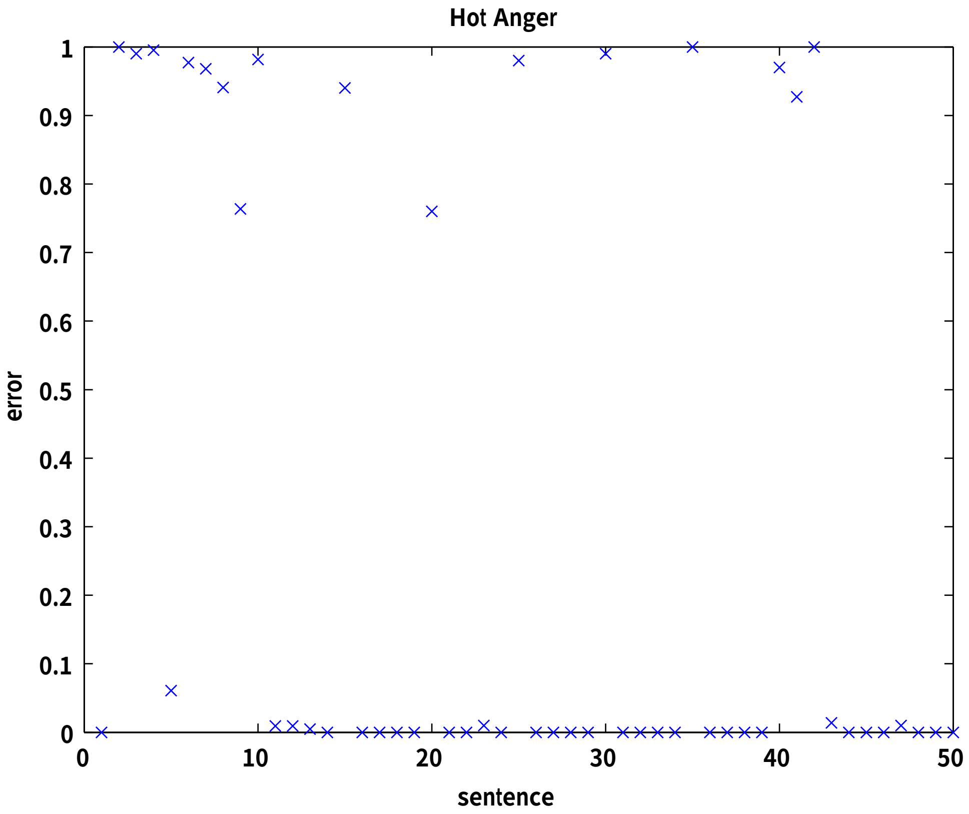


Figure 3-19. Root mean squared error (RMSE) of the 50 sentences of Hot Anger that used to train and check FIS

3.2.6 Results and discussion

The relationship between semantic primitives and expressive speech categories are characterized in Table 3-8. The "top five" semantic primitives were selected for each expressive speech category: three positive correlations (which are the ones that showed the highest correlation values with a positive slope) and two negative ones (which showed the highest correlation values with a negative slope). The reason five semantic primitives were chosen for each expressive speech category was that it is difficult for people to assign a large number of semantic primitives to a specific expressive speech perception, but less than five may not be sufficient to result in a reliable perception.

The table shows a balance of commonly used adjectives with a precise numerical description and seems to be compatible with the way a human responds when they perceive expressive speech. For example, usually a joyful vocal expression sounds bright and clear, but not quiet nor weak. This matches the FIS result for Joy shown in Table 3-7; the other FIS results also seem to be good matches with expressive speech utterances.

3.3 Analysis of a large number of acoustic features

When dealing with how the change of speech signal is related to human's perception in terms of expressive speech by using a computer, doubtlessly an acoustic feature extracted from the speech signal is a cue that can be applied to be an indicator. So far, as numerous studies have investigated this type of effect, there are various acoustic features that can be extracted from the speech signal, some are found to be significant to the perception of expressive but some are not.

Therefore, in this study, before conducting the measurement of acoustic cues, it is necessary to consider what acoustic cues should be measured. The determination is conducted by reviewing previous studies. From the literature review, it is clear that prosody, including voice quality, is suggested to be the most important factor related to the perception of expressive speech; therefore, the explanations in Section 3.3.1 are about the acoustic cues belonging to prosody, including voice quality.

The detailed description of the analysis of each acoustic feature is described in Section 3.3.2. Since the main purpose of acoustic feature analysis is to build the relationship between acoustic features and semantic primitives, in Section 3.3.3, the results of the acoustic feature analysis was applied to build the relationship by calculating the correlation coefficients. As a result, the built relationship between acoustic features and semantic primitives are discussed as well in Section 3.3.3

3.3.1 Acoustic cues related to expressive speech

As mention above, according to much previous work, there is a general agreement that prosody, including voice quality, is considered to be an important factor correlated to the perception of expressive speech.

Prosody

For the perception of expressive speech, prosody is an important cue to study because prosodically different sentences can convey a wide range of information [41]. For example, a simple sentence "I am writing a letter". It can be produced as a declarative (ending with a falling voice), a question (*ending with a rising voice*), a way in which the speaker asks for more information (inserting a long pause before the object of the verb along with a final rising voice.) It can be also produced with different stress patterns, such like, *I am writing a LETTER* (where the stressed item is written in capital letters, thus in this case, indicating that *letter* is emphasized), *I AM WRITING a lette*r (emphasizing *I am writing*), or I -- AM – WRITING – A – LETTER (emphasizing *every spoken word, with pauses in between each word.).* In addition, a person can say, “I am writing a letter” with a sad or happy tone of voice. All these variations of intonation, pause, stress pattern, speed, tone of voice, belong to what is called the *prosody* of a sentence. Clearly, prosody is essential when studying the perception of expressive speech.

Generally, prosody is conveyed by the parameters of fundamental frequency (perceived primarily as vocal pitch), intensity (perceived primarily as loudness), and duration (perceived primarily as length). These features are described below.

- Fundamental frequency

  Speech is normally looked upon as a physical process consisting of two parts: a product of a sound source (the vocal cords) and filtering (by the tongue, lips, teeth, jaw, etc. changing the shape of the vocal tract.).When a person speaks, as air is expelled from lungs through the vocal cords, the vocal cords are caused to vibrate: that is the sound source. The sound source produces a sound wave. Fundamental frequency corresponds to the period of the sound wave which is called the fundamental period. From the aspect of the physical meaning, fundamental frequency reflects the pitch of the voice according to a human’s perception. By analyzing the output speech utterance, the fundamental frequency of the sound wave can be captured. When studying the prosody of connected speech, fundamental frequency is one of the three features that is most consistently used in the literature because it represents the intonation and the pitch contour of a speech utterance. Therefore, when analyzing acoustic correlates to expressive speech, fundamental frequency is crucially important.

- Duration

  From the aspect of the physical meaning, duration is perceived primarily as length, the speed of a speech utterance. In a speech utterance, the durations of phonemes, words, phrases, and pauses compose the prosodic rhythm. It also varies greatly when a speaker is in different emotional states. Consequently, duration is also important to investigate

- Intensity

  Intensity of a speech utterance is perceived primarily as loudness. It is determined by the volume of the air flow of breath sent out by the lungs and is represented by the power contour of the speech signal. A voice with different levels of loudness can be perceived differently. When people are in a good mood, such as happy or cheerful, the voice is usually at a higher level of loudness. Conversely, when people are in a sad mood such as depressed or upset, the voice is usually at a lower level of loudness. Therefore, for studying the perception of expressive speech, intensity is another feature that needs to be investigated.

- Voice quality

In addition to pitch, loudness and length, an additional important perceptual cue that can be used to distinguish different categories of express speech is voice quality. Voice quality refers to the auditory impression that a listener gains while hearing speech. In speech production, the vocal folds have a predominant important role, as also does the shape and length of the vocal tract, including the hypopharyngeal area, just above the vocal tract. The articulators of the vocal tract as well as the characteristics of the vocal folds which are involved with non-paralinguistic factors, e.g. age, healthy, gender, mood, geographic background and so on, directly influence voice quality. As a speech signal is generated by excitation of the vocal tract caused by the modulated flow of air generated by the lungs, the voice quality is represented by the spectrum content of the speech signal. A review of voice quality studies can be found in the paper of Erickson [19].

Previous work has suggested that different categories of expressive speech present various voice qualities. Gobl and Ni Chasaide [26] created stimuli involved with seven voice qualities--tense, harsh, modal, creaky, lax-creaky, breathy, and whispery--and conducted experiments to examine listeners' reactions to the seven voice qualities in terms of eight pairs of opposing affective attributes. For each pair, ratings were obtained on a seven-point scale. Their results confirm that "voice quality adjustments can alone evoke rather different affective colorings". In the study by Patwardhan et al. [59], however, they also pointed out that "there is no simple one-to-one mapping between affective state and voice quality".

There are many possible factors that can influence voice quality. Voice quality may be one of the most important cues for the perception of expressive speech, but also the one most difficult to analyze. It is generally acknowledged that certain acoustic cues measured from the spectrum correspond to the perception of some types of voice quality, such as breathiness, roughness, creaky, whispery and so on. In the following, parameters measured in the spectrum that are considered related to voice quality are described.

Spectrum analysis

- Formants

When speech is produced, the vocal fold vibration causes resonances in the vocal tract. These are usually considered in terms of the spectrum of the speech. Theoretically, there are an infinite number of formants, but only the lowest three or four are of interest for practical purposes. For example, F1, F2, and F3 are useful cues for distinguishing vowels [31]. In the study by Ishii and Campbell [34], they examined the correlation between acoustic cues e.g., fundamental frequency, power value, glottal amplitude quotient (AQ), third and fourth formants (F3 and F4), and some categories of voice quality and speaking manners, e.g., energy and activity, brightness and interest, politeness and considerateness. The reason that F3 and F4 were adopted in their study is because higher formants are more related to the vocal tract length compared with the lower formants. Their results suggested that average F0, average power value, AQ and F4 were found to be the most influential parameters of voice quality. Therefore, formants are one of the meaningful acoustic cues that should be investigated.

However, one problem that arises when measuring formants is that vocal tracts are different among different people. The analysis is more reasonable when the voices analyze are produced by the same people. The database used in this study is justified by this consideration since all the voices are produced by the same voice actress. Another justification is that all acoustic feature measured in this study were normalized to the neutral one. Since each sentence in the database used by this study includes one utterance of each category of expressive speech, it provides the possibility of normalization.

- Spectral shape

Spectral shape can provide cues to relevant aspects of voice quality, for example, H1–A3. This is the ratio of amplitude of the first harmonic (H1) relative to that of the third-formant spectral peak (A3), and has been used by Hanson [28] to characterize spectral tilt. Another example is H1-H2. This is the ratio of amplitude of first (H1) and second (H2) harmonics. As indicated by Menezes et al. [50], H1-A3 reflects glottal cycle characteristics, i.e., speed of closing of vocal folds while H1-H2 is concerned with glottal opening.

Generally, spectral tilt (H1-A3) is related to the brightness of voice quality.

A spectrum with a flatter spectral tilt, stronger energy in the high frequency region, represents a brightness to the voice while a spectrum with a sharper spectral tilt, weaker energy in the high frequency region, represents a darkness to the voice.

As concluded by Schröder [68], "spectral slope is a simple spectral measure of the relation of high-frequency and low-frequency energy." It is used as a simple approximation of the "harshness" vs. "softness" of the voice quality.

- Aperiodic ratio

Aperiodic ratio refers to the ratio of the periodic harmonics component to the aperiodic noise component, i.e., Harmonic-to-noise ratio. Spectra with a higher aperiodic ratio represent a more breathy voice. Krom [43], examining the relationship between spectral parameters and *breathiness* and *roughness* suggested that the harmonic-to-noise ratio was the best single predictor of speech rated as both breathy and rough.

In order to build the relationship between semantic primitives and acoustic features, it is necessary to ascertain what acoustic features possibly exert an influence on the perception of expressive speech,

According to the above discussion, the acoustic features with such properties can be referred to as prosodic features. The acoustic features related to prosody are extracted fundamental frequency contour, power envelope, time duration, and spectrum.

Therefore, fundamental frequency contour (F0 contour), power envelope, spectrum, and time duration are measured. The detailed descriptions of the measurement are presented in the next section.

3.3.2 Acoustic feature analysis

F0 contour, power envelope, and spectrum were calculated by using STRAIGHT [37] with a FFT length of 1024 points and a frame rate of 1ms. The sampling frequency was 22050 Hz. All acoustic features measured in this study were normalized to the neutral. Since each sentence in the database used by this study included one utterance of each category of expressive speech, it provides the possibility for normalization.

F0 contour

The F0 contours for the smaller phrases within the utterance, as well as the entire utterance, were measured. In this study, we refer to the smaller phrasal units as "accentual phrases". We use 'accentual phrase' as a term to refer to a unit is terminated with a drop in F0, and sometimes even a pause. For example, the Japanese sentence /a ta ra shi i ku ru ma o ka i ma shi ta/ (I brought a new car.), forms three accentual phrases: / a ta ra shi i ku ru ma o ka i ma shi ta/. Figure 3-20 shows the F0 contours for the entire utterance as well as for the accentual phrases. Utterances containing the same words (lexical content) but with different expressive speech categories varied greatly in F0 contour and power envelope, both for the accentual phrases as well as for the overall utterance. For each accentual phrase, the measurements made were rising slope (RS), rising duration (RD), falling slope (FS), falling duration (FD) and pitch range in each accentual phrase (RAP) and for each overall utterance, average pitch (AP), pitch range (PR), and highest pitch (HP).

To measure these acoustic features, it is necessary first to separate one utterance into several phrases depended on the content as shown in Figure 3-15. Figure 3-15 also shows the rising slope marked in a red line and the falling slope marked in a back line. In one accentual phrase, it includes high accent parts and low accent parts that are based on Japanese speaking. Rising slope (RS) is the slope of a regression line that fits the raising part of one accentual phrase which is from the start point of one accentual phrase to the highest point of the first high accent part. Falling slope (FS) is the slope of a regression line that fits the falling part of one accentual phrase which is from the highest point of the first high accent part to the end of one accentual part. Rising duration (RD) is the time duration of rising part and falling duration (FD) is the time duration of the falling part. Pitch range in each accentual phrase (RAP) is the difference between the maximum and the minimum F0 in one phrase. Average pitch (AP) is the average of F0 in one utterance. Pitch range (PR) is the difference between the maximum and the minimum F0 in one utterance. Highest pitch (HP) is the maximum F0 in one utterance.

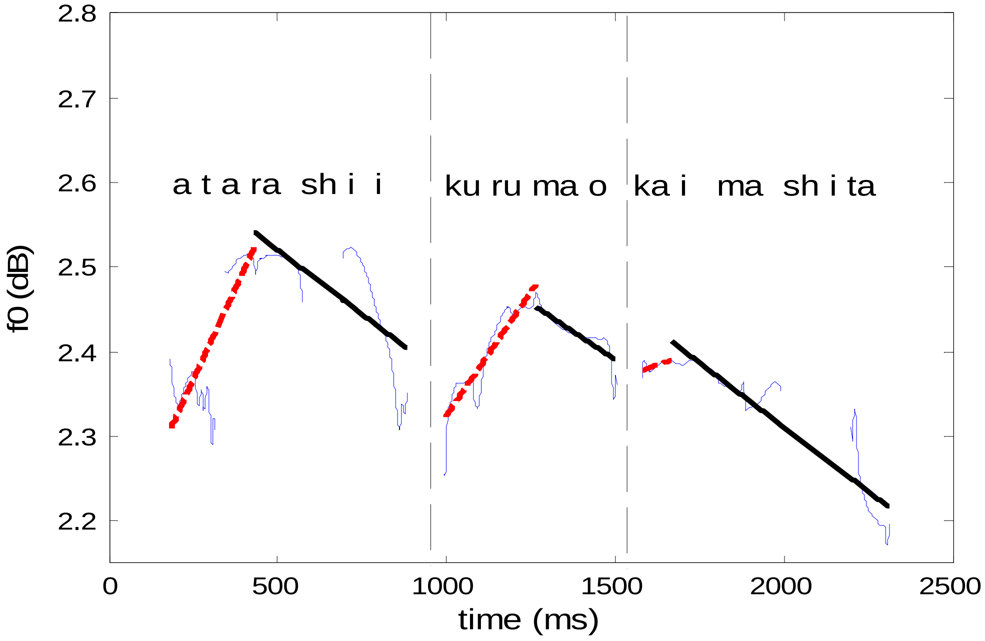


Figure 3-20. F0 Contour for accentual phrases and entire utterance of /a ta ra shi i ku ru ma o ka i ma shi da/.

Power envelope

Power envelope was measured in a way similar to that for the F0 contour. For each accentual phrase, rising slope (PRS), rising duration (PRD), falling slope (PFS), falling duration (PFD), and mean value of power range in accentual phrase (PRAP) were measured. For each overall utterance, power range (PWR), rising slope of the first accentual phrase power (PRS1st), and the ratio between the average power in the high frequency portion (over 3 kHz) and the average power (RHT) were measured.

Each accentual phrase, rising slope (PRS), rising duration (PRD), falling slope (PFS), falling duration (PFD), and rising slope of the first accentual phrase power (PRS1st) were measured based on the same concept as that for measuring F0 acoustic features. Power range (PWR) is the difference between the maximum and the minimum power in one utterance. STRAIGHT analysis provides the average power in the high frequency portion (over 3 kHz). The ratio between the average power in the high frequency portion (over 3 kHz) and the average power is the acoustic feature RHT here.

For spectrum, formants (F1, F2, and F3), spectral tilt (ST) and spectral balance (SB) were measured.

Formant measures were taken approximately at the vowel midpoint of the vowels /a/, /e/, /i/, /o/, and /u/. The sampling frequency of the speech signal was set at 10 kHz. The spectrum was obtained by using STRAIGHT and linear predictive coding (LPC) coefficients were calculated according to the method described in [54]. The method transforms the amplitude spectrum is:

$X[m]$, where $0 \le m \le \frac{M}{2}$ with M is the number of samples in the frequency domain, obtained from STRAIGHT analysis to the power spectrum using Eq. (1)

$$S[m] = \left|X[m]\right|^2, \quad 0 \le m \le \frac{M}{2} \qquad (1)$$

Then, according to the following equation, the inverse Fourier transform is used to calculate the i th autocorrelation coefficient, $R[i]$.

$$R[i] = \frac{1}{M}\sum_{m=0}^{M-1} S[m]\exp\left\{ j\frac{2\pi mi}{M} \right\}$$

Where $S[m] = S[M-m]$ and $0 \le i \le M-1$. Assume that a P th order all-pole model, where $0 < P < M$, can be used to estimated. The reconstruction error is calculated by Eq. (3).

$$P_L = R[0] - \sum_{l=1}^{P} a_l^P R[l] \qquad (3)$$

where $\{a_l^P\}, l = 1,2...P$, are the corresponding LPC coefficients. They can be evaluated by minimizing P_L with respect to a_l^P, where $l = 1,2...P$.

In this study, the first, second, and third formants (F1, F2, and F3) were calculated with LPC-order 12.

To measure voice quality, **spectral tilt** was calculated from H1-A3, where H1 is the level in dB of the first harmonic and A3 is the level of the harmonics whose frequency is closest to the third formant [46]. Other measures of voice quality are H1-H2 [50] or OQ [38, 55], but onlyH1-A3 was done in the study.

Spectral balance (SB) is a parameter that serves for the description of acoustic consonant reduction, and was calculated according to the following algorithm [42]:

$$SB = \frac{\sum f_i . E_i}{\sum E_i}$$

f_i is the frequency in Hz

E_i the spectral power as a function of the frequency

Time duration:

For each sentence, the duration of all phonemes, both consonants and vowels, as well as pauses, were measured. The duration measurements were the following: pause length (PAU), phoneme length (PHN), total utterance length (TL), consonant length (CL) and the ration of consonant duration to vowel duration (RCV).

3.3.3 Correlation analysis of acoustic measurements with semantic primitives

A total of 16 acoustic features were measured: Four involved F0--mean value of rising slope (RS), highest pitch (HP), average pitch (AP) and rising slope of the first accentual phrase (RS1st); four involved power envelope--mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (PRS1st), the ratio between the average power in the high frequency portion (over 3 kHz) and the average power (RHT); five involved the power spectrum-- first formant frequency (F1), second formant frequency (F2), third formant frequency (F3), spectral tilt (SPTL), spectral balance (SB); and three involved duration of the total length (TL), consonant length (CL),the ratio between consonant length and vowel length (RCV).

A correlation between the 16 acoustic features and the 17 semantic primitives was done. Correlation coefficient values that have at least one correlation coefficient over 0.6 are considered significant and are shown as shadowed cells in Table 3-9.

Table 3-9. Correlation coefficients between the semantic primitives and the acoustic features.

PF	bright	dark	high	low	strong	weak	calm	unstable	well mo-dulated	mono-tonous	heavy	clear	noisy	quiet	sharp	fast	slow
RS	0.44	-0.64	0.70	-0.60	0.56	-0.54	-0.74	0.67	0.54	-0.32	-0.40	0.44	0.63	-0.67	0.59	0.50	-0.56
HP	0.69	-0.88	0.90	-0.89	0.42	-0.56	-0.72	0.67	0.50	-0.18	-0.73	0.74	0.60	-0.73	0.44	0.42	-0.62
AP	0.71	-0.88	0.87	-0.91	0.33	-0.54	-0.66	0.60	0.41	-0.10	-0.78	0.76	0.52	-0.70	0.34	0.35	-0.62
RS1st	0.50	-0.79	0.77	-0.78	0.45	-0.58	-0.67	0.60	0.42	-0.10	-0.61	0.66	0.57	-0.72	0.47	0.24	-0.51
PRAP	0.31	-0.67	0.62	-0.56	0.73	-0.67	-0.77	0.73	0.55	-0.26	-0.30	0.47	0.76	-0.78	0.73	0.31	-0.55
PWR	0.43	-0.74	0.74	-0.65	0.70	-0.66	-0.80	0.78	0.59	-0.27	-0.41	0.57	0.76	-0.79	0.69	0.38	-0.57
PRS1st	0.48	-0.80	0.64	-0.70	0.45	-0.78	-0.61	0.51	0.27	-0.01	-0.56	0.64	0.44	-0.78	0.42	0.37	-0.64
RHT	-0.10	-0.05	0.29	0.00	0.68	-0.14	-0.55	0.67	0.52	-0.41	0.24	-0.10	0.72	-0.29	0.68	0.36	-0.16
F1	0.41	-0.64	0.59	-0.60	0.25	-0.39	-0.49	0.52	0.17	0.10	-0.49	0.47	0.43	-0.52	0.29	0.30	-0.29
F2	0.60	-0.41	0.50	-0.56	-0.31	0.07	-0.11	0.07	0.08	0.05	-0.66	0.44	-0.03	-0.09	-0.27	0.11	-0.06
F3	0.60	-0.47	0.61	-0.54	0.01	-0.15	-0.33	0.33	0.33	-0.16	-0.55	0.49	0.23	-0.29	0.02	0.27	-0.10
SPTL	-0.29	0.49	-0.65	0.53	-0.48	0.17	0.62	-0.71	-0.49	0.21	0.30	-0.32	-0.72	0.42	-0.53	-0.23	0.24
SB	0.27	-0.44	0.63	-0.48	0.49	-0.16	-0.55	0.66	0.55	-0.29	-0.28	0.28	0.68	-0.39	0.51	0.20	-0.31
TL	-0.26	0.42	-0.25	0.30	-0.41	0.69	0.52	-0.28	-0.19	0.19	0.21	-0.28	-0.22	0.63	-0.39	-0.59	0.80
CL	-0.36	0.64	-0.39	0.53	-0.34	0.71	0.50	-0.32	-0.10	-0.04	0.47	-0.44	-0.29	0.71	-0.31	-0.37	0.59
RCV	-0.41	0.78	-0.47	0.71	-0.14	0.58	0.29	-0.23	0.02	-0.32	0.66	-0.66	-0.27	0.58	-0.12	0.00	0.28

3.4 Results and discussion

The results of **Error! Reference source not found.** (page 56) are visualized from Figures 3-21 to 3-25. The results of Table 3-9 (page 78) are visualized from Figures 3-26 to 3-42, where each semantic primitive is connected to their "top five" related acoustic features.

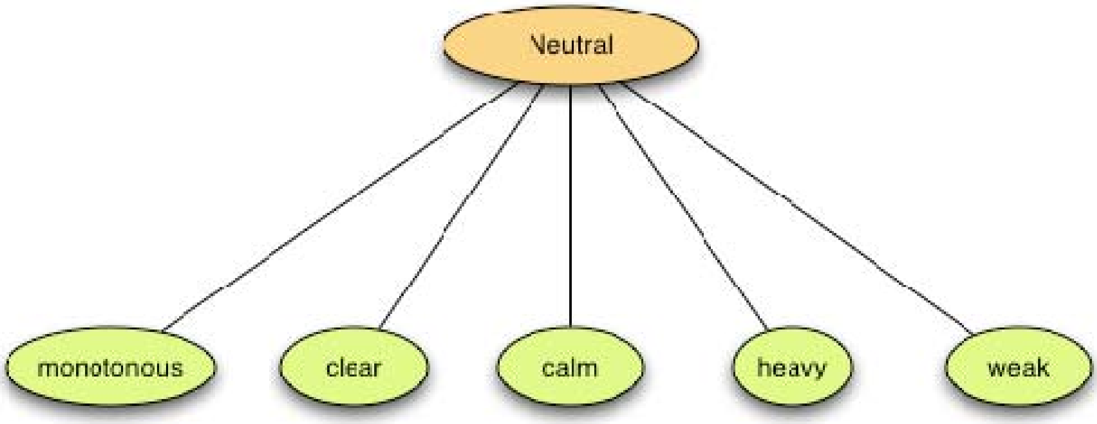


Figure 3-21. Neutral and its related semantic primitives

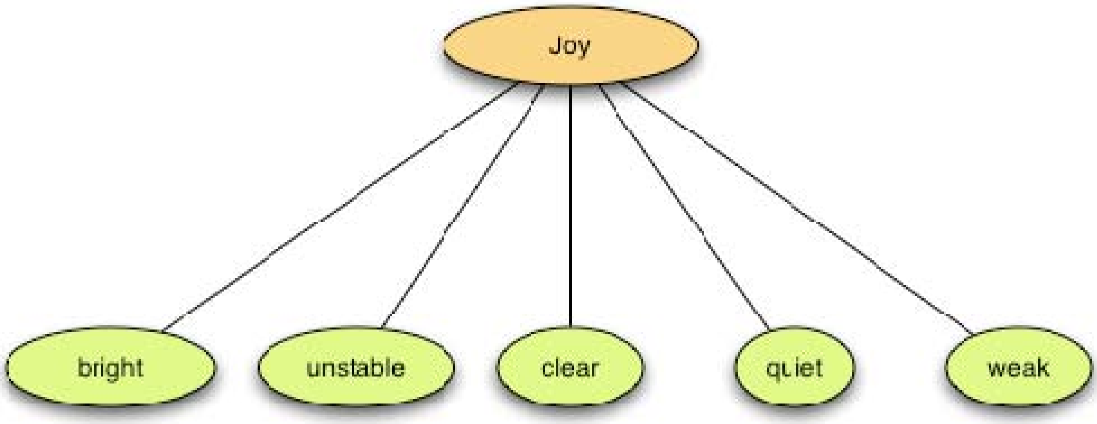


Figure 3-22. Joy and its related semantic primitives

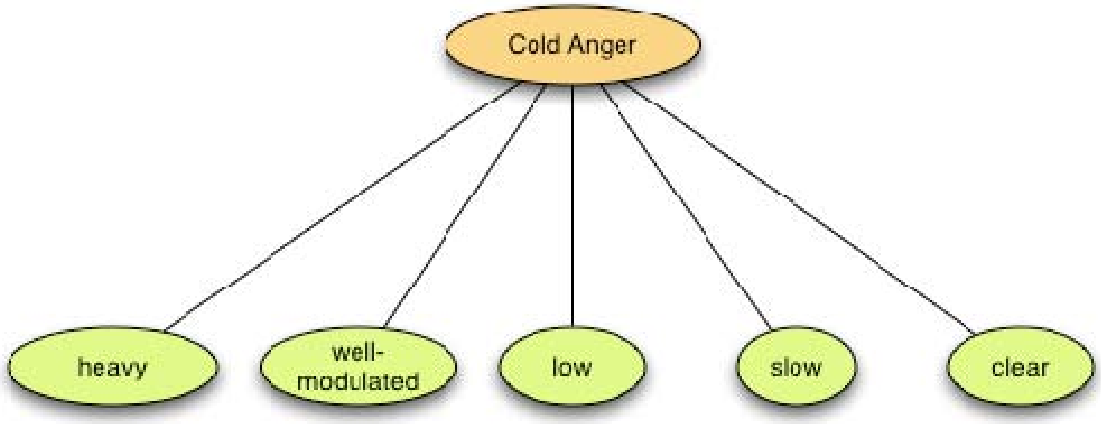


Figure 3-23. Cold Anger and its related semantic primitives

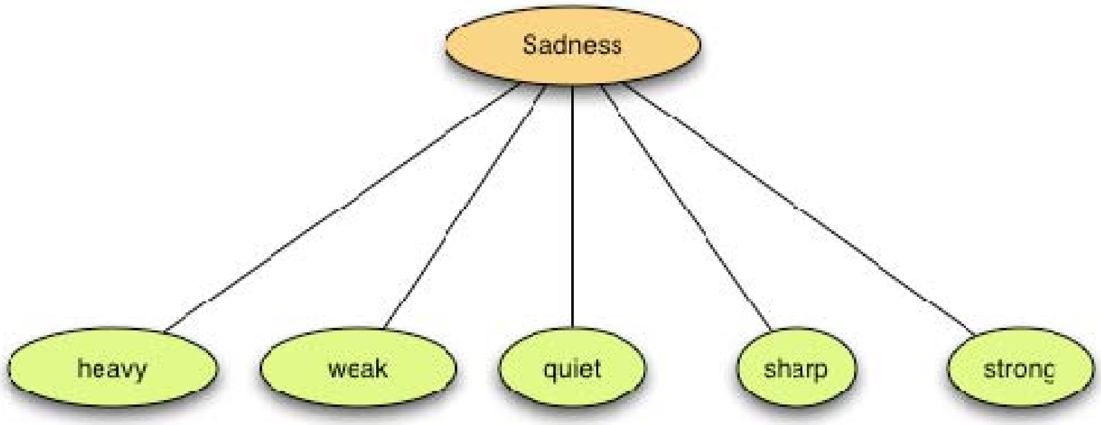


Figure 3-24. Sadness and its related semantic primitives

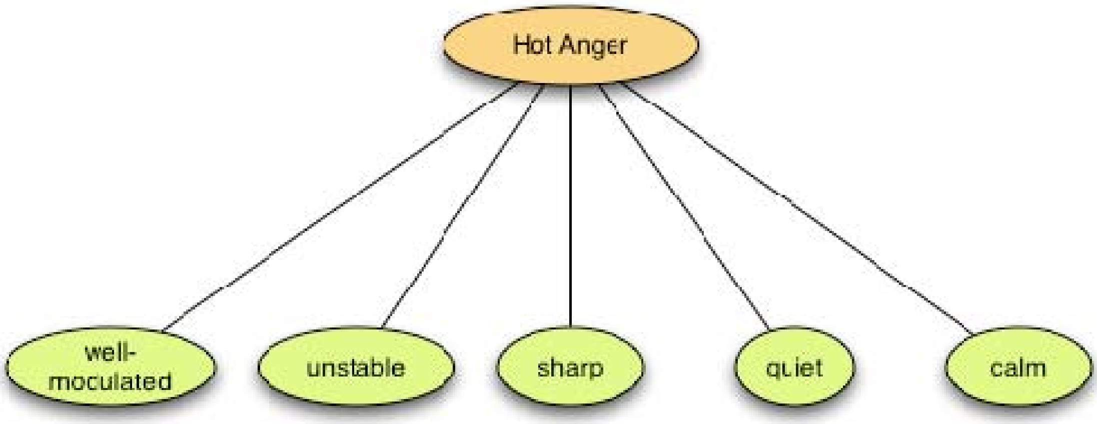


Figure 3-25. Hot Anger and its related semantic primitives

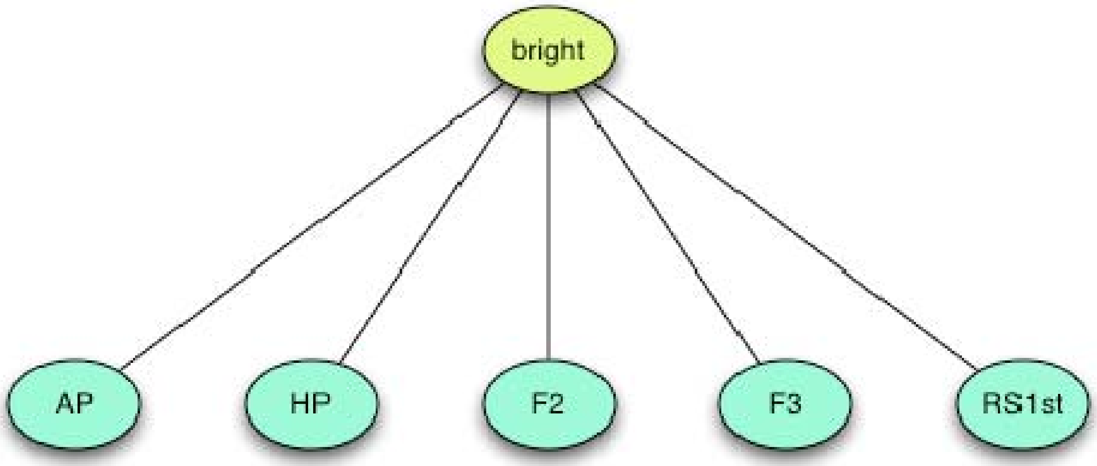


Figure 3-26. Bright and its related acoustic features

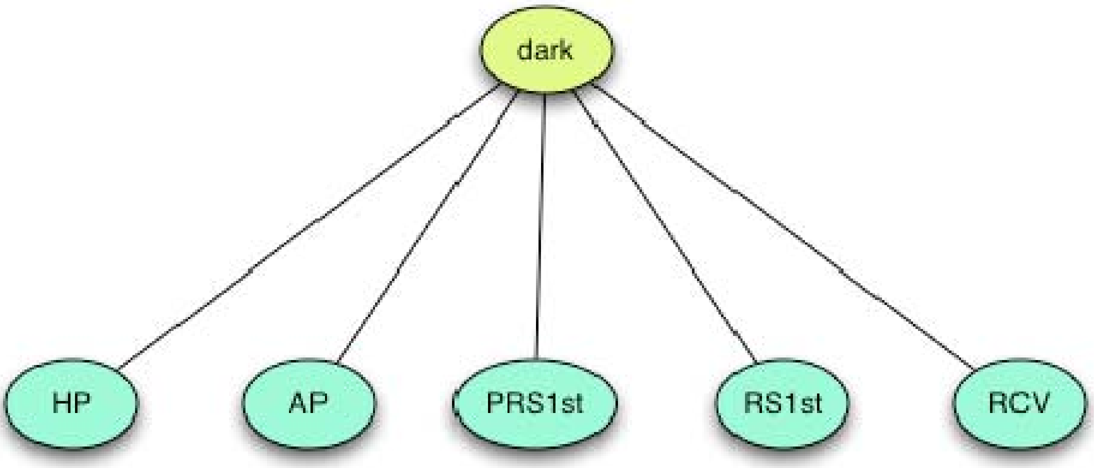


Figure 3-27. Dark and its related acoustic features

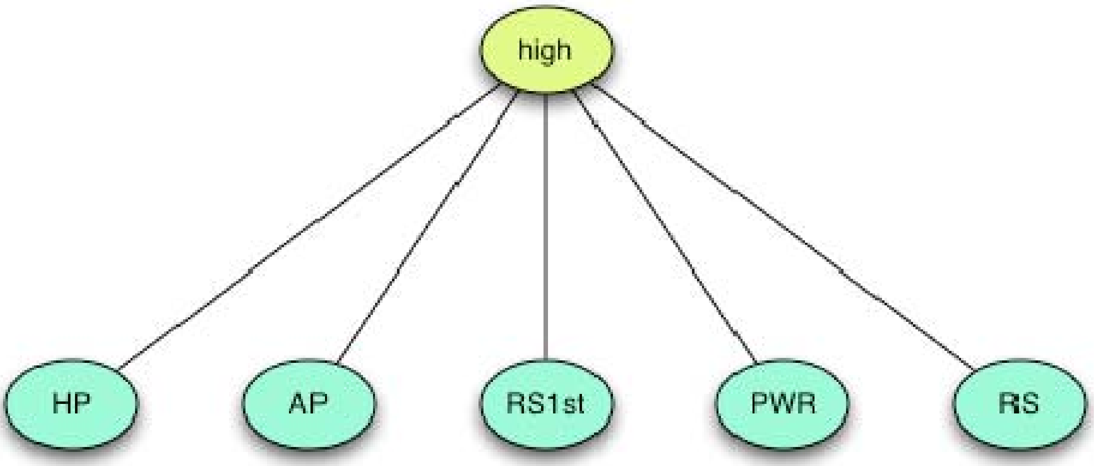


Figure 3-28. High and its related acoustic features

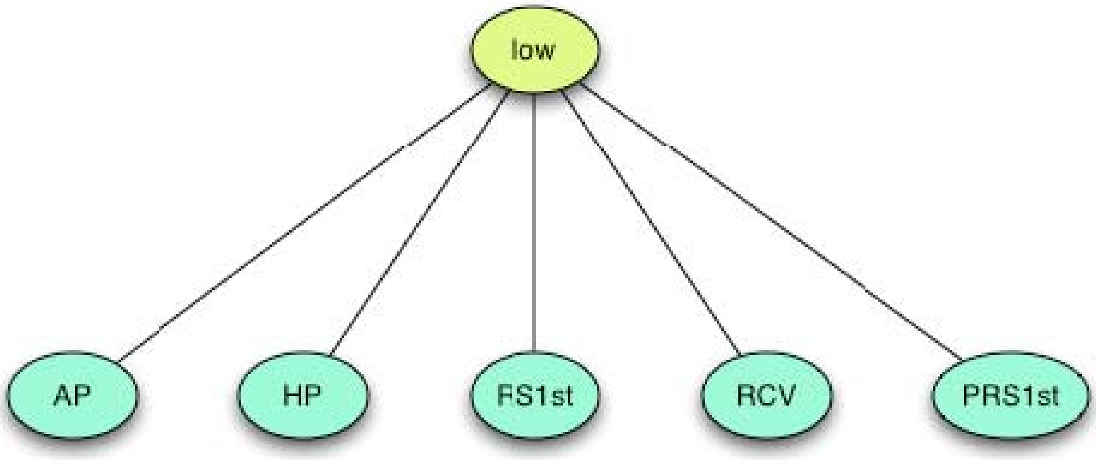


Figure 3-29. Low and its related acoustic features

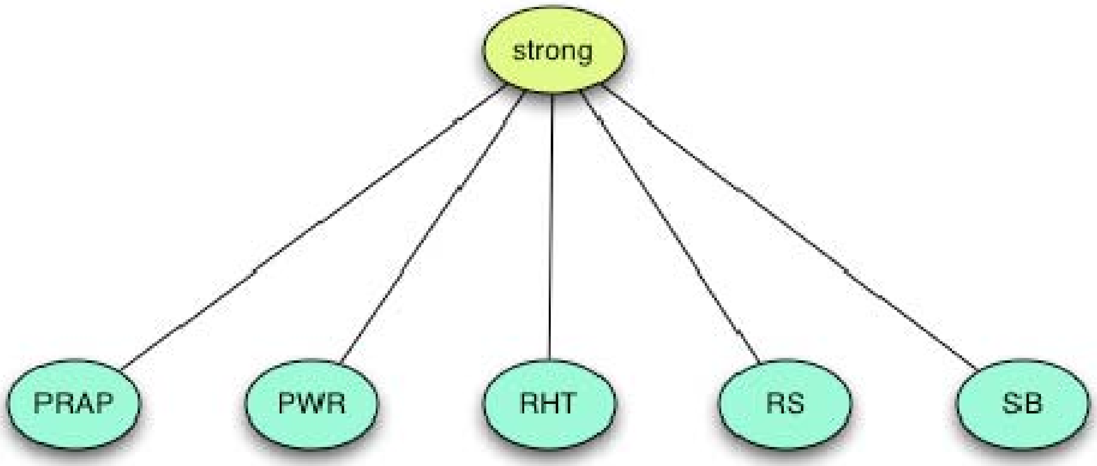


Figure 3-30. Strong and its related acoustic features

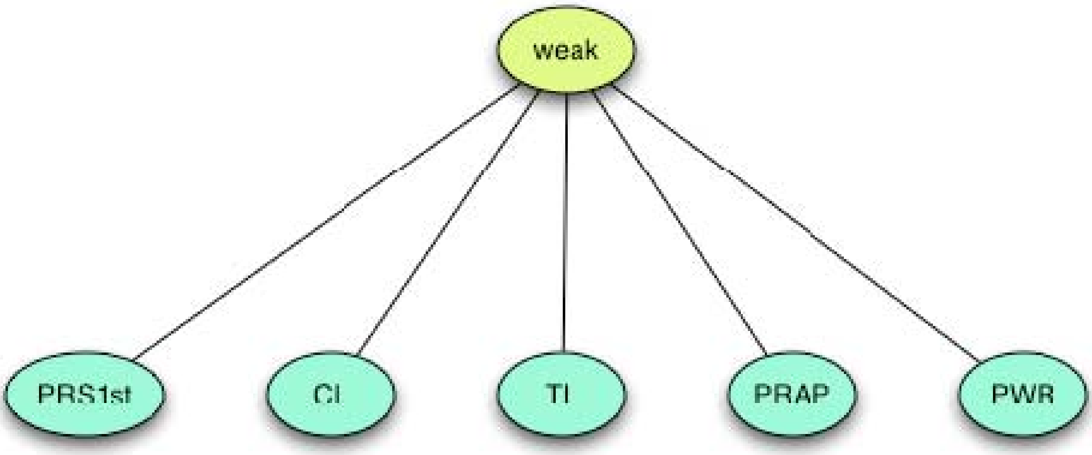


Figure 3-31. Weak and its related acoustic features

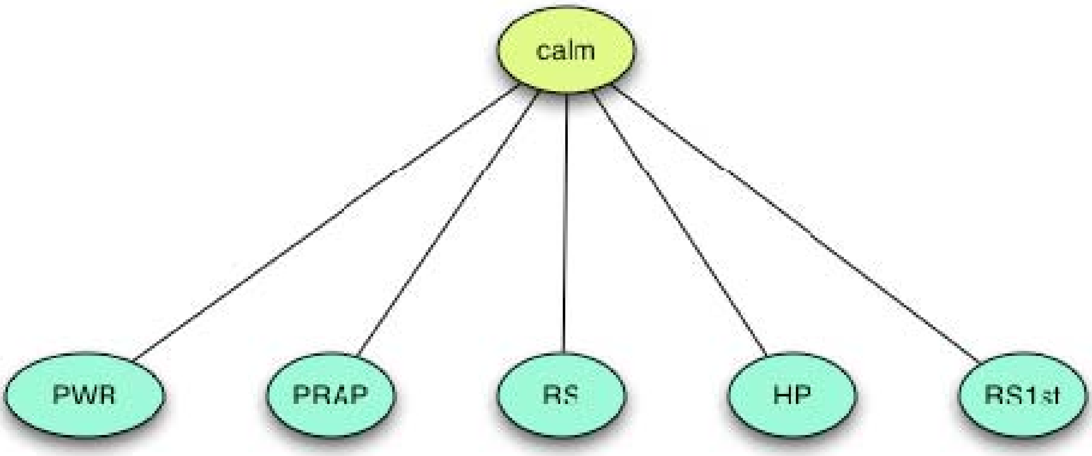


Figure 3-32. Calm and its related acoustic features

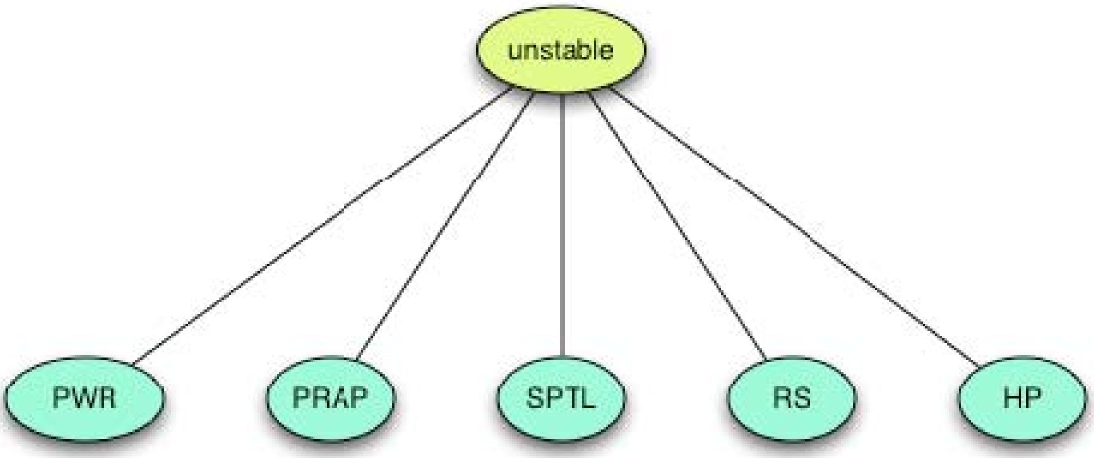


Figure 3-33. Unstable and its related acoustic features

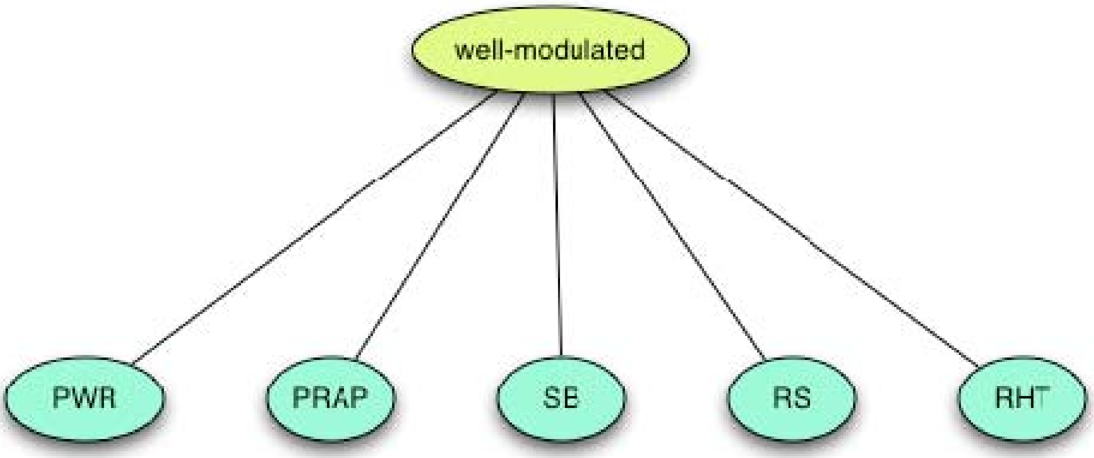


Figure 3-34. Well-modulated and its related acoustic features

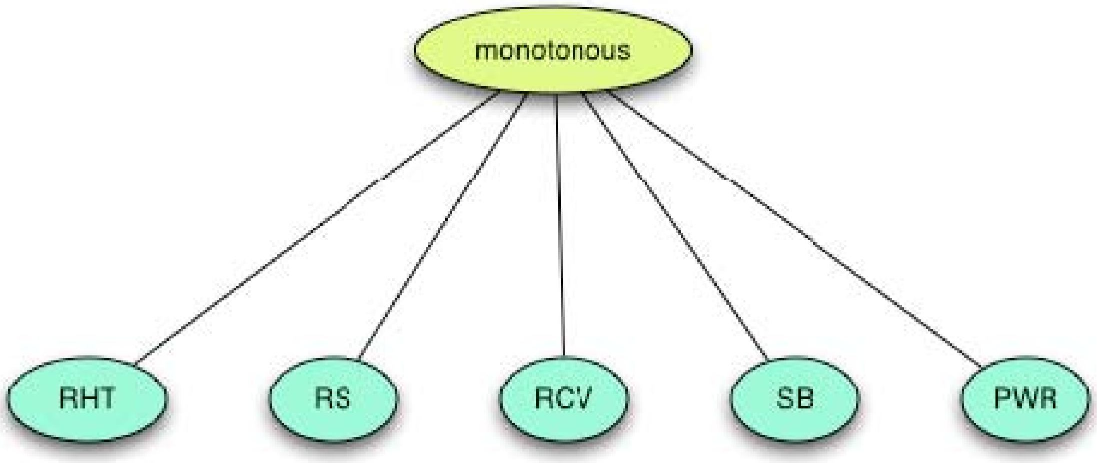


Figure 3-35. Monotonous and its related acoustic features

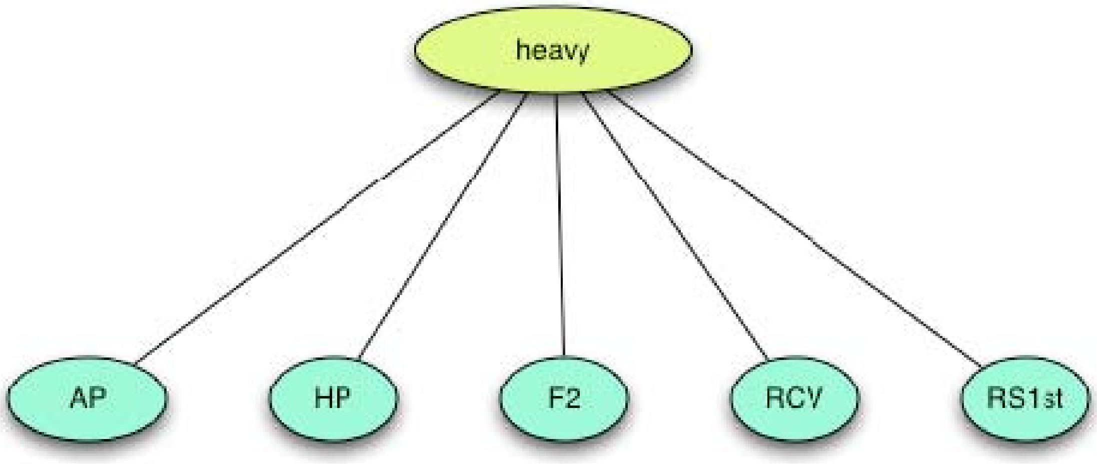


Figure 3-36. Heavy and its related acoustic features

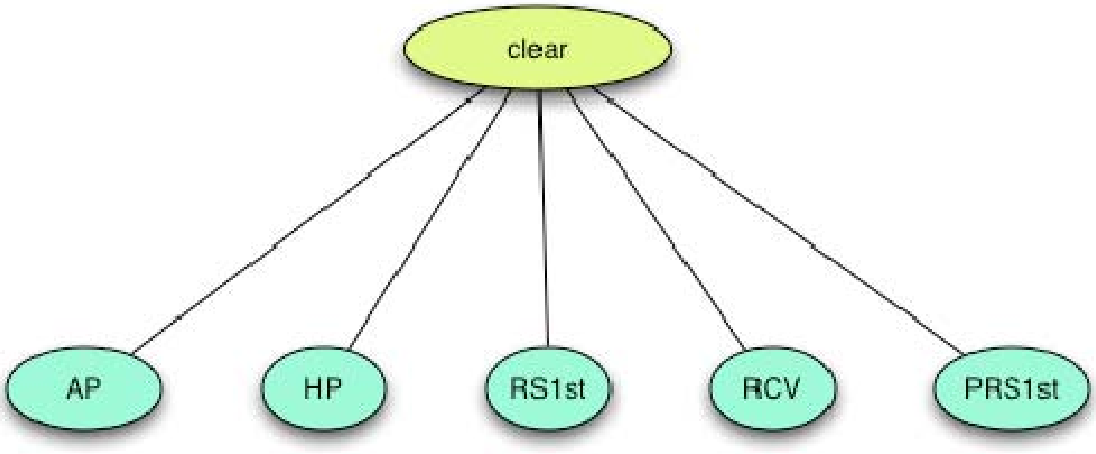


Figure 3-37. Clear and its related acoustic features

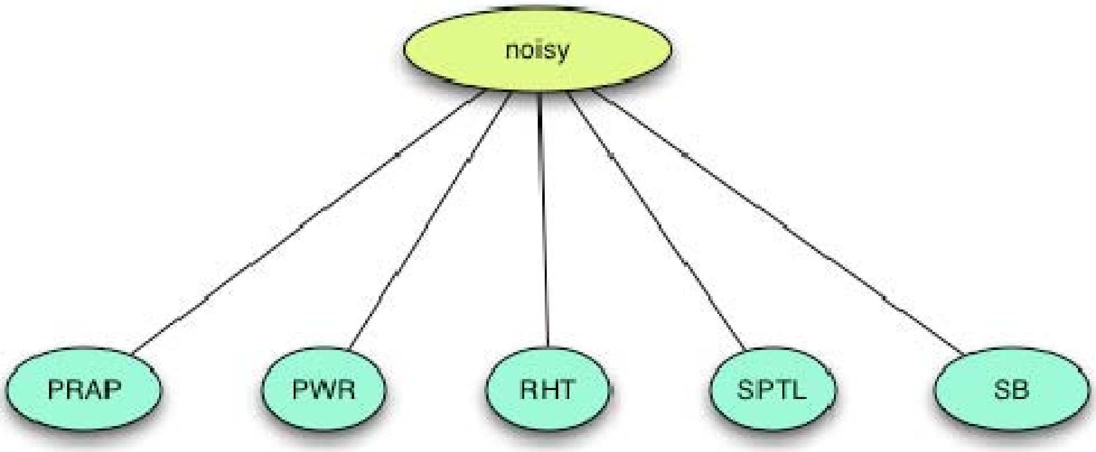


Figure 3-38. Noisy and its related acoustic features

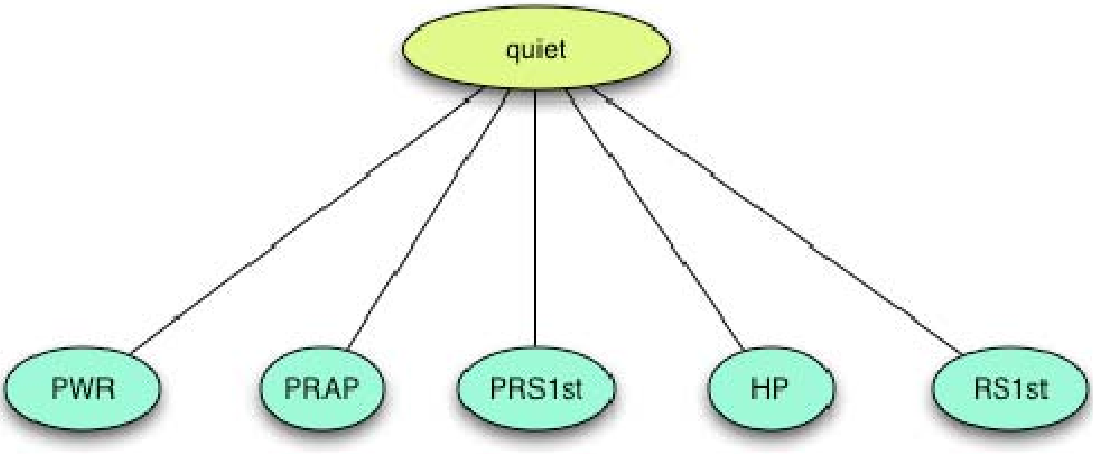


Figure 3-39. Quiet and its related acoustic feature

s

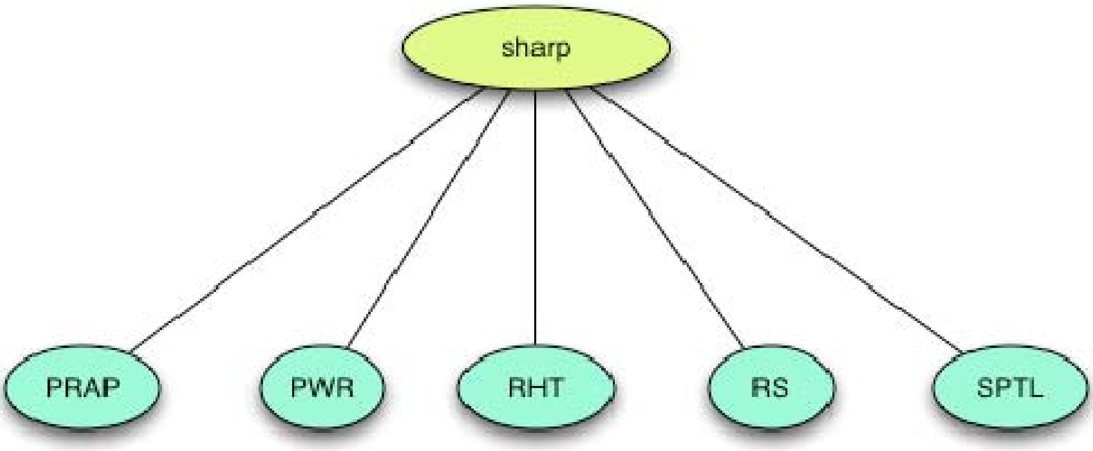


Figure 3-40. Sharp and its related acoustic features

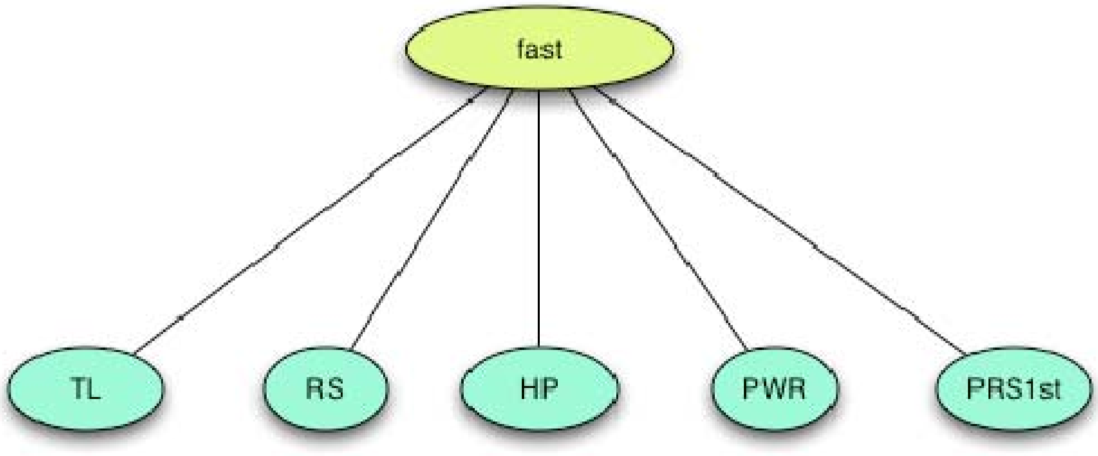


Figure 3-41. Fast and its related acoustic features

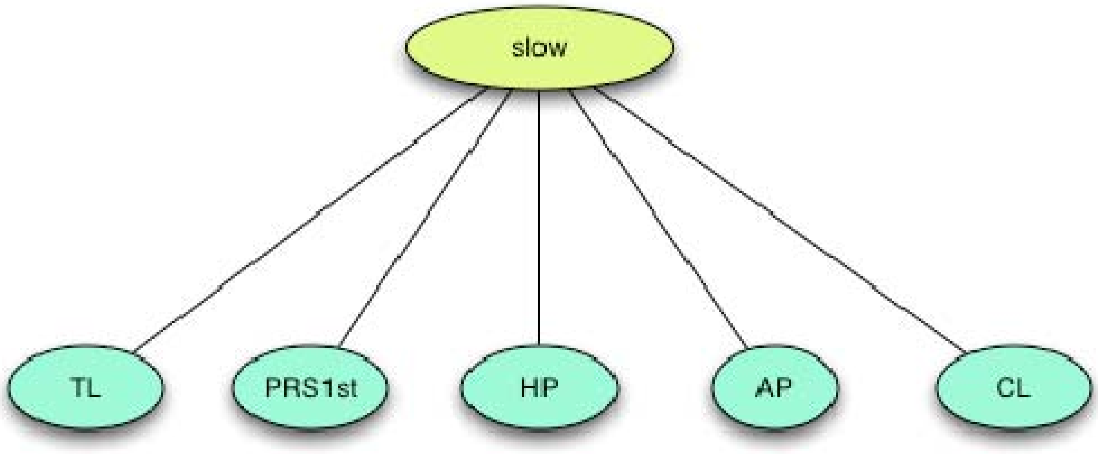


Figure 3-42. Slow and its related acoustic features

Based on the results described in this chapter, a perceptual model for each category of expressive speech was built. Figures 3-43, 3-44, 3-45, 3-46, and 3-47 illustrate the perceptual model for *Neutral*, *Joy*, *Cold Anger*, *Sadness*, and *Hot Anger*, respectively. In the figures, the solid lines indicate a positive correlation, and the dotted ones, a negative correlation. For the relationship between expressive speech and semantic primitives, the highest values are shown in the bold lines; others are shown in non-bold lines. For the relationship between semantic primitives and acoustic features, the highest two values are shown in the bold lines, other are shown in non-bold ones. Simply put, the thicker the line is, the higher the correlation. For example, the model in

Figure 3-44 describes that a *Joy* speech utterance will sound *bright*, *unstable* and *clear* but not *quiet* or *weak*. In the figures, it also shows which acoustic features are most related to which semantic primitives of each category of expressive speech. This visual aid not only expresses the relationships between different layers, but is also used in the verification process, which will be described later.

As mentioned earlier, the semantic primitives involved with the perceptual model for each expressive speech category can be thought to correspond to a human's actual perceptual behavior. The acoustic features which mainly affect the perception of semantic primitives, as can be seen from Table 3-9, are those related to F0 contour and power envelope. From this observation, it would seem that the perception of expressive speech, at least in Japanese, may mainly be affected by the change of F0 contour and power envelope.

However, in addition, spectral characteristics affect perception of some of the semantic primitives, for example, bright, high, unstable, and noisy. These are the more active adjectives. This result suggests that change of spectrum, i.e., change in voice quality, may encourage perception of active semantic primitives.

Table 3-9 also shows that the easier a semantic primitive is perceived, the more acoustic features can be found and the higher is the correlation coefficient. This explains why the correlation coefficients of the semantic primitives well-modulated and monotonous were relatively low compared to other semantic primitives—that is, these two are relatively more "abstract" and not easily quantified by human perception.

Two points that may make people confused from the figures are monotonous (see Figure 3-43) and well-modulated (Figures 3-45 and 3-47), where no acoustic features connect to either of them.

1. For monotonous, it might be due to the fact a neutral voice is usually perceived as plain, in other words, lack of variety, which has identical meaning with monotonous, than other four types of expression speech categories, which made us hard to find acoustic features that are significant to be measured when a voice is already a neutral voice. That is, for a neutral voice, there is no significant acoustic feature that would make a neutral voice to be more monotonous.

2. For well-modulated, this does not imply that there is no acoustic feature that affects the perception of well-modulated. However, if we lower the criteria of coefficients to 0.5 then we can find RS, HP, PRAP, PWR, RHT and SB. The

choice of this criteria 0.6 is to take a moderate results, between too many and too few. For example, when lowering the criteria 0.5, there are three more acoustic features RS, SPTL, and SB found that affect the perception of sharp. Too many acoustic features may lower the significance of each acoustic feature in the model.

The resultant perceptual models are consistent with the first assumption we made in Chapter 1 that people do not perceive expressive speech directly from the change of acoustic features but rather from the perception of semantic primitives. The models also substantiate results from previous work of the importance of F0 and power envelope in perception of expressive speech.

However, verification of the model is necessary. To this end, the analysis by synthesis method was adopted. We resynthesized the expressive speech utterances using the semantic primitive specifications with corresponding acoustic features, as determined by the correlation analysis results described in Section 3.3.2 above. In this way, it is possible to assess the validity of the relationship between semantic primitives and acoustic features assumed by the purposed perceptual model.

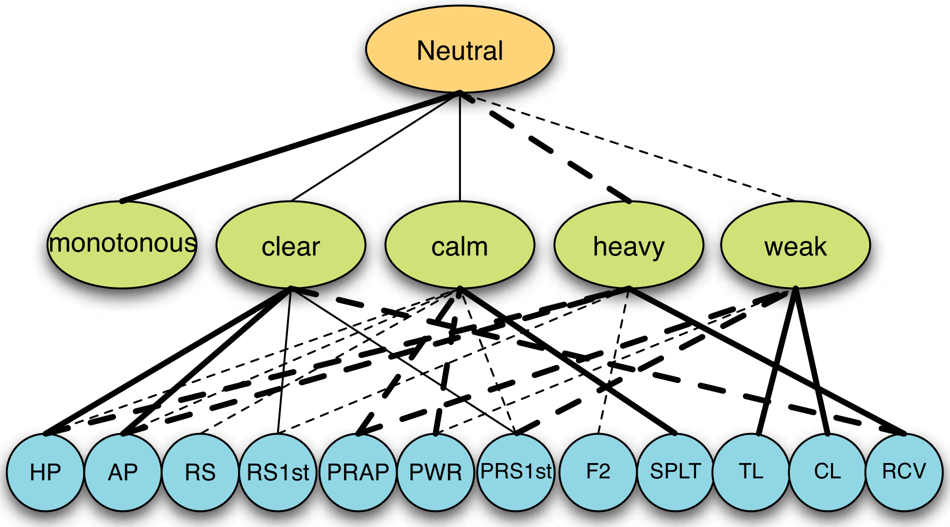


Figure 3-43. Resultant perceptual model of *Neutral*.

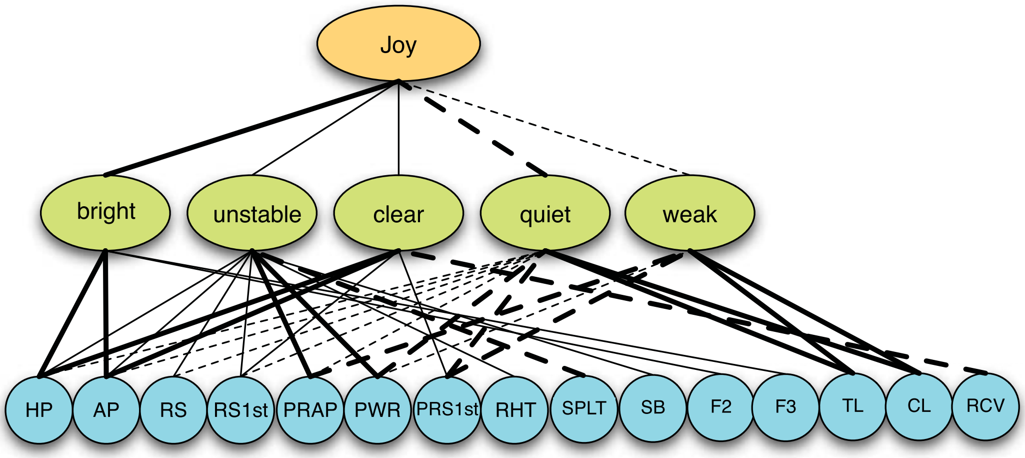


Figure 3-44. Resultant perceptual model of *Joy*.

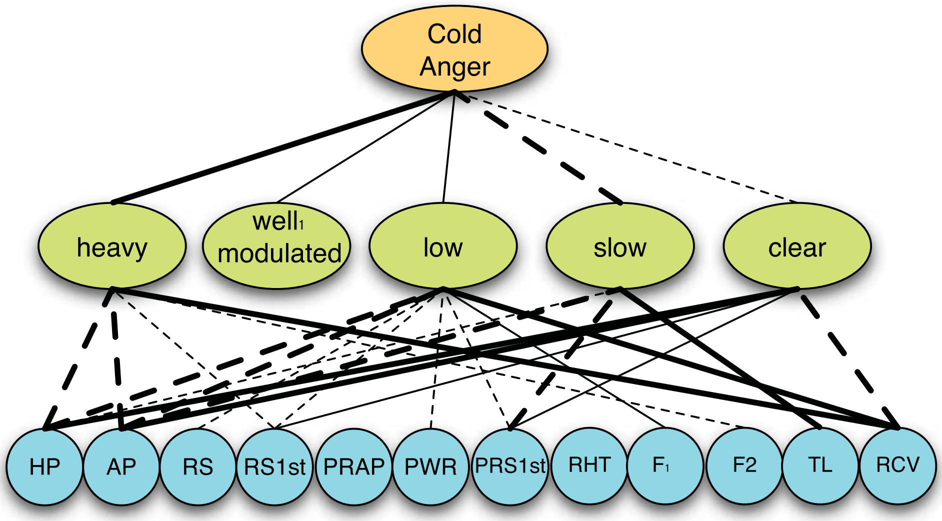


Figure 3-45. Resultant perceptual model of *Cold Anger*.

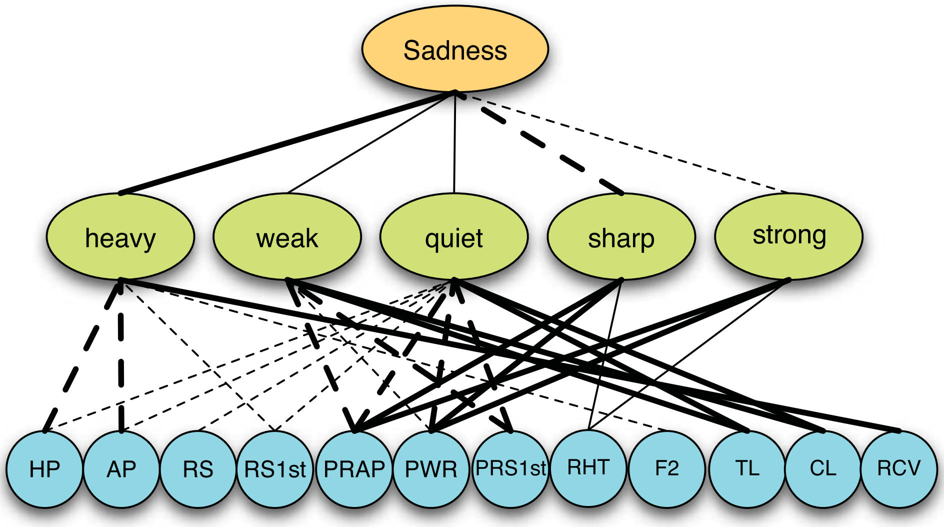


Figure 3-46. Resultant perceptual model of *Sadness*.

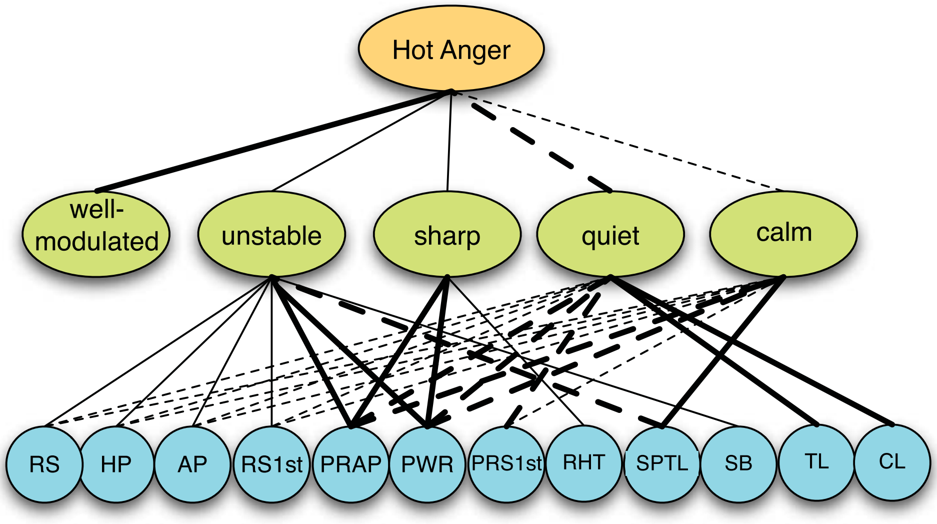


Figure 3-47. Resultant perceptual model of *Hot Anger*.

3.5 Summary

This chapter provides a complete description of the work with regard to the building of the proposed model which consisted of three layers, expressive speech, semantic primitives and acoustic features. In order to accomplish the work, there are three sub tasks accomplished.

- To find those descriptors (e.g., adjectives) that can be used to describe the perception of expressive vocalizations, a thorough selection of semantic primitives will be necessary. Three experiments are carried out in order to determine which adjectives are suitable for describing expressive speech.
- To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech, a fuzzy inference system will be built. In order to build FIS with highly reliable probability, another experiment, Experiment 4, is conducted for collecting more utterances as input data.
- To support the relationship between semantic primitives and acoustic features, analysis of a large number of acoustic features will be necessary.

For the first task, detailed descriptions and discussions of those three experiments conducted for selecting appropriate semantic primitives is provided in Section 3.1. Those three experiments are (1) examination of listeners' perception of expressive speech, (2) construction of a psychological distance model and (3) selection of semantic primitives. The first experiment results in showing that subjects can well perceive the intended emotion of most utterances. Moreover, the perceptual ratings of all utterances in the voice database contribute to building the relationship between expressive speech category and semantic primitives. The second experiment results in a psychological distance model which depicts the position of each category of expressive speech where the distance between each of the two categories represents the similarity among them. Such a psychological distance model conveying similarity information can be utilized when investigating appropriate semantic primitives in Section 3.1.3. The main contribution gained from the third experiment is the 17 semantic primitives which will then be used for representing the perceptual characteristics of expressive speech in the perceptual model.

For the second task, Section 3.2 starts with an explanation regarding the fuzzy inference system which is adopted to build the relationship between expressive speech categories and semantic primitives by modeling those rating results of the three experiments conducted in Section 3.1. It is followed by an introduction to an experiment which is conducted to collect more data in order to build a fuzzy inference system. The next, descriptions indicating the process of a fuzzy inference system construction and evaluation is provided. According to the evaluation results, five semantic primitives with high significant correlation to each category of expressive speech were selected for each of them. The generated relationships are a balance of commonly used adjectives with a precise numerical description which seem to be well-matched with the way a human responds when they perceive expressive speech.

For the third task, Section 3.3 first provides a discussion with a brief literary review regarding acoustic cues related to expressive speech. The acoustic cues involved with prosody parameters are mainly extracted from F0 contour, power envelop duration and spectrum. Consequently, acoustic feature analysis done in this study were in terms of F0 contour, power envelop, duration and spectrum. A detailed introduction to the process of acoustic feature analysis is given in this section as well. The correlation coefficients between semantic primitives and acoustic features were then calculated for depicting the relationship. The built relationship between expressive speech and semantic primitives and that between semantic primitives and acoustic features were combined to formulate a perceptual model for each category of expressive speech, which is illustrated by a figure (see Figures 3-6 to 3-10) for each of them.

A further discussion about those built perceptual models is given in Section 3.4, and concludes with the finding that the resulting models are consistent with the first assumptions made in Chapter 1, i.e., that the perception of semantic primitives has a significant effect on the perception of expressive speech and that people do not perceive expressive speech directly from the change of acoustic features but rather from the perception of semantic primitives. Moreover, the built models also substantiate results from previous work about the importance of F0 and power envelope in perception of expressive speech.

Chapter 4. The Verification of the Three Layered Model

This chapter describes the second-step of the model construction. To provide support for the first assumption that states that before listeners can decide to which expressive speech category a speech sound belongs, they will qualify a voice according to different descriptors, where each descriptor is an adjective for voice description. In this step, the two relationships built in the first-step, which were described in Chapter 3, are verified. By this verification step, the effectiveness of the relationships built by the fuzzy inference system and by acoustic cues analysis can be validated.

The verification is conducted by a rule-based type of morphing approach. Usually, a morphing procedure is to change one recorded sound in gradual steps to another recorded sound, first by using a copy-synthesis program to copy-synthesize each of the original sounds, and then using the morphing procedure developed by Kawahara et al. [37] to gradually “morph” from the one sound to the other sound. In the case of this research, however, we use morphing-type rules to change from one copy-synthesized recorded sound to a synthesized speech sound. Specifically, we copy-synthesize the original neutral utterance and then use rules involving semantic primitives together with a combination of multiple rules to create each expressive speech category. This was done to focus on the verification process for the validation of the results of the analysis. In this way, by using this type of morphing technique, a neutral utterance can be taken as input, and by systematically manipulating acoustic features of the input utterance, it is possible to produce a voice that can be perceived as having different semantic primitives or different expressive speech categories. The choice of using a neutral utterance as the basis for morphing is because the neutral utterances are less dynamic than those utterances perceived as other expressive speech categories and thus the changes made to the neutral utterances by the process of morphing are more easily perceived by listeners.

Two types of information in the relationships were validated. The first type is the significance of one element (acoustic feature or semantic primitive) in a layer to the perception of one element (semantic primitive or expressive speech category) in another layer. The second type is the impact direction and strength of one significant element. For example, Figure 3-7 (Chapter 3) explains that *bright*, *unstable*, *clear*, *quiet*, and *weak* are significant to the perception of *Joy*. Furthermore, the impact direction and strength to *Joy* of these five semantic primitives are different, which can be observed from the styles of the connected lines. The verification of these two types of information consolidates the support for the first assumption we made in this research work.

The work of this step consists of the following three sub tasks.

- To develop morphing rules, which represent the changes of acoustic features in expressive speech, from the analyzed acoustic features and the built FIS. This step embeds the two types of information in the built relationships in the rules.
- To modify a neutral utterance according to the developed morphing rules. STRAIGHT provides the tool for morphing neural utterances by the created rules. This step creates the implementation for the morphing neutral utterances by the rules.
- To conduct perceptual experiments for examining the modified utterance in terms of the perception of semantic primitives and expressive speech categories. This step verifies the built relationships by the morphed utterances.

For the first sub task, morphing rules were developed for the two types of information and how the results built in the first step are transformed to rules are explained in Section 4.1. For the second sub task, the implementation of morphing logic on STRAIGT is described in Section 4.2. For the third sub task, the perceptual experiments and the results are discussed in Section 4.3.

4.1 Morphing rule development

The concept of rule development of this research work is described. This description briefly explains why rules are developed and what information in built relationships should be transformed. After this, the principles of rule development are explained. These principles further clarify what is emphasized during the rule development. Finally, the details of the transformation process are described.

4.1.1 Concept of rule development

To simplify the manipulation of acoustic features in the morphing process, the models shown in Figures 3-7 to 3-10 are represented as rules. To verify the two types of information in the models, two types of rules (i.e., morphing rules) were developed. For verifying the first type of information by considering only those elements (acoustic features or semantic primitives) that are significant to the perception of one element (a semantic primitive or an expressive speech category), (1) base rules are developed. The base rules provide the basis for the verification of the models. For verifying the second type of information by changing the values in the base rules in terms of the impact direction and strength of acoustic features to the perception of one element in a layer, (2) intensity rules are developed.

For assessing which acoustic features are involved in creating the percept of each semantic primitive, it is necessary to morph a neutral utterance into an utterance (i.e., "SU-utterance") which could be perceived in terms of one and only one semantic primitive. "SR-base" rules are created for this purpose. For accessing how the change in the intensity of the acoustic features changed the intensity levels of semantic primitives, it is necessary to morph an utterance in such a way that the intensity of the semantic primitive changed. These rules are called "SR-intensity" rules.

For assessing which semantic primitives are involved in creating the percept of each expressive speech category, it is necessary to morph a neutral utterance into an expressive speech utterance (i.e., "EU-utterance") which could be perceived in terms of one and only one expressive speech category. "ER-base" rules are created for this purpose. For accessing how the change in the intensity of the semantic primitives changed the intensity levels of expressive speech categories, it is necessary to morph an utterance in such a way that the intensity of the expressive speech category changed. These rules are called "ER-intensity" rules.

4.1.2 Principles of rule development

To achieve a better quality of rule development, it is necessary to set a higher-level goal. This goal tells us what should be emphasized and what should be overlooked. It provides a direction when developing rules and clarifies the expected results from the rules.

This higher-level goal consists of three mandating principles and one optional principle. The rules with their explanations are as follows.

Obligatory Rules

- Rules are monotonous. This means that each base/intensity rule generates one single intended perception, which can be one semantic primitive or one expressive speech category.

- Rules are general. This means rules are applied generally. They should not be created for any individual perception.
- Rules are dynamic. This means that each constituent of a combined rule can affect the generation of a different perception, but it will still be monotonous. This principle applies to the development of SR-rules and their combination as ER-rules.

Optional Principle

- The pursuit of naturalness. The focal point of the morphing process is to find an effective way to synthesize expressive utterances using semantic primitives. If pursuit of naturalness takes precedence over the obligatory rules, successful morphing of utterances may be compromised. For this reason, “pursuit of naturalness” is optional.

4.1.3 Rule development

For both relationships, the verification process is identical:

(1) Base rule development (SR-and ER- base rules).

(2) Base rule implementation.

(3) Experiment for evaluating base rule efficiency.

(4) Intensity rule development.(SR- and ER-intensity rules)

(5) Experiment for evaluating intensity rule efficiency.

As described in Section 4.1.1, base rules are for assessing which acoustic features or semantic primitives are involved in creating the percept of each semantic primitive or expressive speech category. Intensity rules are for accessing how the change in the intensity of the acoustic features or semantic primitives changed the intensity levels of semantic primitives or expressive speech categories.

First, the relationship between semantic primitives and acoustic features was examined; then, that between semantic primitives and expressive speech.

Base rule development for semantic primitives (SR-base rules)

To verify the first type of information in the resulting relationship between acoustic features and semantic primitives, two things should be done.

First, we need to select only the acoustic features that are considered “significant” to the percept of semantic primitives. Correlation coefficient values between acoustic features and semantic primitives that have at least one correlation coefficient over 0.6 are considered, see Table 3-9. One problem here is that both *well-modulated* and *monotonous* are without any correlation coefficient over 0.6 in the table. This may be because these two semantic primitives usually involve fewer changes of acoustic features. To overcome this characteristic, those acoustic features with correlation coefficients over 0.5 are selected for *well-modulated* and those with over 0.3 are selected for *monotonous*

Second, we need to obtain the morphing parameters by calculating the difference of acoustic features between the input neutral utterance and the utterances of the intended semantic primitive. There is one base rule for one semantic primitive. One rule has 16 parameters which control the 16 acoustic features of the bottommost layer that are measured in Section 3.3. The values of the parameters are the percentage of changes to an acoustic feature of an input neutral utterance, and are calculated by the following method.

From the 50 utterances that were used when building FIS (see Section 3.1.6), 10 utterances were selected that were well-perceived for that semantic primitive. In order to reduce bias in the data, the utterance that showed the greatest deviation from the mean perception score was discarded, thus leaving 9 utterances. For each of the remaining 9 utterances, the differences between the values of their acoustic features and the values of the acoustic features of the neutral utterance from which it should be morphed were calculated. Then we calculated how much the acoustic features of each utterance varied compared to those of the neutral utterance (i.e., percentage variation) by dividing the differences in the values of the acoustic features with those of the corresponding neutral utterance. Finally, the percentage variations of each of the 9 utterances were averaged to give the values of acoustic features for each semantic primitive. Equation (1) presents the calculation.

$$\frac{\sum_{i=0}^{8} \frac{vaf_i - vnaf_i}{vnaf_i}}{9} \quad (1)$$

Where vaf_i is the value of acoustic features of ith utterance and $vnaf_i$ is the value of the corresponding neutral utterance. One example of the parameters is shown in

Table 4-1. The second column labeled **SR1** lists the variation of percentages that are used for morphing a neutral utterance to an utterance supposedly perceived as bright.

Intensity rule development for semantic primitives (SR-intensity rules)

To verify the second type of information in the resulting relationship between acoustic features and semantic primitives, we need to change the values in the base rules according to the styles of the connected lines shown in Figures 3-7 to 3-10. That is, the solid lines indicate a positive correlation, and the dotted ones, a negative correlation. The thicker the line is, the higher the correlation. In this way, the parameters of the base rules were adjusted such that the parameter with a solid thick line would be changed in a positive direction by a larger amount than that of the solid thin line. The parameter with a dotted thick line would be changed in a negative direction by a larger amount than the dotted thin line.

In order to create the intensity rules, the parameters of the base rules were adjusted so that the morphed speech utterance could be perceived as having different levels of intensity of the semantic primitives. Three intensity rules (SR1, SR2, and SR3) were created. SR1 was directly derived from the base rule without any modification. SR2 and SR3 were derived from SR1 with modification. The utterance morphed by SR2 was supposed to be with stronger perception than that morphed by SR1; the utterance morphed by SR3 was supposed to be with stronger perception than that morphed by SR2. Specifically, SR2 was created by increasing 4% or 2% for the solid thick and thin line, respectively, or decreasing with 4% or 2% for the dotted thick and thin lines, respectively, for each parameter of the acoustic features of SR1. SR3 was created by increasing 4% or 2% for the solid thick and thin line, respectively, or decreasing with 4% or 2% for the dotted thick and thin lines, respectively, for each parameter of the acoustic features of SR2.

For example, in Figure 3-7 the line between *bright* and AP (Average Pitch) is a solid thick line. Therefore, the value of the parameter AP was increased from, 6.9% to 10.9% (see Table 4-1). However, in Figure 3-7 the line between *bright* and F3 is a solid thin line. Therefore, a smaller value is given to the parameter F3 from 4.2% to 6.2%. The parameters of SR1 come from the base rule of *bright*, which was calculated from Equation (1).

Table 4-1. Example of Rule Parameters for *Bright*. Values in the cells are the variation of percentage to the acoustic features of the input neutral utterance. Unlisted acoustic features in the table are not modified.

Acoustic Feature	SR1	SR2	SR3
Highest F0 (HP)	6.9%	10.9%	14.9%
Average F0 (AP)	7.5%	11.5%	15.5%
F2	3.3%	5.3%	7.3%
F3	4.2%	6.2%	8.2%

Base rule development for expressive speech (ER-base rules)

To verify the first type of information in the resulting relationship between expressive speech and semantic primitives, it is needed (1) to select the base rules of the significant semantic primitives to the percept of expressive speech, and (2) to consider the combination of the selected base rules. Both (1) and (2) can be considered from Figures 3-7 to 3-10. For (1), only those semantic primitives shown in Figures 3-7 to 3-10 are selected. For (2), they are represented as the weight and weight combination of semantic-primitive rules. That is, a higher weight value leads to a better perception of the expressive speech utterance. As explained previously in Section 3.4, the widths of the lines between the two layers of the model shown in the diagrams represent the weight values of the combinations. The weight value is higher for a thicker line and lower for a thinner line. The base rules of the semantic primitives were combined to form base rules for each expressive speech category and the values of these weight combinations, which are the slope of regression line that are shown in Table 3-8, which are in turn the slope of the regression line fitting the output of the fuzzy inference system that illustructed the relationship between semantic primitives and expressive speech categories. For example, the base rule for *Joy* is calculated by adding the various base rules of the appropriate semantic primitives, and then multiplying these by the appropriate weight values as shown below

ER-Base rule of *Joy* = (base rule of *Bright* * 0.101 + base rule of *Unstable* * 0.063 + base rule of *Clear* * 0.034 + base rule of *Quiet* * (-0.039) + base rule of *Weak* * (-0.036)) / 0.123

This formula is a linear function based on the non-linear fuzzy logic.

Intensity rule development for expressive speech (ER-intensity rules)

In order to create the intensity rules, the parameters of the semantic-primitive intensity rules were combined so that the morphed speech utterance could be perceived as having different levels of intensity of expressive speech. Three intensity rules (ER1, ER2, and ER3) were created. The utterance morphed by ER2 was supposed to be with stronger perception than that morphed by ER1; the utterance morphed by ER3 was supposed to be with stronger perception than that morphed by ER2. Thus, the changes to the value of each parameter of ER1 should be lower than the change to the value of each parameter of ER2, which in turn should be lower than ER3. For example, ER1 should combine weaker intensity rules of positively-correlated semantic primitives and stronger intensity rules of negatively- correlated semantic primitives.

More specifically, for each expressive speech category, intensity rule ER1 was created by combining intensity rule SR1 of the positively-correlated semantic primitives with intensity rule SR3 of the negatively-correlated semantic primitives. Intensity rule ER2 was created by combining intensity rule SR2 of positively-correlated semantic primitives with intensity rule SR2 of the negatively-correlated semantic primitives. Intensity rule ER3 was created by combining intensity rule SR3 of the positively-correlated semantic primitives with intensity rule SR1 of the negatively-correlated semantic primitives. Notice that because the perception of expressive speech categories has a different scheme of intensity rule combination than the perception of semantic primitives, intensity rules ER1 are not identical to expressive-speech base rules (ER-base rules). Table 4-2 shows an example of this way of combining intensity rules. As can be seen from Table 3-8 and Figure 3-7, *Joy* is positively correlated with *bright*, *unstable* and *clear*, but negatively correlated with *heavy* and *weak*. Therefore, ER1 for *Joy* can be created by combing the intensity rule SR1 of *bright*, SR1 of *clear*, SR1 of *unstable*, SR3 of *heavy*, and SR3 of *weak*. Along similar lines, ER2 for *Joy* can be created by combing intensity rules SR2 of *bright*, of *clear*, of *calm*, and of *weak*. The same weight and weight combination values were used when creating expressive-speech base rules for combining the expressive-speech intensity rules here.

Table 4-2. An example of semantic primitive rule combination.

Joy	Bright	Unstable	Clear	Heavy	Weak
ER1	SR1	SR1	SR1	SR3	SR3
ER2	SR2	SR2	SR2	SR2	SR2
ER3	SR3	SR3	SR3	SR1	SR1

4.2 Rule implementation

A speech morphing process was developed in order to implement the rules for morphing a neutral utterance into an utterance which could be perceived in terms of one and only one semantic primitive or expressive speech category. Figure 4-1 shows that for the speech morphing process, F0 contour, power envelope, and spectrum were extracted from the neutral speech signal by using STRAIGHT while segmentation information was measured manually. Next, acoustic features in terms of F0 contour, power envelope, spectrum and duration were modified according to the morphing rule. Finally, the modified F0 contour, power envelope, spectrum and duration were re-synthesized by using STRAIGHT to produce a morphed utterance.

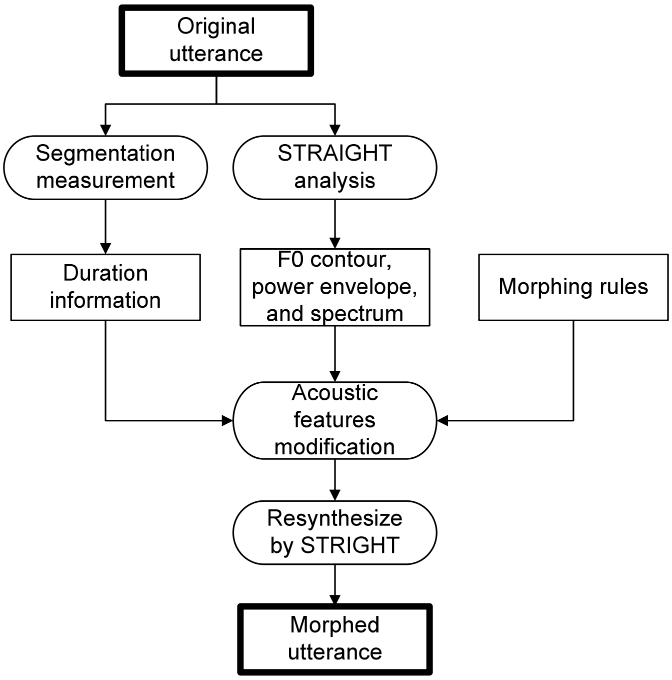


Figure 4-1. Process of morphing voices in STRAIGHT

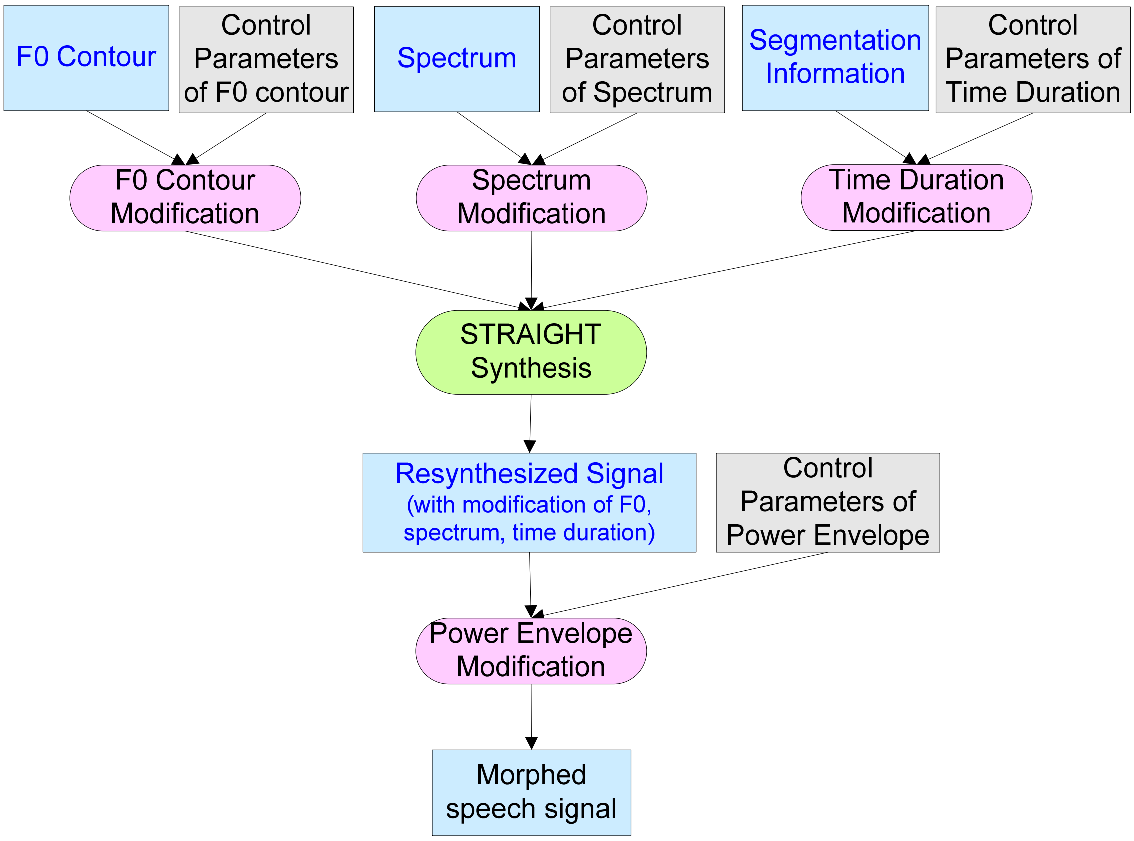
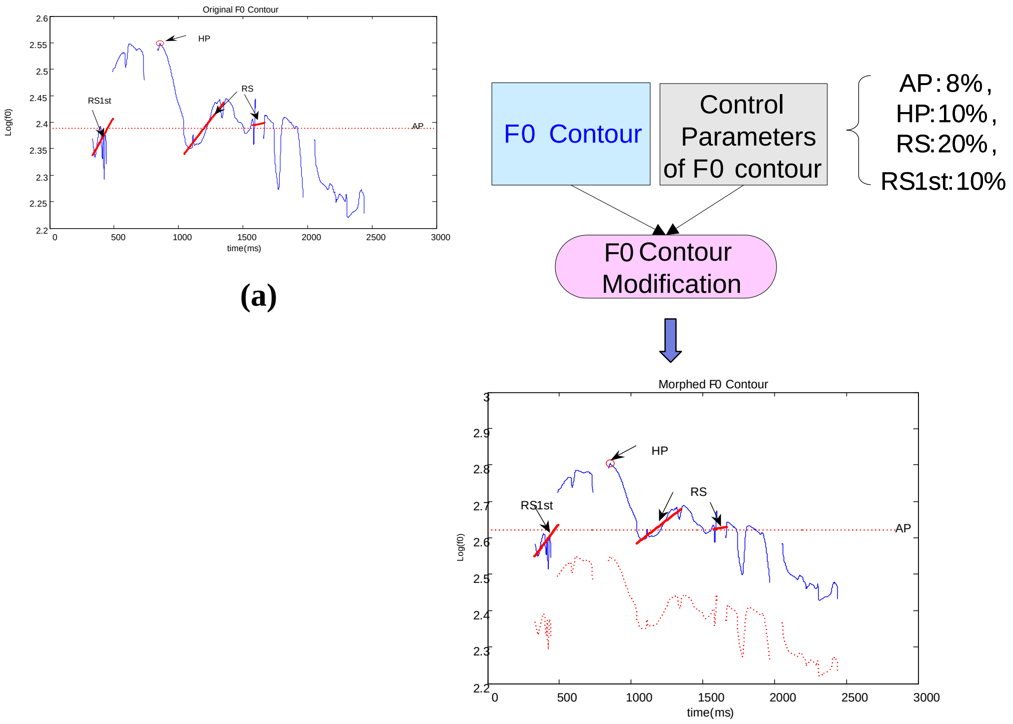

F0 Contour
Control Parameters of F0 contour
Spectrum
Control Parameters of Spectrum
Segmentation Information
Control Parameters of Time Duration
F0 Contour Modification
Spectrum Modification
Time Duration Modification
STRAIGHT Synthesis
Resynthesized Signal (with modification of F0, spectrum, time duration)
Control Parameters of Power Envelope
Power Envelope Modification
Morphed speech signal

Figure 4-2. Process of acoustic features modification.

Figure 4-2 illustrates a detailed processing flow of *acoustic features modification* in Figure 4-1. First, F0 contour modification, spectrum modification, and duration modification were processed. The modified F0 contour, spectrum, and duration are combined to produced a resynthesized voice which was followed by the power envelope modification. Sections 4.2.1 to 4.2.4 demonstrate more specifically about how F0 contour modification, spectrum modification, duration modification, and power envelope modification change the related acoustic parameters.

4.2.1 F0 contour modification

Figure 4-3 illustrates an example of F0 contour modification. The process changes four parameters, which are AP (average pitch), HP (highest pitch), RS (rising slope), and RS1st (rising slope of the first accentual phrase). The left-upper graph (a) shows the original F0 contour with those four acoustic features notated in red. By changing AP with 8%, HP with 10%, RS with 20%, and RS1st with 10%, the right-bottom graph (b) shows the modified F0 contour and its corresponding acoustic features. Also, the original F0 contour is shown in the dotted red line for purposes of comparison.



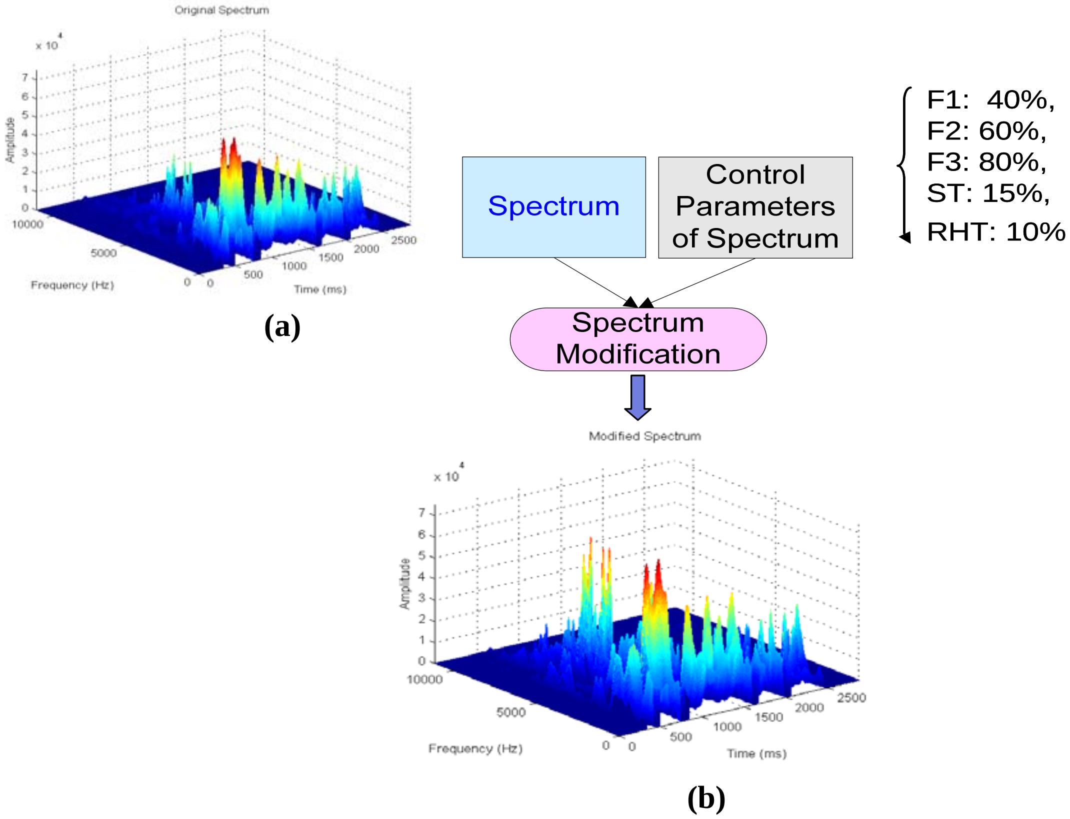
(a)

(b)

Figure 4-3. Demonstration of F0 contour modification.

4.2.2 Spectrum modification

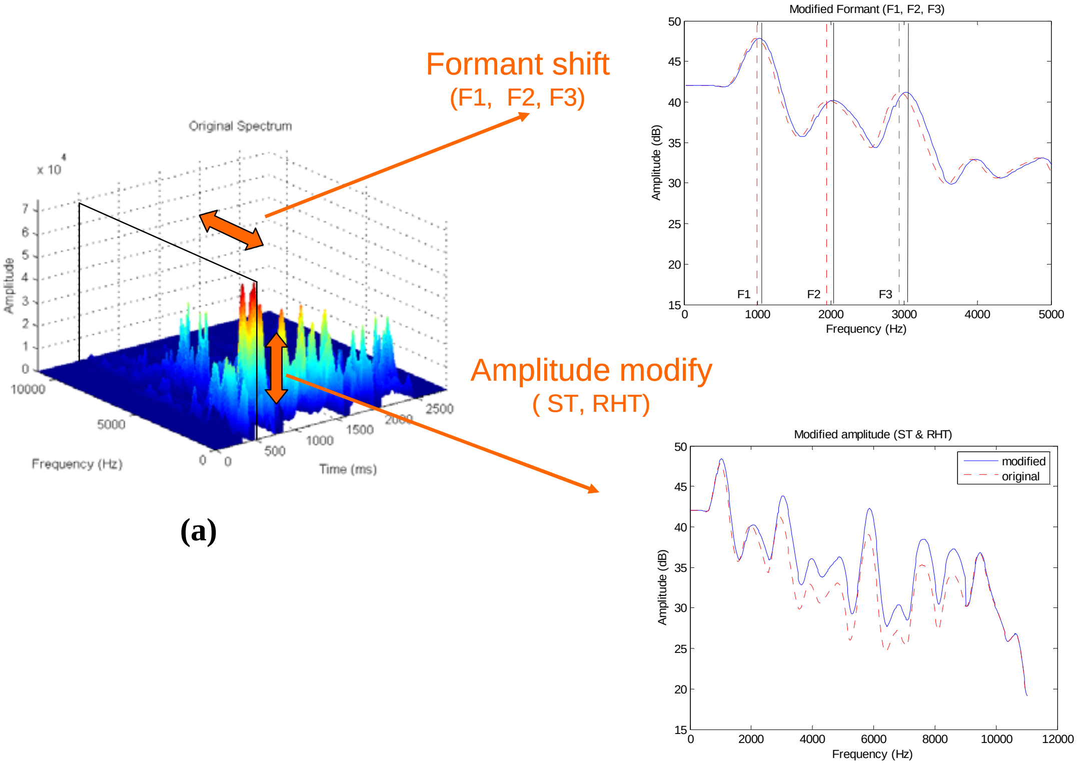
Figure 4-4 illustrates an example of spectrum modification. Spectrum modification involves modifying four parameters, which are F1 (the first formant), F2 (the second formant), F3 (the third formant), ST (spectral tilt), and RHT (the ratio between the average power in high frequency portion, which is over 3 kHz, and the average power). Especially, an algorithm developed by Nguyen et al. [84] is used in the formant shift. By changing parameters of the graph 4-4 (a), F1 with 40%, F2 with 60%, F3 with 80%, ST with 15%, and RHT with 10%, the graph 4-4 (b) in the right-bottom corner shows the modified spectrum. Figure 4-5 shows the fine changes of modified spectrum compared with the original spectrum. The upper graph (a) shows the formant frequency shift which is controlled by F1, F2, and F3. The bottom graph (b) shows changes of energy amplitude which are controlled by ST and RHT. In both graphs 4-5 (a) and 4-5 (b), the modified result is shown in the solid line and original contour is shown in the dotted line.



(a)

(b)

Figure 4-4. Demonstration of spectrum modification.



(a)

(b)

Figure 4-5. Formant shift and amplitude modify in spectrum modification

4.2.3 Duration modification

With respect to duration modification, the information of time segmentation of the original utterances should be manually measured first. This part of the work was the same as that for the acoustic feature analysis described in Section 3.3.2. The measurement included phone (i.e., segment) number, time (ms), phrase, vowel, and accent. Figure 4-6 illustrates an example of duration modification. Figure 4-6 (a) shows an example of time segmentation. In the table, the first row indicates the phone; the second row indicated the order of a phone, noted by -1 before the first phone; the third row indicates the start time of the second phone; the fourth row represents the phrase number that the phone belongs to; the fifth row indicates the vowel and consonant of the phone: 1:/a/, 2: /e/ 3: /i/ 4:/o/, 5:/u/ and 6: consonant; the sixth row indicates the accent of the phone: 1:high and 0:low. The duration modification first calculates the impact function according to the parameters of total length (TL), consonant length (CL), ratio between consonant length, and vowel length (RCV). In this example, TL has changed 14%, RCV 12%, and CL 3%; the impact function is calculated and shown in the blue line in the figure. The red dotted line is the non-changed function for comparison purposes. The impact function will be applied when re-synthesizing the modified F0 contour and the modified spectrum by STRAIGHT.

phone		/a/		t	a	r	a	sh	i	i
phone No.	-1	1	0	2	2	3	3	4	4	5
timePoint	328	403	481	493	585	609	727	839	942	1039
phrase	-1	1	1	1	1	1	1	1	1	1
vowel	-1	1	0	6	1	6	1	6	3	3
accent	-1	0	0	1	1	1	1	0	0	0

(a)

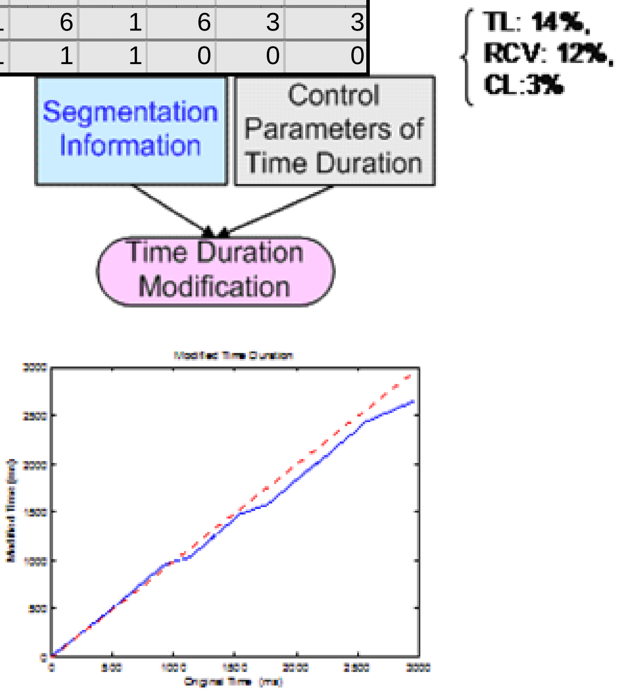


Figure 4-6. Duration modification

4.2.4 Power envelope modification

Figure 4-7 shows an example of the power envelope modification. It involves modifying four parameters, which are the power range in the accentual phrase (PRAP), the power range (PWR) and the rising slope of the first accentual phrase (PRS1st). The graph (a) in the left-upper corner shows the original signal. By changing the parameters of PWR with -8%, PRAP with -8%, PRS1st with -3%, the graph (b) in the right-bottom corner shows the modified speech signal.

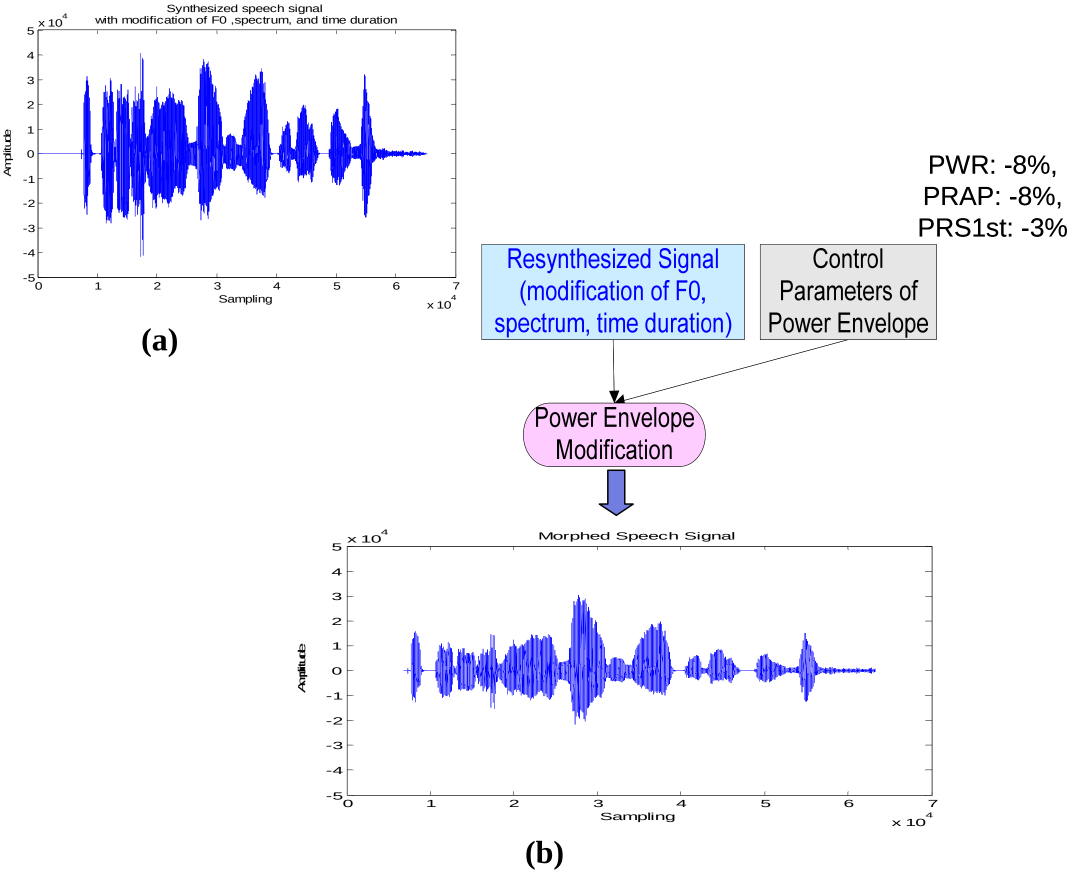


Figure 4-7. Power envelope modification

4.3 Perceptual experiments

4.3.1 Experiment for evaluating SR-base rule efficiency

To examine whether the selected acoustic features are significant to the percept of semantic primitives or not, an experiment was conducted for subjects to evaluate the morphed utterances by comparing them to the neutral utterances from which they were morphed.

Method

In this experiment, 17 morphed speech utterances were produced by implementing the created semantic primitive base rules, giving one morphed speech utterance for each semantic primitive. In addition, there was one neutral speech utterance. Subjects were ten male Japanese graduate students, who were different from those subjects of the three experiments described in Section 3.1. They had normal hearing ability and were asked

(a) to compare a morphed speech utterance with the neutral speech utterance and (b) to choose which utterance was most associated with a particular semantic primitive. The question was of the type, "Is (a) or (b) more 'bright'?" Paired stimuli were presented randomly to each subject through binaural headphones at a comfortable sound pressure level.

Results and discussion

The results shown in Table 4-3 indicate that most of the morphed speech utterances were perceived as the semantic-primitive intended by the morphed speech utterance. From these results, we can see which of the selected acoustic features significantly influenced the perception of the semantic primitives. These results suggest that the created base rules are effective. A comment about the two exceptional cases, semantic primitive *well-modulate*d and *monotonous*, which base rules were developed based on different conditions to select acoustic features (see Section 4.1.3). The results show that the speech morphed by the base rule of *well-modulated* can be well-perceived. However, the semantic primitive *monotonous* showed the lowest rate of accuracy. Perhaps this was because *monotonous* is most similar to the neutral voice, and it is difficult to morph a *monotonous* utterance into a "more" *monotonous* utterance.

Table 4-3. Experimental results of semantic-primitive rule evaluation.

Semantic Primitive	Accuracy Rates
bright	100%
dark	100%
high	100%
low	100%
strong	100%
weak	80%
calm	100%
unstable	100%
well-modulated	100%
monotonous	60%
heavy	90%
clear	80%
noisy	100%
quiet	90%
sharp	90%
fast	100%
slow	100%

4.3.2 Experiment for evaluating SR-intensity rule efficiency

To examine whether the impact direction and strength of the selected acoustic features are valid or not, an experiment was conducted in which subjects evaluated the morphed utterances by comparing the utterances created by the base rules (SR1) with those made by the intensity rules (SR2 and SR3).

Method

This experiment evaluates the validity of the intensity rules (SR1, SR2, and SR3) where for each semantic primitive, the utterance (SU2) created by SR2 should have stronger perception than the utterance (SU1) created by SR1, and the utterance (SU3) created by SR3 should have stronger perception than SU2. Stimuli include three morphed utterances (SU1, SU2, and SU3) for each semantic primitive and one neutral utterance.

SU1, SU2, and SU3 were morphed by following the intensity rules SR1, SR2, and SR3, respectively, which were described above. Scheffe’s method of paired comparison was used to evaluate the intensity of the semantic-primitive. Subjects were the same as in the previous experiment and were asked to evaluate which stimulus (A or B) had a stronger intensity (0 to 2 for B and 0 to -2 for A) of the semantic primitive according to a five-grade scale. A template of the questionnaire is shown in Figure 4-8.

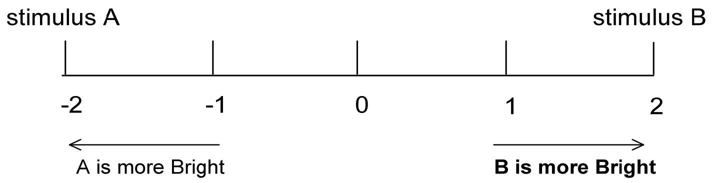


Figure 4-8. Template of the questionnaire for assessing intensity of the semantic primitive "bright"

Results and discussionFigure 4-9 shows the results of how listeners perceived the levels of intensity of each of the 4 utterances, e.g., one neutral and three morphed utterances. The numbers under the horizontal axis indicate the intensity levels of the semantic-primitive; labeled arrows indicate the level of intensity of perception of each of the morphed utterances (plus *neutral*). The results show that listeners were able to perceive four levels of intensity for each semantic primitive, except for *quiet*, for which only three levels were perceived. The difficulty in perceiving different intensity levels for *quiet* could be because the neutral utterances are intrinsically *quiet*.

Note that intensity levels of perception are congruent with what was intended by the intensity rules, i.e., neutral, SU1, SU2, SU3. These results suggest that by adjusting the parameters of the semantic-primitive rules, it is possible to control the intensity of the perception of the semantic primitives. They also suggest that the relationship between semantic primitives and acoustic features is a valid one.

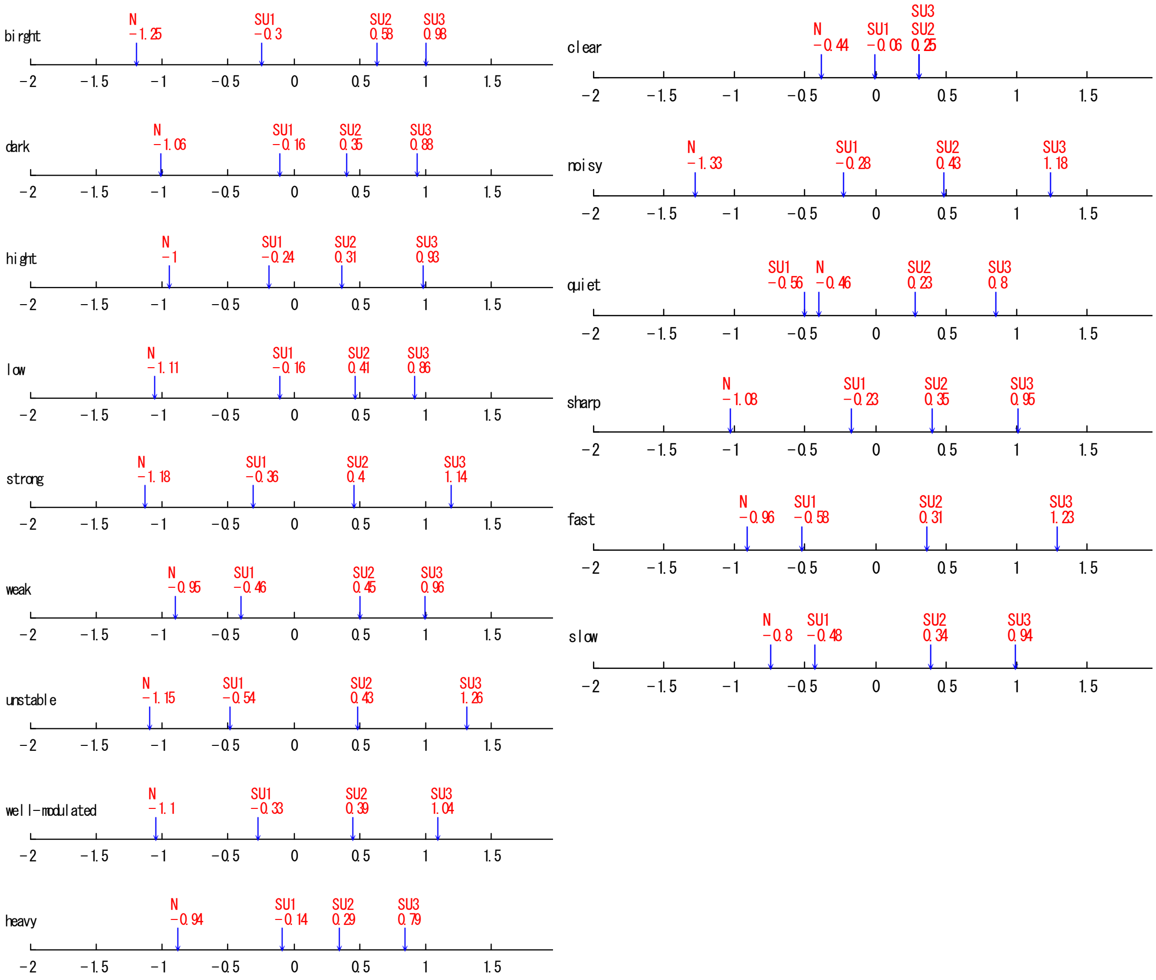


Figure 4-9. Experimental results of semantic-primitive rule evaluation. The intended intensity levels are N < SU1 < SU2 < SU3. That is, SU3 should have a higher level of intensity perception than SU2 than SU1 than N. N is the neutral utterance, and SUn represent semantic primitive utterances created by the semantic primitive intensity rules.

4.3.3 Experiment for evaluating ER-base rule efficiency

To examine whether the combination of semantic primitive rules are valid or not, an experiment is conducted so that subjects evaluate the morphed utterances by comparing them to the neutral utterances from which they were morphed.

Method

The experimental conditions including subjects were the same as in the previous experiments described above. The speech utterances were one neutral utterance and 4 morphed utterances that were morphed from the neutral utterance, one for each of the four expressive speech categories. Subjects were asked to compare (a) a morphed utterance with (b) the neutral voice and to choose which utterance was most associated with a specific expressive speech category. The questions asked were like “Is (a) or (b) more ‘joyful’? “

Results and discussion

Table 4-4 shows that the morphed speech utterances were perceived as the expressive speech category intended by the morphing process--100% accuracy rate for each of the expressive speech categories, except Sad (90%). This result suggests that the created base rules are effective, and moreover, that the combinations are appropriate.

Table 4-4. Experiment results of base rule evaluation

Expressive Speech Category	Accuracy Rate
Joy	100%
Cold Anger	100%
Sadness	90%
Hot Anger	100%

4.3.4 Experiment for evaluating ER-intensity rule efficiency

To examine whether the impact direction and strength of the selected semantic primitives are valid or not, an experiment is conducted so that subjects evaluate the morphed utterances by comparing them with the utterances created by the intensity rules (ER1, ER2 and ER3).

Method

This experiment evaluates the validity of the intensity rules (ER1, ER2, and ER3) where for each expressive speech category, the utterance (EU2) created by ER2 should have stronger perception of intensity than the utterance (EU1) created by ER1, and the utterance (EU3) created by ER3 should have stronger perception of intensity than EU2.

Stimuli include three morphed utterances (EU1, EU2, and EU3) for each expressive speech category plus one neutral utterance. EU1, EU2, and EU3 were morphed by following the intensity rules ER1, ER2, and ER3, respectively, which were described above. Scheffe’s method of paired comparison was used to evaluate the expressive speech intensity of the utterances. Subjects were the same as in the previous experiments, and evaluated the intensity of each utterance according to a five-scale rating.

Results and discussion

The results in Figure 4-10 show the four stimuli of each expressive speech category listed in ascending order of the intensity perception. The order is congruent with what was intended by the intensity rules. These results suggest that it is possible to control the intensity of emotional perception by adjusting the intensity of semantic primitives. Moreover, the perception of expressive speech categories appears to be related to the perception of semantic primitives. These results lend validity to the model proposed here in that it substantiates the relationship between semantic primitives and expressive speech categories.

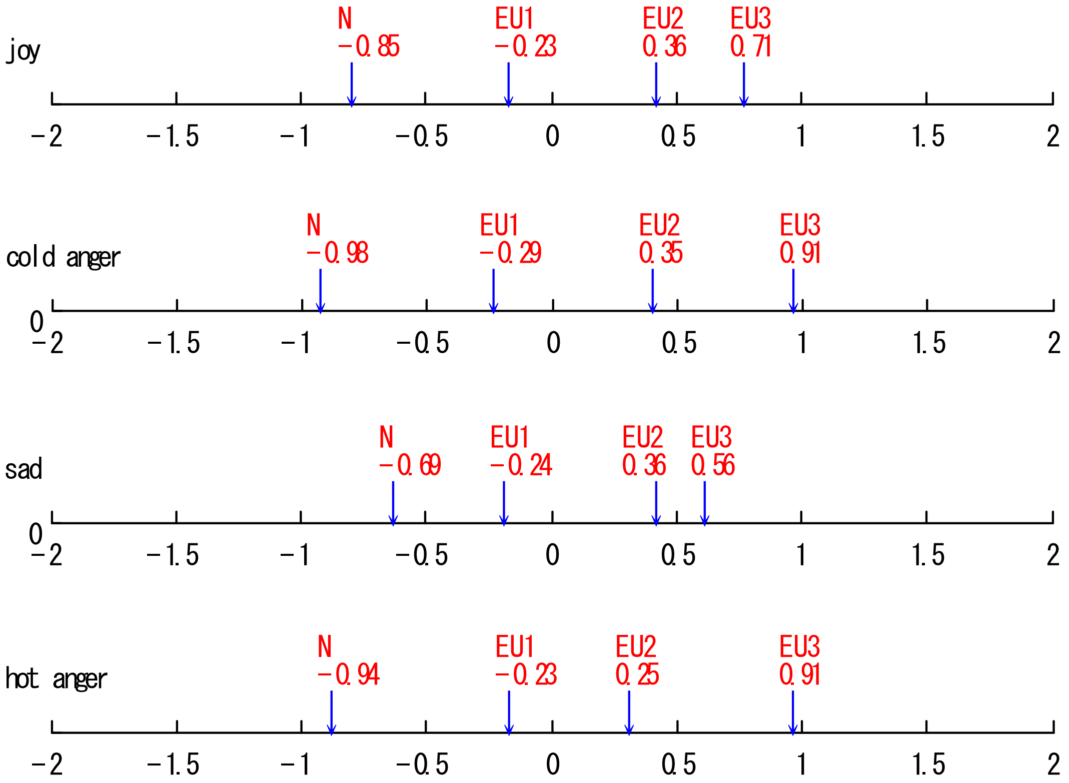


Figure 4-10. Experimental result of intensity rule evaluation. The intended intensity levels are N < EU1 < EU2 < EU3. That is, EU3 should have a higher level of intensity perception than EU2 than EU1 than N. N is the neutral utterance, EUn represent expressive speech utterances created by the expressive speech intensity rules.

4.4 General discussion

The relationships built in Chapter 3 are verified in this chapter. The results of the perceptual experiments show that neutral utterances morphed by the rules developed from the fuzzy inference systems and measurements of acoustic features can generate the perception of intended semantic primitives and expressive speech categories. Moreover, the changes of strength and impact direction also can change the intended intensity levels of perception. Therefore, combining the results of the building and the verification of the model provides a solid support for the first assumption we made in Chapter 1, that is, that people do not perceive expressive speech directly from the change of acoustic features but rather from the perception of semantic primitives.

4.5 Summary

This chapter provides a complete description of the work with regard to the verification of the proposed model. In order to accomplish the work, there are three sub tasks accomplished.

- To simplify the manipulation of acoustic features in the morphing process, rules are created from the built relationships. There were two types of information embedded in these rules. The first type is the significance of one element (acoustic feature or semantic primitive) in a layer to the perception of one element (semantic primitive or expressive speech category) in another layer. The second type is the impact direction and strength of one significant element.
- To modify a neutral utterance according to the developed morphing rules. STRAIGHT provides the tool for morphing neural utterances by the created rules.
- To conduct perceptual experiments for examining the modified utterance in terms of the perception of expressive speech categories and semantic primitives.

For the first task, Section 4.1 starts with the concept description. This concept clarified what information in the built relationships should be transformed to the rules. The principles of rule development are then proposed. These principles clearly stated what should be empathized when developing rules. From the concept and the principles, four groups of rules are created:

(1) Base rules for semantic primitives (SR-base rules)

(2) Intensity rules for semantic primitives (SR-intensity rules)

(3) Base rules for expressive speech (ER-base rules)

(4) Intensity rules for expressive speech (ER- intensity rules)

For the second task, the process of speech morphing and the implementation details are described in Section 4.2. It starts with a diagram that shows that F0 contour, power envelope, and spectrum were extracted from the neutral speech signal by using STRAIGHT while segmentation information was measured manually. It then uses another diagram to show the details of acoustic feature modification. Different algorithms are developed for modifying F0 contour, power envelope, duration and spectrum.

For the third task, four perceptual experiments are conducted. Each perceptual experiment verifies one group of rules. The methods, results, and discussion of the experiments are described in Section 4.3. One significant aspect of this research work is that not only the intended perception is verified, thus providing a more solid support to our assumption, but also we can know the intensity level of the intended perception. The experiment results shown in Sections 4.3.1 to 4.3.4 suggest that the SR-rules and ER-rules created from the built relationships support the first assumption we made in this research work. The results also showed that the obligatory principles are followed.

- The rules are monotonous because subjects can identify the intended perception clearly.
- The rules are general because the intended perception of the morphed utterances can be clearly identified by different subjects.
- The rules are dynamic because the combination of SP-rules to ER-rules can create morphed utterances with intended categories of expressive speech.

Combining the results of the building and the verification of the model provides a solid support for the first assumption we made in Chapter 1, that is, that people do not perceive expressive speech directly from the change of acoustic features but rather from the perception of semantic primitives.

Chapter 5. Application of the Three-Layer Model

This chapter describes the model application. It provides support for the second assumption that states people who are from different cultures/native-language background have common characteristics in perceiving non-linguistic information of expressive speech categories. To this end, the three-layer model proposed in this research work is applied. In this application, instead of using Japanese utterances and Japanese listener subjects, listeners with a different native language/culture background were used, while at the same time, using the same procedure as described in the first part of the thesis. By comparing the results of the different sets of application environments, we expect to obtain a systematic understanding of common features in perception of expressive speech catteries by people with different culture/native language backgrounds. The significance of this is that it will help further clarify the vagueness nature of human perception.

The background and related work of this application is introduced in Section 5.1. This introduction reviews previous studies about commonalities and differences in cross-linguistic/cross-cultural perception of expressive speech. However, most of the previous work overlooked the vagueness nature of human perception, although many of their results showed that specific acoustic cues are common to the judgment of expressive speech across people with different culture/native-language background. After this review, the motivation of this application is described.

The application process is similar to that described in Chapter 3. However, instead of using Japanese utterances and Japanese listener subjects, we used the same process with the same stimuli but Taiwanese listeners. By comparing the results of these two sets of application environments, we expect to find support for the assumptions made about similarities in human perception of expressive speech. This process includes

- Three experiments are carried out in order to determine which adjectives are suitable for describing expressive speech.
- To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech, a fuzzy inference system will be built.
- In order to support the relationship between semantic primitives and acoustic features, analysis of a large number of acoustic features will be necessary.

This process and the comparison are described in Sections 5.2 and 5.3. Section 5.5 discusses the results.

5.1 Introduction of model application

In speech communication, linguistic information requires listeners getting acquainted with one language; however, non-linguistic information may be different. Even without the acquaintance of one language, we can still judge a speaker's age and gender or even perceive the expressive speech categories of a voice, such as *joy* or *sadness*. However, sometimes we also find misunderstanding of emotional communication between people. Therefore, one question is what common features (or different features) exist in non-linguistic verbal information to help (or hamper) people make this judgment.

5.1.1 Related work

As reviewed in Chapter 1, much of the previous work focused on the relationship between acoustic features and expressive speech categories. In addition, some research also showed that women can perceive emotion better than men [17]. There is a growing body of work on the study of the expressive speech perception among different native languages/cultures (e.g., Erickson [19]), which indicate that there are both similarities and differences in expressive speech perception across languages/cultures. For instance, Erickson and Maekawa [20] reported that even without the acquaintance to a language, listeners can often perceive the intended emotion from listening to just the voice, especially when the intended emotion of the foreign utterance has an F0 contour similar to that of the listener's native language. However, sometimes this very similarity creates problems in that the same intonational contour can be interpreted by speakers of different languages very differently and leads to confusions, as is discussed also by Shochi et al. [69]. In addition, Erickson et al. [21] found that the acoustic features, especially the parameters AQ (amplitude quotient) and F0, can be perceived differently depending on the listener's native language. Sakuraba [63], on the other hand, showed that high-pitched voices are perceived by both Japanese and English to express *surprise* and *joy*. Dromey et al. [17] suggested that there is a difference among English mother tongue polyglots, other mother tongue listeners, and English mother tongue monoglots when perceiving emotion in speech. However, Beier and Zautra [7] found agreement across cultures on expressive speech categories perception and suggested the possibility of universal characteristics in expressive speech categories perception.

5.1.2 Motivation

As stated before, one important characteristic in human perception is the vagueness nature of humans, which has been overlooked in previous research. Therefore, a next step in the investigation of expressive speech perception is a model for systematically understanding the common features existing between people with different native languages/cultures, which also takes the vagueness nature into consideration. The benefit of this would be for improving global communication and developing tools for multi-language expressive speech recognition and synthesis. For example, it helps us to control the common features for creating an expressive speech synthesizer or category detection system. The three-layered model, which was proposed in this research work, is used to study the common features. This multi-layered approach can clarify the relationship between physical characteristics of voice and vagueness nature of human perception.

More specifically, we hope to better understand the relationship between semantic primitives and acoustic features, as well as that between semantic primitives and expressive categories. In this way we will have a better understanding of the similarities and differences in how Japanese and Taiwanese listeners perceive the physical characteristics of a voice as well as in how they perceive expressive speech categories.

5.2 Three experiments

To achieve the research purpose, the first relationship of the perceptual model between expressive speech and semantic primitives for Taiwanese listeners is built by conducting three experiments. For clarity, they are named Experiment 5, 6, and 7, which are analogous to Experiments 1, 2, and 3 described in Section 3.1.

The goal of Experiment 5 is to show that even without the acquaintance of one language, people can perceive the intended expressive speech categories of that language. The goal of Experiment 6 is to show the reliability of the voice database used for Taiwanese subjects. The results of this experiment should show that Taiwanese can also clearly distinguish the expressive speech categories of the Japanese utterances they heard, in a way similar to how Japanese subjects do. The goal of Experiment 7 is to determine suitable semantic primitives for the perceptual model. To clarify which adjective is more appropriate for describing expressive speech and how each adjective was related to each category of expressive speech, the selected semantic primitives are superimposed into the perceptual space built in Experiment 2 by applying a multiple regression analysis.

To achieve each of these goals, in Experiment 5, Taiwanese subjects are asked to evaluate the perceived categories of utterances. In Experiment 6, Taiwanese subjects are asked to evaluate utterances selected according to the results of Experiment 1. In Experiment 7, Taiwanese subjects are asked to evaluate utterances by the semantic primitives they perceive. Finally, from the experimental results, fuzzy inference systems are built for representing the relationship between expressive speech and semantic primitives.

As a reference, some details are repeated again in the following sections. In addition, the results of both sets of experiments are listed together for comparison.

5.2.1 Experiment 5

This experiment aims at showing that even without the acquaintance of one language, people can perceive the intended expressive speech categories of that language.

Method

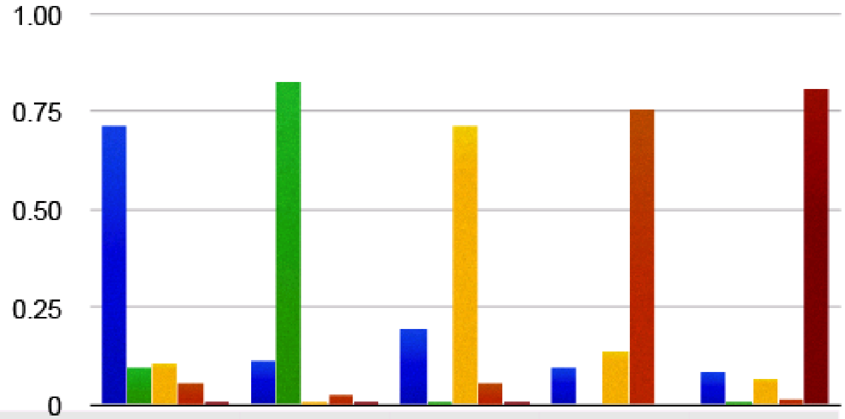
Stimuli were selected from the same database produced and recorded by Fujitsu Laboratory in Section 3.1. Subjects were 20 Taiwanese, 10 males and 10 females, who did not understand Japanese. They rated the utterances according to how strongly they perceived each of the five expressive speech categories, on a scale of 1 to 5. If a subject perceived that an utterance belonged to one expressive speech category without any doubt, then the subject gave it 5 points. Conversely, if the subject was confused within two or even more expressive speech categories, then the subject divided the 5 points among the five expressive speech categories according to what seemed appropriate. The stimuli were randomly presented to each subject through binaural headphones at a comfortable sound pressure level.

Results

Table 5-1 lists the results of Experiment 5 as a confusion matrix of each intended expressive speech category. From the diagonal lines shown in Table 5-1 and Table 5-2 (which give the evaluations by Japanese listeners), we know that for each category, the highest percentage belongs to the intended category. When comparing Table 5-1 and Table 5-2 both Taiwanese and Japanese have similar patterns of confused categories: the one most confused with *Joy* was *Neutral* (both are 12%), the one most confused with

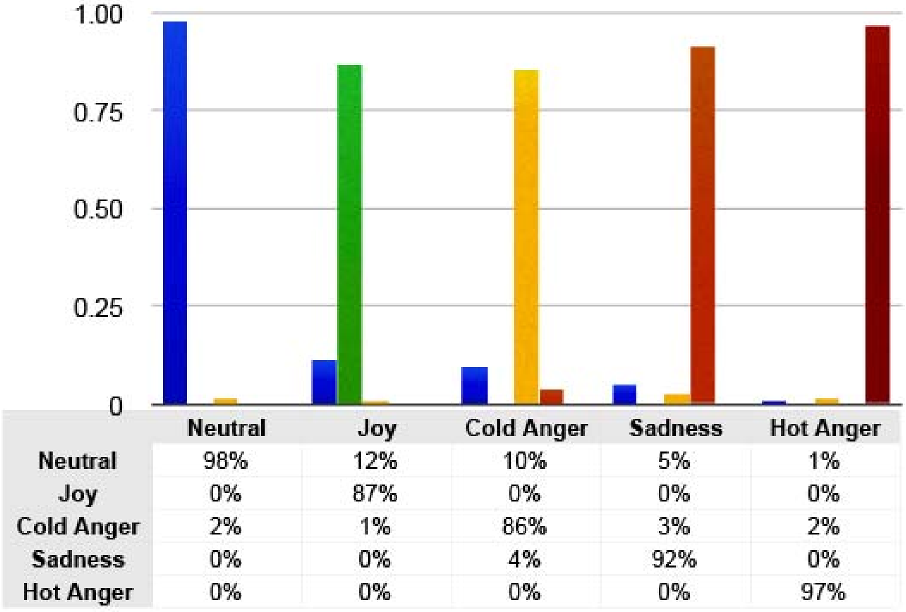
Cold Anger was *Neutral* (20% and 10%), and the ones most confused with *Sadness* are *Neutral* (10% and 5%) and *Cold Anger* (14% and 3%). These results suggest that even without the acquaintance of one language, people can perceive the intended categories of that language. This also implies the existence of common features in non-linguistic information. However, Japanese listeners can well perceive *Neutral* (98%) and *Hot Anger* (97%), while Taiwanese listeners confused *Neutral* with *Joy* (10%) and *Cold Anger* (11%), and *Hot Anger* with *Neutral* (9%) and *Cold Anger* (7%).

Table 5-1. Percentage of ratings of the 5 intended categories. Subjects were Taiwanese.



	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	72%	12%	20%	10%	9%
Joy	10%	83%	1%	0%	1%
Cold Anger	11%	1%	72%	14%	7%
Sadness	6%	3%	6%	76%	2%
Hot Anger	1%	1%	1%	0%	81%

Table 5-2. Percentage of ratings of the 5 intended categories. Subjects were Japanese.



	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	98%	12%	10%	5%	1%
Joy	0%	87%	0%	0%	0%
Cold Anger	2%	1%	86%	3%	2%
Sadness	0%	0%	4%	92%	0%
Hot Anger	0%	0%	0%	0%	97%

5.2.2 Experiment 6

This experiment aims at showing the reliability of the voice database used for Taiwanese subjects.

Method

Stimuli were 15 utterances chosen according to the ratings in Experiment 5. For each of the five expressive speech categories, three utterances were selected: (1) one that was most confused, (2) one that was least confused, and (3) one that fell in the middle. The subjects were identical to those who participated in Experiment 1. Scheffe's method of paired comparison was used. Subjects were asked to rate each of the $15 \times 14 = 210$ utterance pairs on a 5-point Liker-type scale (from -2 to 2, including 0, -2 = totally different, 2 = extremely similar) according to how similar they perceived them to be. The pair-wise stimuli were randomly presented to each subject through binaural headphones. The SPSS 11.0 MDS ALSCAL procedure, non-metric model of Shepard and Kruskal using the symmetric, matrix conditional and ordinal options, was applied to the ratings.

Results

Figure 5-1 shows the distribution of utterances in the resulting 3-dimensional perceptual space (STRESS value was 7%). In the figure, one circle represents one utterance. The symbols ’N’, ’J’, ’C’, ’S’, and ’H’ stands for the distributions of the utterances of the five emotional categories. Figure 5-2 illustrates the results of the corresponding Experiment 2, where subjects were Japanese. As the distribution shows, all categories of expressive speech are separated clearly and the utterances of the same category are close to each other, which means the distribution in the perceptual space can appropriately represent the similarity of the utterances and the position of each emotion. Therefore, it is reliable for using to determine the semantic primitives suitable for Experiment 7. The results of Experiment 6 also suggest that even without the acquaintance of one language, people can for the most part clearly identify the utterances with the same intended category.

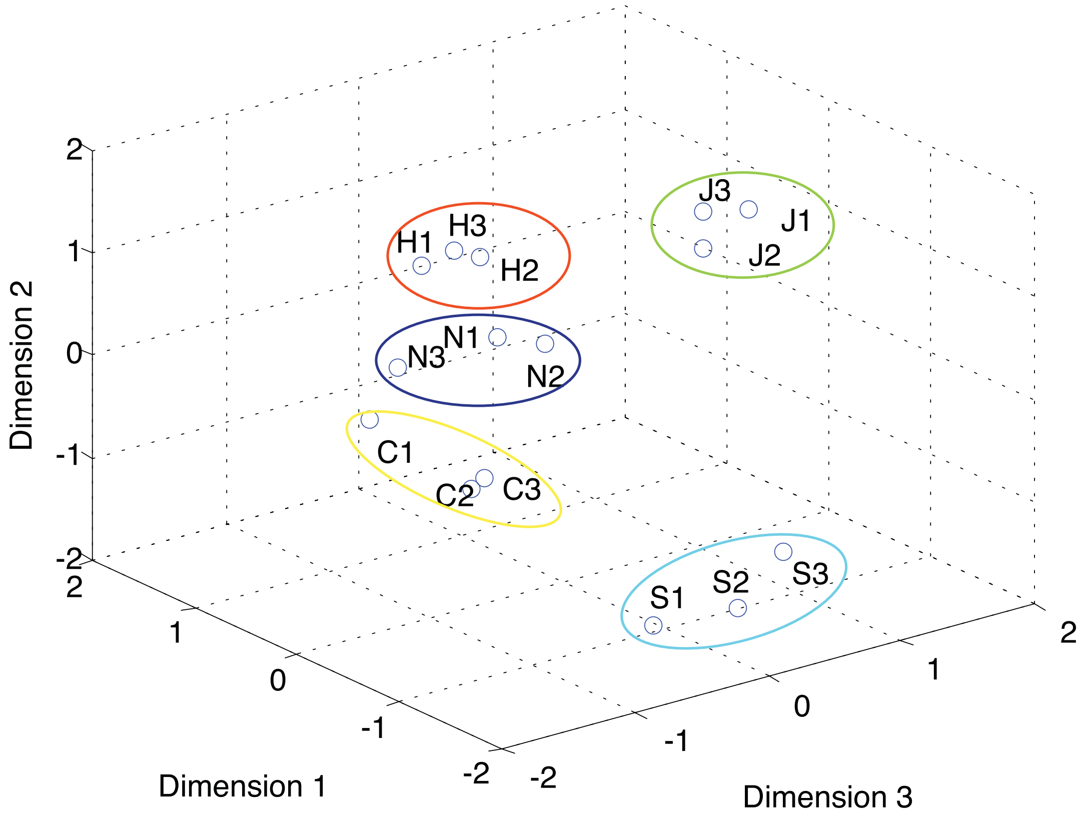


Figure 5-1. The resulting perceptual space of utterances in different categories of expressive speech of Experiment 6. Subjects are Taiwaness.

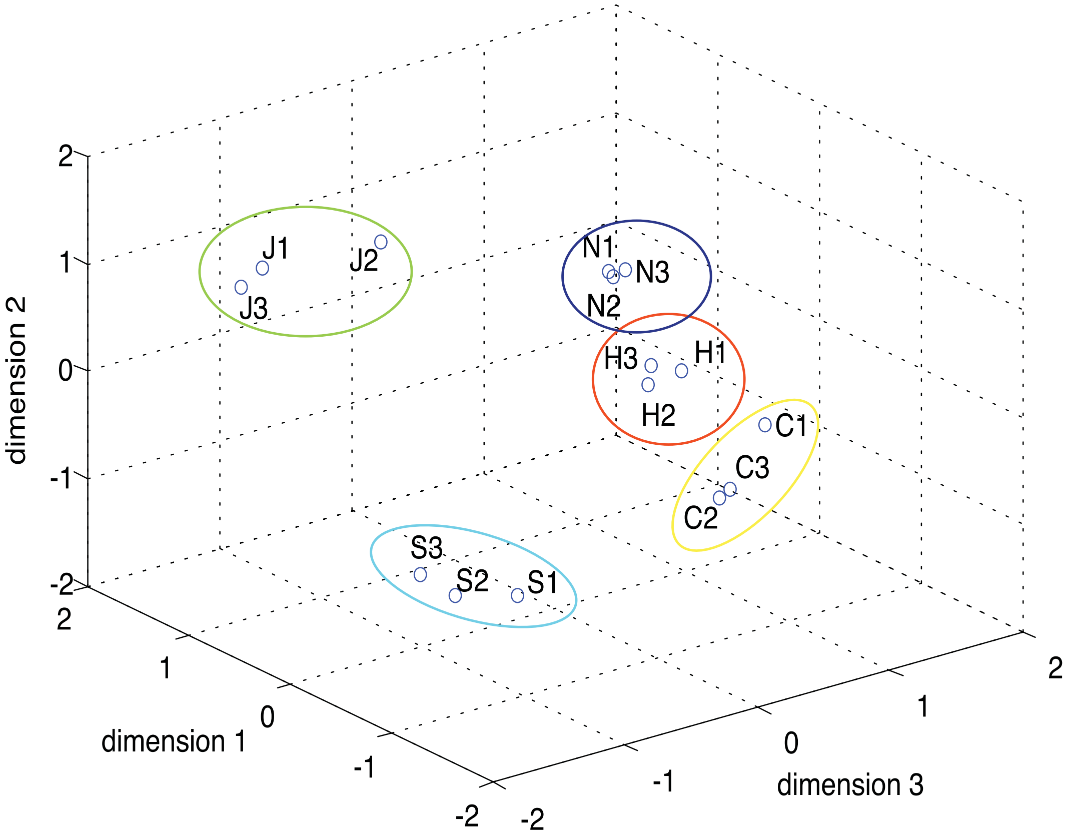


Figure 5-2. The resulting perceptual space of utterances in different categories of expressive speech of Experiment 2. Subjects are Japanese.

5.2.3 Experiment 7

Experiment 7 aims at determining suitable semantic primitives for the perceptual model. To clarify which adjective is more appropriate for describing expressive speech and how each adjective was related to each category of expressive speech, the selected semantic primitives are superimposed into the perceptual space built in Experiment 7 by the application of a multiple regression analysis.

Method

The stimuli and subjects were the same as in Experiment 6. The same 34 adjectives which were judged by Japanese in the previous work were translated into Mandarin. The stimuli were randomly presented to each subject through binaural headphones at a comfortable sound pressure level. Subjects were asked to rate each of the 34 adjectives on a 4-point scale (0: very appropriate, 3: very not appropriate) when they heard each utterance, indicating how appropriate the adjective is for describing the utterance they heard. Equation (1) in Section 3.1.3 is the regression equation for the superimposing.

Results

Semantic primitives were selected according to the same three criteria described in Section 3.1.3. The selected semantic primitives by Taiwanese subjects are shown in Table 5-3. Japanese results are shown again in Table 5-4. The first ten semantic primitives that are shown in Table 5-3 and Table 5-4 are identical. These results suggest that semantic primitives are a suitable tool for expressive speech description because people with different native languages/cultures tend to use the same set of semantic primitives for expressive utterance description. These results are used to build a fuzzy inference system for representing the relationship between expressive speech and semantic primitives.

Table 5-3. Semantic primitives selected in Experiment 7 by Taiwanese subjects

Adjective (Mandarin)	Adjective (English)
明亮的	bright
灰暗的	dark
低的	low
重的	heavy
明亮的	clear
強的	strong
弱的	weak
沉著的	calm
浮動的	unstable
慢的	Slow
堅硬的	Hard
鈍的	Dull
流暢的	Fluent
光滑的	Smooth
沙沙的	Raucous
混濁的	Muddy

Table 5-4. Semantic primitives selected in Experiment 3 by Japanese subjects

Adjective (Japanese)	Adjective (English)
明るい	bright
暗い	dark
声の低い	low
重い	heavy
明らかな	clear
強い	strong
弱い	weak
落ち着いた	calm
落ち着きのない	unstable
ゆっくり	slow
声の高い	high
抑揚のある	well-modulated
単調な	monotonous
うるさい	noisy
静かな	quiet
鋭い	sharp
早い	fast

5.3 Fuzzy inference system

To clarify the common semantic primitives used when deciding expressive speech categories by Taiwanese and Japanese, the relationship between expressive speech and semantic primitives is built by using the same method of fuzzy inference system (FIS) as described in Section 3.2. Therefore only the results are shown here.

Table 5-5 lists the five semantic primitives for each expressive speech category. For each semantic primitive, there are three positive correlations (which are the ones that showed the highest correlation values with a positive slope) and two negative ones (which showed the highest correlation values with a negative slope). The corresponding results of Japanese subjects are listed in Table 5-6.

Table 5-5. The related semantic primitives of each expressive speech category selected by Taiwanese S's

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
dull	-0.181	weak	-0.254	fluent	-0.498	bright	-0.122	muddy	-0.26
heavy	-0.117	low	-0.185	bright	-0.121	smooth	-0.295	heavy	-0.186
bright	0.115	clear	0.178	slow	0.164	heavy	0.179	fluent	0.118
clear	0.234	calm	0.288	weak	0.212	strong	0.181	unstable	0.17
smooth	0.256	smooth	0.44	muddy	0.384	raucous	0.267	hard	0.325

Table 5-6. The related semantic primitives of each expressive speech category selected by Japanese S's.

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
heavy	-0.329	quiet	-0.039	sharp	-0.079	slow	-0.231	calm	-0.063
weak	-0.181	weak	-0.036	strong	-0.049	monotonous	-0.073	quiet	-0.047
calm	0.103	clear	0.034	quiet	0.044	well-modulated	0.091	sharp	0.103
clear	0.127	unstable	0.063	weak	0.074	fast	0.153	unstable	0.12
monotonous	0.27	bright	0.101	heavy	0.061	heavy	0.197	well-modulated	0.124

Comparison between Table 5-5 and Table 5-6 helps us find the common features of semantic primitives. Those semantic primitives that influence the perception of the expressive speech category for both Taiwanese and Japanese listeners appear in Table 5-5 and Table 5-6 as shaded. The semantic primitives *heavy* and *clear* influence the perception of *Neutral* for both Taiwanese and Japanese listeners; specifically, it is positively correlated to *clear*, and negatively correlated to *heavy*. Similarly, *Joy* is correlated with *weak* and *clear* but in opposite directions. *Cold Anger* is positively correlated to *weak*. *Sadness* is positively correlated to heavy. *Hot Anger* is positively correlated to *unstable*. These results suggest that for people with different native languages/cultures, the perception of expressive speech categories are affected by a common set of semantic primitives.

5.4 Analysis of a large number of acoustic features

To clarify the common acoustic features that can influence the semantic primitives used by Taiwanese and Japanese, the method is identical to that described in Section 3.3.2, which calculates the same set of acoustic features - F0 contour, power envelope, and power spectrum.

5.4.1 Correlation analysis of acoustic measurements with semantic primitives

The same 16 acoustic features in Section 3.3.3 were also measured. As a reference, they are repeated again here. Four involved F0–mean value of rising slope (RS), highest pitch (HP), average pitch (AP) and rising slope of the first accentual phrase (RS1st); four involved power envelope–mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (PRS1st), the ratio between the average power in high frequency portion (over 3 kHz) and the average power (RHT); five involved the power spectrum–first formant frequency (F1), second formant frequency (F2), third formant frequency (F3), spectral tilt (SPTL), spectral balance (SB); and three involved duration total length (TL), consonant length (CL), ratio between consonant length and vowel length (RCV). Four acoustic features are measured in each accentual phrase of an utterance, i.e. RS1st, PRAP, RS, and PRS1st, and the other 12 acoustic features are in an utterance.

A correlation between the 16 acoustic features and the 16 semantic primitives was done. Correlation coefficient values that have at least one correlation coefficient over 0.6 are considered significant and are shadowed in Table 5-7. Comparison with Table 5-7 (Taiwanese listeners) and

Table 5-8 (which gives the correlation coefficients by Japanese listeners) suggests that F0 plays an important role in the perception of semantic primitives. This is consistent with previous work that showed F0 is significant for the judgment of expressive speech perception [30, 44 58, 65, 82], which suggests the possibility of the role semantic primitives plays in expressive speech perception. 6 of the 10 semantic primitives are associated with the same two acoustic features that have the highest correlations: *bright*, *dark*, *low*, *heavy*, and *clear* are associated with average pitch (AP) and highest pitch (HP), and *strong* is associated with power range (PWR) and mean value of power range in accentual phrase (PRAP). The first 10 semantic primitives that were shared by both Taiwanese and Japanese listeners have the same valence (i.e., positive or negative correlation).

Table 5-7. Correlation coefficients between semantic primitives and acoustic features, Taiwanese S's.

	bright	dark	low	heavy	clear	strong	weak	calm	unstable	slow	hard	dull	fluent	smooth	raucous	muddy
AP	0.88	-0.84	-0.88	-0.83	0.72	0.61	-0.59	-0.45	0.47	-0.84	0.21	-0.82	0.77	0.39	0.61	-0.76
HP	0.85	-0.82	-0.87	-0.80	0.68	0.68	-0.63	-0.52	0.56	-0.86	0.29	-0.81	0.74	0.32	0.67	-0.73
RS	0.62	-0.64	-0.67	-0.54	0.46	0.69	-0.67	-0.55	0.57	-0.70	0.49	-0.63	0.48	0.09	0.59	-0.53
RS1st	0.73	-0.75	-0.80	-0.69	0.61	0.66	-0.65	-0.46	0.47	-0.79	0.37	-0.72	0.61	0.17	0.60	-0.68
PWR	0.66	-0.73	-0.78	-0.66	0.50	0.83	-0.76	-0.62	0.64	-0.80	0.58	-0.73	0.57	-0.01	0.77	-0.62
RHT	-0.01	-0.06	-0.12	0.04	-0.19	0.63	-0.35	-0.67	0.76	-0.26	0.66	-0.05	-0.12	-0.53	0.71	0.11
PRS1st	0.74	-0.80	-0.80	-0.74	0.61	0.58	-0.73	-0.34	0.35	-0.75	0.35	-0.80	0.69	0.22	0.47	-0.75
PRAP	0.58	-0.70	-0.73	-0.58	0.45	0.83	-0.80	-0.60	0.61	-0.77	0.67	-0.68	0.47	-0.10	0.70	-0.57
F1	0.63	-0.61	-0.64	-0.61	0.48	0.49	-0.44	-0.33	0.41	-0.58	0.23	-0.58	0.53	0.15	0.48	-0.46
F2	0.48	-0.32	-0.32	-0.41	0.42	-0.05	0.08	-0.02	0.02	-0.30	-0.37	-0.29	0.42	0.57	0.06	-0.29
F3	0.56	-0.42	-0.40	-0.44	0.43	0.22	-0.15	-0.20	0.24	-0.41	-0.10	-0.44	0.48	0.42	0.29	-0.32
SPTL	-0.46	0.45	0.54	0.39	-0.23	-0.67	0.39	0.67	-0.74	0.52	-0.48	0.38	-0.24	0.13	-0.76	0.29
SB	0.39	-0.39	-0.47	-0.34	0.22	0.60	-0.33	-0.63	0.68	-0.51	0.42	-0.34	0.25	-0.10	0.73	-0.26
TL	-0.34	0.50	0.47	0.44	-0.38	-0.36	0.63	0.19	-0.12	0.58	-0.26	0.50	-0.46	-0.12	-0.17	0.55
CL	-0.56	0.65	0.63	0.64	-0.52	-0.41	0.60	0.23	-0.17	0.67	-0.21	0.65	-0.59	-0.19	-0.29	0.69
CL/VL	-0.71	0.74	0.72	0.75	-0.66	-0.40	0.47	0.10	-0.12	0.61	-0.09	0.73	-0.66	-0.32	-0.30	0.71

Table 5-8. Correlation coefficients between semantic primitives and acoustic features, Japanese S's.

	Bright	dark	low	heavy	clear	strong	weak	calm	unstable	slow	hight	well-modulated	monotonous	noisy	quiet	sharp	fast
AP	0.71	-0.88	-0.91	-0.78	0.76	0.33	-0.54	-0.66	0.60	-0.62	0.87	0.41	-0.10	0.52	-0.70	0.34	0.35
HP	0.69	-0.88	-0.89	-0.73	0.74	0.42	-0.56	-0.72	0.67	-0.62	0.90	0.50	-0.18	0.60	-0.73	0.44	0.42
RS	0.44	-0.64	-0.60	-0.40	0.44	0.56	-0.54	-0.74	0.67	-0.56	0.70	0.54	-0.32	0.63	-0.67	0.59	0.50
RS1st	0.50	-0.79	-0.78	-0.61	0.66	0.45	-0.58	-0.67	0.60	-0.51	0.77	0.42	-0.10	0.57	-0.72	0.47	0.24
PWR	0.43	-0.74	-0.65	-0.41	0.57	0.70	-0.66	-0.80	0.78	-0.57	0.74	0.59	-0.27	0.76	-0.79	0.69	0.38
RHT	-0.10	-0.05	0.00	0.24	-0.10	0.68	-0.14	-0.55	0.67	-0.16	0.29	0.52	-0.41	0.72	-0.29	0.68	0.36
PRS1st	0.48	-0.80	-0.70	-0.56	0.64	0.45	-0.78	-0.61	0.51	-0.64	0.64	0.27	-0.01	0.44	-0.78	0.42	0.37
PRAP	0.31	-0.67	-0.56	-0.30	0.47	0.73	-0.67	-0.77	0.73	-0.55	0.62	0.55	-0.26	0.76	-0.78	0.73	0.31
F1	0.41	-0.64	-0.60	-0.49	0.47	0.25	-0.39	-0.49	0.52	-0.29	0.59	0.17	0.10	0.43	-0.52	0.29	0.30
F2	0.60	-0.41	-0.56	-0.66	0.44	-0.31	0.07	-0.11	0.07	-0.06	0.50	0.08	0.05	-0.03	-0.09	-0.27	0.11
F3	0.60	-0.47	-0.54	-0.55	0.49	0.01	-0.15	-0.33	0.33	-0.10	0.61	0.33	-0.16	0.23	-0.29	0.02	0.27
SPTL	-0.29	0.49	0.53	0.30	-0.32	-0.48	0.17	0.62	-0.71	0.24	-0.65	-0.49	0.21	-0.72	0.42	-0.53	-0.23
SB	0.27	-0.44	-0.48	-0.28	0.28	0.49	-0.16	-0.55	0.66	-0.31	0.63	0.55	-0.29	0.68	-0.39	0.51	0.20
TL	-0.26	0.42	0.30	0.21	-0.28	-0.41	0.69	0.52	-0.28	0.80	-0.25	-0.19	0.19	-0.22	0.63	-0.39	-0.59
CL	-0.36	0.64	0.53	0.47	-0.44	-0.34	0.71	0.50	-0.32	0.59	-0.39	-0.10	-0.04	-0.29	0.71	-0.31	-0.37
CL/VL	-0.41	0.78	0.71	0.66	-0.66	-0.14	0.58	0.29	-0.23	0.28	-0.47	0.02	-0.32	-0.27	0.58	-0.12	0.00

5.5 General Discussion

In this study, by using the same process with the same stimuli from a previous perception experiment (with Japanese listeners) but different subjects (Taiwanese listeners), we discuss some common features in the perception of expressive speech categories. Comparison of the results from the three experiments with Taiwanese listeners with the results of the three previous experiments with Japanese listeners suggests the following. The first point is that people who are from different cultures/native-language background show a certain similarity in the way they use semantic primitives. These results suggest that the two language speakers use the same semantic primitives as the primary ones. The findings discussed in the last section suggest the possibility of some type of universality of acoustic cues associated with semantic primitives.

However, the non-primary semantic primitives are different between the languages: Taiwanese associates *Neutral* with *dull*, *bright*, and *smooth* while Japanese associates it with *weak*, *calm* and *monotonous*; Taiwanese associates *Joy* with *low*, *calm*, and *smooth*, while Japanese associates it with *quiet*, *unstable* and *bright*, etc. This finding of a difference in the non-primary semantic primitives may account for why communication of emotions between people of different languages and cultures may go relatively smoothly for the most part yet may suddenly fall apart - a phenomena sometimes experienced in cross-linguistic communication situations. This finding may also be related to the difference in intonation between Japanese and Taiwanese: Japanese is a pitch-accent language, whereas Taiwanese is a tone language. Moreover, the speaking style is different between the two languages. For example, a sentence which may sound *Neutral* to Taiwanese speakers may sound like *Hot Anger* to Japanese, perhaps because Taiwanese speakers tend to use louder and more dynamic voices than do Japanese. More research is needed to explore this further. For example, we might use the same multi-layer approach and the same process described in this paper, but the stimuli would be Mandarin utterances, and the subjects, Japanese and Taiwanese.

5.6 Summary

To support the assumption of the existence of common characteristics in non-linguistic verbal information for expressive speech categories perception across different languages/cultures, the three-layer model is applied. In this application, we used

Japanese utterances and compared the resulting models of Japanese subjects, which was described in Chapter 3, with the model of Taiwanese subjects. To construct the three-layer model of Taiwanese subjects, the same three experiments were conducted again. The experimental results were used to build a fuzzy inference system, which revealed the relationship between semantic primitives and expressive speech of the model. Finally, 16 acoustic features were measured to reveal the relationship between semantic primitives and acoustic features.

The results show that certain common features exist among subjects from different linguistic backgrounds. The findings of a commonality also suggest that the proposed model is effective and appropriate for investigating common features that exist in speech communication, not only across different languages and cultures, but also within the same language.

Chapter 6. Summary and Future Work

A research work for understanding the role that non-linguistic information plays in the perception of expressive speech is described in this dissertation. In this research work, we successfully used a model to show the role that non-linguistic acoustic features and the role that humans' vagueness nature plays in the perception of expressive speech, where this model is built by perceptual experiments, evaluated by speech morphing, and applied to the analysis of non-linguistic verbal information.

Solid support of the two assumptions made for of expressive speech perception is found.

(1) The first assumption states that before listeners can decide to which expressive speech category a speech sound belongs, they will qualify a voice according to different descriptors, where each descriptor is an adjective for voice description.

(2) The second assumption states that people who are from different cultures/native-language background have some common characteristics for perception of expressive speech categories as well as some differences.

The two assumptions are firstly represented as a layer-structure model. This model illustrates that people perceive expressive speech categories (i.e. the topmost layer) not directly from a change of acoustic features (i.e. the bottommost layer), but rather from a composite of different types of "smaller" perceptions that are expressed by semantic primitives (i.e. the middle layer). A conceptual diagram is illustrated in Section 2.1. By this three-layer model, the focal points of the research work become the building and evaluation of these two relationships. After this, the model is applied to find the common features as well as differences of people who are or are not acquainted with the language of the voice they heard.

6.1 Contributions

Compared with other traditional statistical approaches, the proposed multi-layered approach gives three major advantages

- It takes the vague nature exhibited in human's perception into account and provides the mechanisms to deal with this vagueness.
- It helps to understand expressive speech perception from an aspect that is more

closely related to humans' behavior, i.e. the use of semantic primitives to judge the perception of expressive speech, which is then useful to improve a synthesizer to create a more natural sounding voice.

- It improves the application of this research topic. One example, as the application demonstrated in this dissertation, is to investigate common features in expressive speech perception among people with different culture/native-language backgrounds. [85].

6.2 The building of the model

Chapter 2 starts with a conceptual description about the multi-layered approach and a three-layered model for handling the human perception vagueness. Subsequently, it provides a brief overview of a two-phase approach for building and evaluating the perceptual model. In the building process, the relationship between expressive speech and semantic primitives is built, taking into account which adjectives should be selected as semantic primitives in the perceptual model; especially, the adjectives should be those suitable for describing expressive speech. This order is important because without finding suitable semantic primitives for expressive category judgments, it is not possible to conduct analysis of acoustic features related to semantic primitives. In the verification process, the relationship between semantic primitives and acoustic features is verified, taking into account that acoustic features may be the only features involved in perception of speech (but see AIM/correolograms, that purport to represent "perceived acoustic features") that can be quantified and modified, and that in turn, makes it possible to verify the built model by the method of rule-based speech morphing.

Chapter 3 describes the building of the perception model. The relationship between expressive speech categories and semantic primitive is built first and then is followed by building the relationship between semantic primitive and acoustic feature. The work was accomplished by three sub tasks.

Three experiments were carried out for examining voice data in terms of expressive speech categories and descriptors. Selection of semantic primitives was done in order to find those descriptors (e.g., adjectives) that can be used to describe the perception of expressive vocalizations. Those three experiments are (1) examination of listeners' perception of expressive speech, (2) construction of a psychological distance model and (3) selection of semantic primitives. The first experiment results in showing that subjects can perceive the intended emotion/affect of most utterances well. Moreover, the perceptual ratings of all utterances in the voice database contribute to building the relationship between expressive speech category and semantic primitives. The second experiment results in a psychological distance model which depicts the position of each category of expressive speech, where the distance between each of two categories represents the similarity among them. Such a psychological distance model conveying similarity information can be utilized when investigating appropriate semantic primitives. The main contribution gained from the third experiment is the 17 semantic primitives which are then used for representing the perceptual characteristics of expressive speech in the perceptual model.

To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech, a fuzzy inference system (FIS) is built for each category of expressive speech. Section 3.2 starts with an explanation about what fuzzy logic is and why it is applied in this study. In order to build an FIS with a highly reliable probability, another experiment, Experiment 4, is conducted for collecting more utterances as input data. Fuzzy inference systems were built by applying the model to the rating results of the experiments. An evaluation selected the five semantic primitives having a high correlation with each category of expressive speech. The generated relationships are a balance of commonly used adjectives with a precise numerical description, which seems to be well-matched with the way humans respond when they perceive expressive speech.

In order to support the relationship between semantic primitives and acoustic features, analysis of a large number of acoustic features was carried out. Section 3.3 first provides a discussion with a brief literature review regarding acoustic cues related to expressive speech. From the literature review, it can be understood that prosody, including voice quality, has an effect on the perception of expressive speech. The prosody related acoustic features are mainly extracted from F0 contour, power envelope, duration and spectrum. Therefore, the acoustic features analyzed in this study involve F0 contour, power envelope, duration, and spectrum. The correlation coefficients between semantic primitive and acoustic features were then calculated for depicting the relationship. The results suggest that the perception of expressive speech, at least in Japanese, may mainly be affected by the change of F0 contour and power envelope. In addition, spectral characteristics affect perception of some of the semantic primitives, for example, bright, high, unstable, and noisy. These are the more active adjectives. The change of spectrum, i.e., change in voice quality, may encourage perception of active semantic primitives.

The built relationship between expressive speech and semantic primitives and that between semantic primitives and acoustic features were combined to constitute a perceptual model for each category of expressive speech, which can be illustrated graphically in a figure. The results concluded that the perception of semantic primitives has a significant effect on the perception of expressive speech and that people do not perceive expressive speech directly from the change of acoustic features but rather from the perception of semantic primitives. Moreover, the built models also substantiate results from previous work of the importance of F0 and power envelope in perception of expressive speech.

6.3 The evaluation of the model

Chapter 4 verified the effectiveness of the built perceptual model by using the method of rule-based speech morphing and perceptual testing. Two types of information embedded in the relationship were verified. The first type is the significance of one element (acoustic feature or semantic primitive) in a layer to the perception of one element (semantic primitive or expressive speech category) in another layer. The second type is the impact direction and strength of one significant element. The work is accomplished by three sub tasks.

To simplify the manipulation of acoustic features in the morphing process, two types of rules for verifying the two types of information are created from the built relationships. Section 4.1 starts with the concept description that clarified what information in the built relationships should be transformed to the rules. The principles of rule development are then proposed. These principles clearly stated what should be used when developing rules. From the concept and the principles, base rules were created for evaluating the first type of information and intensity rules were created for evaluating the second type of information.

In order to modify a neutral utterance according to the morphing rules, a process was created for morphing expressive speech utterances (from the neutral utterance) by the created rules. The process of speech morphing used in this research first extracts F0 contour, power envelope, and spectrum from the neutral speech signal by using STRAIGHT while segmentation information was measured manually. Next, acoustic features in terms of F0 contour, power envelope, spectrum and duration were modified according to the morphing rule. Finally, the modified F0 contour, power envelope, spectrum and duration were re-synthesized by using STRAIGHT to produce a morphed utterance.

The perceptual experiments for examining the modified utterance in terms of the perception of expressive speech categories and semantic primitives (e.g., the methods, results, and discussions of the experiments), are described in Section 4.3. The significance of this work is that not only the intended perception is verified, but also it provides a more solid support to our assumption about how expressive speech is perceived; also, the intensity level of the intended perception can be described. The experiment results suggest that the rules created from the built relationships support the first assumption we made in this research work, i.e. that before listeners can decide to which expressive speech category a speech sound belongs, they will qualify a voice according to different descriptors, where each descriptor is an adjective for voice description. The results also showed that the obligatory principles are followed

- The rules are monotonous because subjects can identify the intended perception clearly.
- The rules are general because the intended perception of the morphed utterances can be clearly identified by different subjects.
- The rules are dynamic because the combination of SR-rules to ER-rules can create morphed utterances with intended categories of expressive speech.

6.4 The application of the model

Chapter 5 applies the three-layered model proposed in this research work to provide support for the second assumption, i.e., the existence of common characteristics in non-linguistic verbal information for expressive speech categories perception across different languages/cultures. By comparing the perceptual model of the different sets of application environments, in which Japanese and Taiwanese subjects listen to Japanese utterances, resultant perceptual models show a certain similarity in the way they use semantic primitives. It suggests that the two language speakers use essentially the same semantic primitives as the primary ones. This in turn suggests the possibility of some type of universality of acoustic cues associated with semantic primitives, which support the second assumption of this thesis. However, the non-primary semantic primitives may be different between the languages: Taiwanese associates Neutral with dull, bright, and smooth while Japanese associates it with weak, calm and monotonous; Taiwanese associates Joy with low, calm, and smooth, while Japanese associates it with quiet, unstable and bright, etc. This finding of a difference in the non-primary semantic primitives may account for why communication of emotions between people of different languages and cultures may go relatively smoothly for the most part yet may suddenly fall apart - a phenomena sometimes experienced in cross-linguistic communication situations.

6.5 Future work

Future research involves applications for synthesizing good expressive speech. The advantage of the perceptual model built by the multi-layered approach proposed in this research is that it provides information about how acoustic features have an effect on the perception of voice characteristics, i.e., semantic primitives, and how the perception of voice characteristics have an effect on listener’s judgments about expressive speech. The information about semantic primitives may approximate human perception, and may be especially useful for dealing with human perception vagueness. While the performance of current synthesizers of expressive speech are still limited by the statistical relationship between acoustic features and expressive speech categories, information that more closely approximates the object, i.e. human’s perception, can provide greater possibilities for solving the problem of how to create a more natural sounding synthesizer for expressive speech.

Regarding the aspect of the second assumption about common features across languages and cultures in perception of expressive speech, this research work has provided some interesting findings. Specifically, this research suggests that the primary semantic primitives may be common across speakers of different languages/ backgrounds, but non-primary semantic primitives may be different. In the case of Japanese and Taiwanese, these differences may be related to the difference in intonation between Japanese and Taiwanese: Japanese is a pitch-accent language, whereas Taiwanese is a tone language. Moreover, the speaking style is different between the two languages. Clearly more research is needed to explore this further. One approach would be to use the same multi-layer approach and the same process described in this paper, but use Mandarin stimuli with Japanese, as well as Taiwanese listeners.

Rather than a three-layer model, we also need to ask if it is possible that more than one layer exists between acoustic features and the perception of expressive vocalizations. A multi-level approach for expressive speech perception as we are proposing should consider the possibility of more levels that are in or between the existing levels in the currently proposed model. One possible level would be one that considers a much more fine-grained taxonomy system of expressive speech categories as described by [15] and [27]. The resulting model could become a four-layer model in order to consider the additional relationships between semantic primitives and a fine-grained category classification, as well as between a fine-grained category classification and basic expressive speech categories. We consider the construction of such a four-layer model future work.

6.6 Conclusion

In this research work, two assumptions of expressive speech perception are made from the observation of our daily life. The results of the building and the evaluation of the model provide solid support for the first assumption. The results of the model application provide solid support for the second assumption. It is hoped these results will help to develop better tools for expressive speech synthesis and recognition, as well as to advance our understanding of human perception of expressive speech. It is also hoped that the research work will be a stepping stone for future work in the ongoing exploration of expressive speech.

References

[1] Alter, K, Erhard R., Sonja K., Erdmut P., Mireille B., Angela F., and Johannes M., “On the relations of semantic and acoustic properties of emotions”, Proceedings of the 14th International Congress of Phonetic Science, San Francisco, 1999, pp.2121-2124.

[2] Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E. Cox, C., “AST for emotional speech: Clarifying the issues and enhancing performance,” Neural Networks Volume 18, Issue 4, May 2005, pp. 437-444.

[3] Austermann, A. Esau, N. Kleinjohann, L. Kleinjohann, B. Paderborn Univ., Germany. “Fuzzy emotion recognition in natural speech dialogue,” IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005.

[4] Banse, R. and Scherer, K. R., “Acoustic profiles in vocal emotion expression,” Journal of Personality and Social Psychology, 1996, pp.614–636.

[5] Banziger, T., Scherer, K.R., “The role of intonation in emotional expressions,” Speech Communication 46, 2005, pp.252-267.

[6] Batliner, A., Fischer, K., Huber, R., Spilker, J. and Nöth, E., 2003. How to find trouble in communication, Speech Communication 40, Issues 1-2, April 2003, pp. 117-143.

[7] Beier, E. G., and Zautra, A.J., "Identification of vocal communication of emotions across cultures", J. Consult. Clin. Psychol 39, 1972, pp. 166.

[8] Terri, L. B., Jeri, L. T., Daniel, W. L., "Leger, Gender stereotypes in the expression and perception of vocal affect," Sex Roles 34, 1996, pp429-445.

[9] Cahn, J. E., Generating expression in synthesized speech, master’s thesis, MIT, 1990, Media Laboratory.

[10] Chiu, S., “Fuzzy model identification based on cluster estimation,” Journal of Intelligent and Fuzzy Systems 2, 1994, pp. 267-278.

[11] Lee, C. M., and Narayanan, S., “Emotion Recognition Using a Data-Driven Fuzzy Inference System,” Proc. Eurospeech, 2003.

[12] Chung, S.-J. : " L'expression et la perception de l'emotion extradite de la parole spontanee : evidence du coreen et de l'anglais. (Expression and perception of emotion extracted from the spontaneous speech in Korean and in English)," Doctoral Dissertation, ILPGA, Sorbonne Nouvelle University, 2000, Septentrion Press.

[13] Albesano, D., Gemello, R., Mana, F., "Hybrid HMM-NN for speech recognition and prior class probabilities", Proceedings of the 9th International Conference on Volume 5, Issue , 18-22 Nov. 2002 pp. 2391 – 2395.

[14] Darke, G., "Assessment of Timbre Using Verbal Attributes," Proceedings of the Conference on Interdisciplinary Musicology, Montreal, 200.

[15] Devillers, L., Vidrascu, L., Lamel, L., "Challenges in real-life emotion annotation and machine learning based detection," Neural Networks 18, Issue 4, May 2005, pp. 407-422

[16] Douglas-Cowie, E. Campbell, N, Cowie, R, Roach, P., 2003. "Expressive speech: towards a new generation of databases," Speech Communication 40, 2003, pp. 33-60.

[17] Dromey, C., Silveira J., and Sandor, P., "Recognition of affective prosody by speakers of English as a first or foreign language," Speech Communication 47, Issues 3, 2005, pp.351-359.

[18] Ehrette T., Chateau N., D'Alessandro C., Maffiolo V. "Prosodic parameters of perceived emotions in vocal server voices," 1st International Conference on Speech Prosody. Aix-en-Provence, France, April 11-13, 2002.

[19] Erickson, D., "Expressive speech: Production, perception and application to speech synthesis," Acoust. Sci. & Tech, 26, 2005, pp.317-325..

[20] Erickson, D., and Maekawa, K., "Perception of American English emotion by Japanese listeners," Proc. Spring Meet. Acoust. Soc. Jpn., 2001, pp. 333-334.

[21] Erickson, D., Ohashi, S., Makita, Y., Kajimoto, N., Mokhtari, P., "Perception of naturally-spoken expressive speech by American English and Japanese listeners", Proc. 1st JST/CREST Int. Workshop Expressive Speech Processing, Kobe, Feb. 21-22, 2003, pp. 31-36.

[22] Fauconnier, G., "Mappings in Thought and Language," Cambridge University

Press.1997.

[23] Friberg, A., “A fuzzy analyzer of emotional expression in music performance and body motion,” Proc. Music and Music Science, Stockholm, 2004.

[24] Friberg, A., Bresin, R. and Sundberg, J., “Overview of the KTH rule system for music performance,” Advances in Cognitive Psychology 2006, vol. 2, no. 2-3, pp.145-161.

[25] Fujisaki, H., “Manifestation of Linguistic, Para-linguistic, non-linguistic Information in the Prosodic Characteristics of Speech”, IEICE, 1994.

[26] Gobl, C. and Ni Chasaide, A. “Testing affective correlates of voice quality through analysis and resynthesis,” Proceedings of the ISCA Workshop on Speech and Emotion, pates Northern Ireland, 2000, pp.178-183.

[27] Grimm, M., Mower, E., Kroschel, K., and Narayanan, S. “Primitives based estimation and evaluation of emotions in speech,” Speech Communication 49, 2007, pp 787-800.

[28] Hanson, H., Glottal characteristics of female speakers: acoustic correlates, J. Acoust. Soc. Am. 101, 1997, pp. 466–481.

[29] Hashizawa, T., Takeda, S., Hamzah, M. D., Ohyama, G., “On the differences in prosodic features of emotional expressions in Japanese speech according to the degree of emotion,” Proc. Speech Prosody, 2004, pp. 655–658.

[30] Hayashi, Y., “Recognition of vocal expression of mental attitudes in Japanese : Using the interjection "eh",” Proc. Int. Congr. Phonetic Sciences, San Francisco, 1999, pp. 2355-2358.

[31] Howard, D., and Angus, J., Acoustics and Psychoacoustics, 2nd ed, Focal Press, 2006.

[32] Huttar, G. L. “Relations between prosodic variables and emotions in normal American English utterances,” Journal of Speech and Hearing Research 11, n3, Sep. 1968, pp481-487.

[33] Ishii C.T., and Campbell N., "Acoustic-prosodic analysis of phrase finals in Expressive Speech," JST/CREST Workshop 2003, pp85-88,

[34] Ishii, C.T., and Campbell, N.. “Analysis of Acoustic-Prosodic Features of

Spontaneous Expressive Speech," Proceedings of 1st International Congress of Phonetics and Phonology 19. 2002.

[35] Jang, J.-S. R., Sun, C.-T., Mizutani, E., Neuro-Fuzzy and Soft Computing. Prentice Hall, 1996.

[36] Juslin, P. N, "Communication of emotion in music performance: A review and a theoretical framework," In P. N. Juslin & J. A. Sloboda (eds.), Music and emotion: Theory and research (pp. 309 – 337). New York: Oxford University Press.

[37] Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., "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 27, 1999, pp.187-207.

[38] Keating, P. and Esposito, C. "Linguistic Voice Quality," invited keynote paper presented at 11th Australasian International Conference on Speech Science and Technology in Auckland, Dec. 2006,

[39] Kecman, V., Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. 2001, MIT Press.

[40] Kendon, A. (ed.), Nonverbal Communication, Interaction and Gesture: Selections from Semiotica, (Approaches to Semiotics). Mouton De Gruyter, August 1981.

[41] Kent, R. D., and Read, C., Acoustic analysis of speech, 2nd ed., Singular, 2001.

[42] Kienast, M., Sendlmeier, W. F., "Acoustical analysis of spectral and temporal changes in expressive speech", ISCA workshop on speech and emotion, Belfast, 2001.

[43] Krom, de G., "Some Spectral Correlates of Pathological Breathy and Rough Voice Quality for Different Types of Vowel Fragments," Journal of Speech and Hearing Research 38, August 1995, pp.794-811.

[44] Leinonen, L., "Expression of emotional-motivational connotations with a one-word utterance," J. Acoust. Soc. Am., 102, 1997. pp.1853-1863.

[45] Maekawa, K., "Phonetic and phonological characteristics of paralinguistic information in spoken Japanese," Proc. Int. Conf. Spoken Language Processing, 1998, pp.635-638.

[46] Maekawa, K., “Production and perception of ‘paralinguistic’ information,” Proceedings of Speech Prosody, Nara, pp.367-374.

[47] Maekawa, K., Kitagawa, N., “How does speech transmit paralinguistic information?” Congnitive Studies 9, 2002, pp. 46-66.

[48] Manning, P. K., Symbolic Communication: Signifying Calls and the Police Response, 2002. The MIT Press

[49] Mehrabian, A., Nonverbal communication, Aldine-Atherton, Chicago, Illinois.

[50] Menezes, C., Maekawa, K., and Kawahara, H., "Perception of voice quality in pralinguistic information types: A preliminary study," Proceedings of the 20th General Meeting of the PSJ, 2006, pp.153-158.

[51] Mozziconacci, Sylvie, Speech variability and emotion: production and perception. Eindhoven: Technische Universiteit Eindhoven, 1998.

[52] Murray, I. R., Arnott, J. L., “Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion,” Journal of the Acoustical Society of America 93 (2), 1993, pp.1097-1108.

[53] Murray, I.R., Arnott, J. L., “Implementation and testing of a system for producing emotion-by-rule in synthetic speech,” Speech Communication 16, 1995, pp. 369-390.

[54] Nguyen, P. C., Ochi, T., Akagi, M., “Modified restricted temporal decomposition and its application to low rate speech coding”, IEICE Trans. Inf. & Syst., E86-D (3), 2003, pp.397-405.

[55] Ní Chasaide, A., Gobl, C., Voice source variation. In: Hardcastle, W.J., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Oxford, 1997. , pp. 427-461.

[56] Ofuka, E., Valbret, H., Waterman, M., Campbell, N., and Roach, P., “The role of F0 and duration in signaling affect in Japanese: Anger, kindness and politeness,” Proceedings of the Third International Conference on Spoken Language Processing, Yokohama, 1994.

[57] Öster, A.-M., and Risberg, A., The identification of the mood of a speaker by hearing impaired listeners” In the Speech Transmission Laboratory, Quarterly Progress and Status Report 4, Royal Institute of Technology, Stockholm, 1986, 1986,

pp. 79-90.

[58] Paeschke, A.,"Global Trend of Fundamental Frequency in Emotional Speech", Proc. Speech Prosody 2004 Nara, 2004, pp. 671–674.

[59] Patwardhan, P. P., and Raoa, P., “Effect of voice quality on frequency-warped modeling of vowel spectra,” Speech Communication 48, Issue 8, August 2006, pp. 1009-1023.

[60] Pell, M. D. “Influence of emotion and focus location on prosody in matched statements and questions, “J. Acoust. Soc. Am., 109, 2001, pp.1668-1680.

[61] Pittam, J., Voice in social interaction. Sage Publications, Thousand Oaks, CA, 1994.

[62] Robbins, S. and Langton, N., Organizational Behaviour: Concepts, Controversies, Applications (2nd Canadian ed.). Upper Saddle River, NJ: Prentice-Hall, 2001.

[63] Sakuraba, K., “Emotional expression in Pikachuu”, J. Phonet. Soc. Jpn., 8, 2004, pp.77-84.

[64] Scherer, K. R., "Vocal communication of emotion: A review of research paradigms," Speech Communication 40, 2003, pp. 227-256.

[65] Scherer, K. R., Banse, R., Wallbott, H. G., Goldbeck, T., “Vocal cues in emotion encoding and decoding,“ Motivation and Emotion 15, 1991, pp. 123-148.

[66] Scherer, K. R., Ladd, D.R., Silverman, K. A., “Vocal cues to speaker affect: Testing two models”, Journal of the Acoustical Society of America 76, 1994, pp. 1346-1356.

[67] Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S., 2001. “Acoustic correlates of emotion dimensions in view of speech synthesis,” Proc. Eurospeech 2001, Denmark, pp. 87-90.

[68] Schröder, M., “Speech and Emotion Research - An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis,” (Ph.D thesis). Vol. 7 of Phonus, Research Report of the Institute of Phonetics, Saarland University.

[69] Shochi, T., Auberg, V., and Rilliard, A., “How prosodic attitudes can be false friends: Japanese vs. French social affects,” Proc. Speech Prosody 2006, Dresden,

Germany, 2006.

[70] Robinson, P. and Shikler T. S.,”Visualizing dynamic features of expressions in speech,” InterSpeech, Korea, 2004.

[71] Kecman, V., Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models, the MIT Press, 2001.

[72] Sugeno, M., Industrial Applications of Fuzzy Control. Elsevier Science Inc., New York. 1985.

[73] Takeda, S., Ohyama, G., Tochitani, A., Nishizawa, Y., “Analysis of prosodic features of “anger” expressions in Japanese speech”. J. Acoust. Soc. Jpn. 58(9), 2002, pp. 561-568. (in Japanese).

[74] Tolkmitt, F. J. and Scherer, K. R., “Effect of experimentally induced stress on vocal parameters”, Journal of Exp Psychol Hum Percept Perform 12(3), 1986 Aug, 302-13.

[75] Traube, C., Depalle, P., Wanderley, M., “Indirect acquisition of instrumental gesture based on signal, physical and perceptual information,” Proceedings of the 2003 conference on new interfaces for musical expression.

[76] Ueda, K., “Should we assume a hierarchical structure for adjectives describing timbre?” Acoustical Science and Technology 44(2), 1988, pp102-107 (in Japanese).

[77] Ueda, K., “A hierarchical structure for adjectives describing timbre,” Journal of the Acoustical Society of America. 100(4), 275.

[78] Ueda, K., and Akagi, M., “Sharpness and amplitude envelopes of broadband noise,” Journal of the Acoustical Society of America, vol. 87, no. 2, 1990, pp. 814-819.

[79] Van Bezooijen, R., “The Characteristics and Recognizability of Vocal Expression of Emotions, The Letherlands, Foris, 1984.

[80] Vickhoff, B., and Malmgren, H., “Why Does Music Move Us?” Philosophical Communication, Web Series, No. 34. 2004.

[81] Williams, C. E., and Stevens, K. N., “On determining the emotional state of pilots during flight: An exploratory study.” Aerospace Medicine 40(12), 1969. pp. 1369-1372.

[82] Williams, C. E., and Stevens, K. N., “Emotions and speech: some acoustical correlates,” Journal of the Acoustical Society of America 52, 1972, pp. 1238-1250.

[83] Wolkenhauer, O., Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis, Wiley, 2001.

[84] Nguyen B. H. and Akagi, M., “A Flexible Spectral Modification Method based on Temporal Decomposition and Gaussian Mixture Model,” Proc. InterSpeech, 2007.

[85] Huang, C. F. Erickson, D. Akagi, M., "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners," Acoustics, 2008.

[86] Carter R, “Mapping the mind”, University of California Press, 2000.

[87] Hawkins J. and Blakesle S., “On Intelligence”, Holt Paperbacks, 2005.

[88] Narendranath M., Murthy H. A., Rajendran S. and Yegnanarayana B., "Transformation of formants for voice conversion using artificial neural networks," Speech Communication Volume 16, Issue 2, February 1995.

[89] Kain E. and Macon M. W., "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," Proc. of the ICASSP'01.

[90] Kain A. B., "High Resolution Voice Transformation," PhD thesis.

[91] Ye H. and Young S. “High quality voice morphing,” Proc. ICASSP 2004.

[92] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., Voice Conversion Through Vector Quantization, Proceedings of the IEEE ICASSP 1988, pp. 565-568.

[93] Kain A., Macon M., "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction". Proceedings of ICASSP, May 2001.

[94] Rentzos D., Vaseghi S., Yan Q., Ho C.H., Turajlic E. "Probability Models of Formant Parameters for Voice Conversion", in Proc. Eurospeech 2003, pp. 2405-2408.

[95] Pfitzinger H. R., “Unsupervised Speech Morphing between Utterances of any Speakers,” Proc. of the 10th Australian Int. Conf. on Speech Science and Technology (SST 2004), pp. 545-550.

[96] Jacques Toen, Elan Tts, "Generation of Emotions by a Morphing Technique in English, French and Spanish," In Proc. Speech Prosody 2002.

[97] Tomoko Yonezawa, Noriko Suzuki, Kenji Mase, Kiyoshi Kogure, "Gradually Changing Expression of Singing Voice based on Morphing," In Proc. Interspeech 2005

[98] Tomoko Yonezawa, Noriko Suzuki, Shinji Abe, Kenji Mase, and Kiyoshi Kogure, "Perceptual Continuity and Naturalness of Expressive Strength in Singing Voices Based on Speech Morphing", EURASIP Journal on Audio, Speech and Music Processing, vol. 2007,

Publication List

Journal

[1] Huang, C. F. and Akagi, M., “A three-layered model for expressive speech perception”. Speech Communication (To be appeared).

International conferences

[2] Huang, C. F. Erickson, D. and Akagi, M., "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners," Acoustics, 2008.

[3] Huang, C. F. and Akagi, M., “A rule-based speech morphing for verifying an expressive speech perception model,” Proc. Interspeech2007, pp.2661-2664, 2007.

[4] Huang, C-H. and Akagi, M., “The building and verification of a three-layered model for expressive speech perception,” Proc. JCA2007, CD-ROM, 2007.

[5] Chun-Fang Huang and Masato Akagi. Toward a rule-based synthesis of emotional speech on linguistic descriptions of perception. Proc. ACII 2005.

[6] Chun-Fang Huang and Masato Akagi. A multi-layer fuzzy logical model for emotional speech perception. Proc. EuroSpeech 2005.

Domestic conferences

[7] Huang C. F., Erickson, D., and Akagi M., “A study on expressive speech and perception of semantic primitives: Comparison between Taiwanese and Japanese,” 電子情報通信学会技術報告, SP2007-32, 2007.

[8] Huang, C. F., Erickson, D., and Akagi, M., “Perception of Japanese expressive speech: Comparison between Japanese and Taiwanese listeners,” Proc. ASJ '2007 Fall Meeting, 1-4-6, 2007.

[9] Huang C. F., Nguyen B. P., and Akagi M., “Rule-Based Speech Morphing for Evaluating Emotional Speech Perception Model,” Proc. ASJ '2007 Spring Meeting, 3-8-9, 2007.

[10] C. F. Huang and M. Akagi, "Rule-Based Speech Morphing for Evaluating Linguistic Descriptions of Emotional Speech Perception," ASJ '2005 Fall Meeting, 1-6-3, 2005.

[11] Huang, C-F., Akagi, M., "Rule-Based Speech Morphing for Verification of Emotional Perception Model," IEICE Technical Report, Jul 2005.

[12] Chun-Fang Huang and Masato Akagi. A multi-layer fuzzy logical model for emotional speech perception. Trans. Tech. Comm. Psychol. Physiol. Acoust., The Acoustical Society of Japan, Vol. 34, No.8, H-2004-95, pp.547-552, Kanazawa, Oct. 2004.

[13] Chun-Fang Huang and Masato Akagi. A perceptual model of emotional speech build by fuzzy logic. Proc. ASJ'2004 Fall Metting, pp.287-289, Okinawa, Sep. 2004.