JAIST Repository
https://dspace.jaist.ac.jp/

Title	A three-layered model for expressive speech perception
Author(s)	Huang, Chun-Fang; Akagi, Masato
Citation	Speech Communication, 50(10): 810-828
Issue Date	2008-10
Type	Journal Article
Text version	author
URL	http://hdl.handle.net/10119/4904
Rights	NOTICE: This is the author's version of a work accepted for publication by Elsevier. Chun-Fang Huang and Masato Akagi, Speech Communication, 50(10), 2008, 810-828, http://dx.doi.org/10.1016/j.specom.2008.05.017
Description	

Japan Advanced Institute of Science and Technology

A three-layered model for expressive speech perception

Chun-Fang Huang and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), 1-1 Asahidai, Nomi, Ishikawa 923-1211, Japan

Abstract

This paper proposes a multi-layer approach to modeling perception of expressive speech. Many earlier studies of expressive speech focused on statistical correlations between expressive speech and acoustic features without taking into account the fact that human perception is vague rather than precise. This paper introduces a three-layer model: five categories of expressive speech constitute the top layer, semantic primitives constitute the middle layer, and acoustic features, the bottom layer.

Three experiments followed by multidimensional scaling analysis revealed suitable semantic primitives. Then, fuzzy inference systems were built to map the vagueness of the relationship between expressive speech and the semantic primitives. Acoustic features in terms of F0 contour, time duration, power envelope, and spectrum were analyzed. Regression analysis revealed correlation between the semantic primitives and the acoustic features. Parameterized rules based on the analysis results were created to morph neutral utterances to those perceived as having different semantic primitives and expressive speech categories. Experiments to verify the relationships of the model showed significant relationships between expressive speech, semantic primitives, and acoustic features.

Keywords:

expressive speech, perception, multi-layer model, fuzzy inference system, acoustic analysis, rule-based

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 receivers. The type of information shared between senders and receivers can be verbal, non-verbal, and/or symbolic (Fauconnier 1997, Mehrabian 1972, Manning 1989), and may not necessarily be all that apparent. For example, music not only gives listeners beautiful sounds to hear, but also communicates the emotion of the composers and performers (Vickhoff 2004). Gestures, another example, communicate both information and emotion (Kendon 1981). With regard to communication using speech, receivers not only decode linguistic information, but also paralinguistic information (Fujisaki 1994). According to one definition, paralinguistic information refers to the acoustic information that is conveyed in speech, such as voice pitch, amplitude, rate, and voice quality, rather than the sheer lexical meaning of the word (Robbins and Langton 2001). Perception of paralinguistic information is the focus of much research, especially the topic of the perception by receivers of a sender's emotional state.

Previous work on expressive speech [1] focuses on two major research areas: recognition and synthesis of expression in speech. Expressive speech recognition is a method for automatically identifying the emotional state of speakers. Recently, researchers recognize that the types of voice database used may result in different detection systems (Batliner et al. 2003; Devillers et al., 2005). Expressive speech synthesis involves manipulating parameters of the acoustic features of the voice data to produce perception of different emotions by receivers (Scherer 2003). Both recognition and synthesis work require analysis of perception of expressive speech. Work heretofore has been concentrated on measuring acoustic features in the speech signal and then using statistical methodology to select the most significant features for the classification or identification of emotions (e.g.,Van Bezooijen, 1984; Tolkmitt and Scherer, 1986, Williams and Stevens, 1969, 1972; Scherer, 1984, 1991; Kienast and Sendlmeier, 2000; Cahn, 1990; Murray and Arnott, 1993, 1995; Schroder et al., 2001; Banziger and Scherer, 2005). Based on the results of the analysis, this field has progressed markedly, even including applications to real-life situations (Douglas-Cowie 2003).

---

[1] Emotional and expressive speech both have been long considered as two interchangeable words. However, in this paper we regard them as different. We only use *emotion* when referring to the mental state of people. Otherwise, the term "expressive speech" is used.

However, results may be limited by this method due to the fact that reproduction of expressive speech either for purposes of recognition or synthesis has not been effectively achieved, even though it is known that certain categories of expressive speech may be characterized by specific acoustic features (Sobol and Robinson, 2004). Clearly a model to approach the study of expressive speech is needed (Scherer 2003).

In this paper, a multi-layered model is proposed to approach the study of expressive speech, which is based on an assumption; namely, 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. 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.

This idea can be understood better by the following example. 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. In this study, we refer to such acoustic characteristics of the voice, e.g., pitch, duration, loudness, etc., as "acoustic features", and such listeners' labelings of the sounds, e.g., "bright-sounding" or "fast-sounding", as "semantic primitives". Given this assumption, we propose a three-layered model that is different from the traditional two-layered models, which only considered the direct correlation between expressive speech and acoustic features. Our perceptual model adds a new layer between expressive speech and acoustic features which consists of semantic primitives.

Using qualitative descriptors to describe expressive utterances is also found with other types of communication. In music, the use of adjectives for describing sound (timbre) has been conducted by many researchers. For example, qualitative descriptions of loud, fast, and staccato music performance give listeners a perception of anger (Juslin 2001). The perception of these three qualitative descriptions is actually related to acoustic features of tempo, sound level, and articulation. Ueda (1996) discussed the possibility of a hierarchical structure of qualitative adjectives for timbre. The psychological perception of sound can be characterized by acoustic features. Ueda and Akagi (1990) 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 (2005) asked musicians to use 12 verbal adjectives for assessing a number of music sounds, and found there was an agreement among the adjectives listeners used to describe their perception of timbre. Traube, Depalle, and Wanderley (2003) 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 (2004) 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 (Friberg., Bresin, and Sundberg, 2006). 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.

A large body of research supports the validity in our proposed model of the relationships between (1) emotional perception and semantic primitives and also that between (2) semantic primitives and acoustic features. The construction of such a model should provide concrete support for the proposed ideas. There are two approaches for achieving this goal. One is to construct a general-purpose model which can explain the perception of expressive speech categories, across a number of speakers, based on examination of voices of multiple actors. The other is to construct a specific-purpose model based on the voice of a single actor. The purpose of this model would be to validate the method proposed here, to demonstrate that expressive voices can be simulated and controlled by using a multi-layered model. This specific-purpose model could then be extended to include a general-purpose model based on multiple speakers to have wider applicability in explaining perception of expressive speech

We propose a two-step approach to construct this specific-purpose model. The details of these two steps are:

1. The first step involves a top-down approach to build the above two relationships by analyzing the connections between expressive speech perception and semantic descriptors about voices, and between semantic descriptors about voices and acoustic features.
2. The second step involves a bottom-up method in order to verify what was built by step 1 described above.

In the first step, 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. Also, in order to support the relationship between semantic primitives and acoustic features, analysis of a large number of acoustic features will be necessary. To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech a fuzzy inference system will be built.

After the first step, the second step will verify that the two relationships built in the first step can simulate and control the expressive voices production. In the verification, we will use the analysis results of the first step to simulate and control the expressive voices by using the neutral utterances from the same actress. We use the semantic primitive specifications with corresponding acoustic features of the first step to resynthesize 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.

Rules are created with parameters that morph the acoustic characteristics of a neutral utterance to the perception of certain semantic primitives or expressive speech categories. These parameters come from the fuzzy inference system and the correlation of these two relationships. Speech morphing is a technique described by Sherer (2003), and technically developed by Kawahara et al. (1999). It is also called copy synthesis, resynthesis or speech-morphing, and essentially refers to a method that 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 speech-morphing and base-rule development. The relationships revealed in the building process then can be verified by morphing a neutral utterance to other utterances which results in different types and intensities of perception.

This paper is organized as follows. In Section 2 the proposed model is introduced. In Sections 3, the details of how the model is built are described, specifically, the construction of the two types of relationships-- the relationship between expressive speech and semantic primitives and the relationship between semantic primitive and acoustic feature. In Section 4 the resulting model is presented. In Section 5, the methods and results of verification are presented, and in Section 6, discussion and suggestions for future work are put forth.

2. Overview of the model

Fig. 1 shows a conceptual diagram of the proposed perceptual model. It consists of three layers: expressive speech, semantic primitives, and acoustic features. Five categories of expressive speech are studied, Neutral (N), Joy (J), Cold Anger (CA), Sadness (S), and Hot Anger (HA). 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.

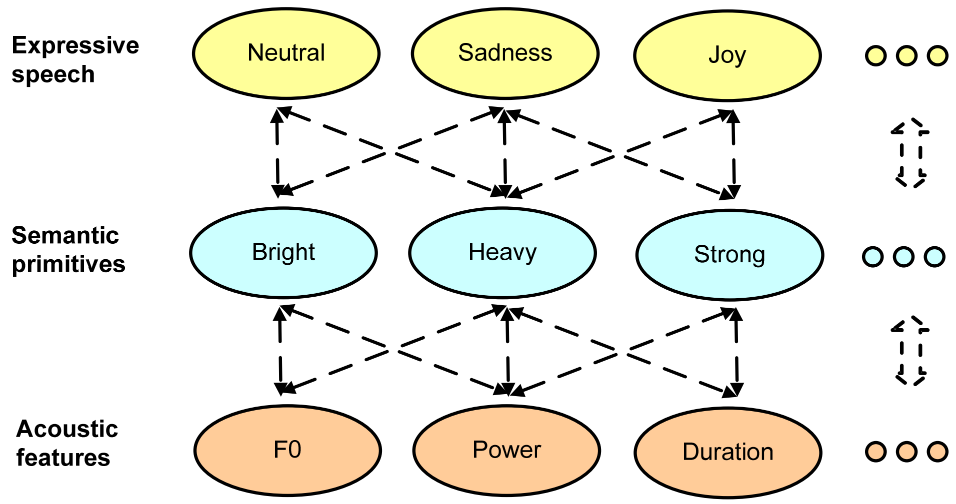


Fig. 1. Conceptual diagram of the perceptual model.

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. The hypothesis also implies that the model describes the process of the perception of expressive speech in a way which is close to a human's behavior. Humans' perceptions always have a quality of vagueness; for example, humans use vague linguistic forms which do not have precise values, such as "slow" or "slightly slow". Therefore, a computable methodology is also proposed for precisely analyzing this natural vagueness.

3. Development of three layered model

We build two relationships in the model: (1) the relationship between expressive speech and semantic primitives and (2) the relationship between semantic primitives and acoustic features.

3.1 The relationship between expressive speech and semantic primitives

In order to build the relationship between expressive speech and semantic primitives, three experiments were conducted, after which a fuzzy inference system was applied. The first experiment examined all 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. The second experiment constructed a perceptual space of categories of expressive speech by using a multidimensional scaling (MDS) technique. (See also Maekawa and Kitagawa (2002) for a similar approach.) Examination of the perceptual space assisted in selecting suitable adjectives as semantic primitives, which were then fine- tuned in the third experiment to arrive at a set of semantic primitives used in the model. The results of the three experiments were then used in training and checking data for the fuzzy inference system that maps the relationship between expressive speech and semantic primitives.

3.1.1 Three psychoacoustical experiments

3.1.1.1 Experiment 1

Experiment 1 was conducted to examine listeners' perception of expressive speech utterances.

- Method

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. 12 graduate students, native male Japanese speakers, average age 25 years old, with normal hearing participated in the experiment. 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 1. 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.

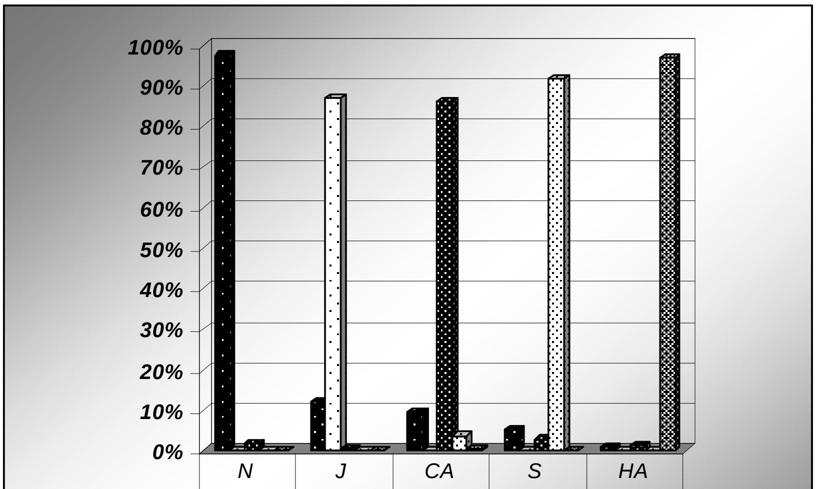
Table 1. Evaluaton form used in Experiment 1.

Sentence 1 Atarasiime-rugatodoiteimasu (A new e-mail has arrived)				
N	J	C A	S	H A

- Results and discussion

Table 2 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 was Joy (J).

Table 2. Percentage of ratings of the 5 intended categories. The columns are the intended categories and the rows, the categories rated by subjects.



	N	J	CA	S	HA
N	98%	12%	10%	5%	1%
J	0%	87%	0%	0%	0%
CA	2%	1%	86%	3%	2%
S	0%	0%	4%	92%	0%
HA	0%	0%	0%	0%	97%

3.1.1.2 Experiment 2

Experiment 2 was conducted to construct a perceptual space of utterances in the different expressive speech categories, and the results were analyzed using MDS. The resulting perceptual space was then used in Experiment 3 to select suitable semantic primitives for the perceptual model.

- Method

In Experiment 2, 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. The subjects were the same as those in Experiment 1. Scheffe's method of paired comparison was used. Subjects were asked to rate each of the 15 x 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 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

Fig. 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. 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.

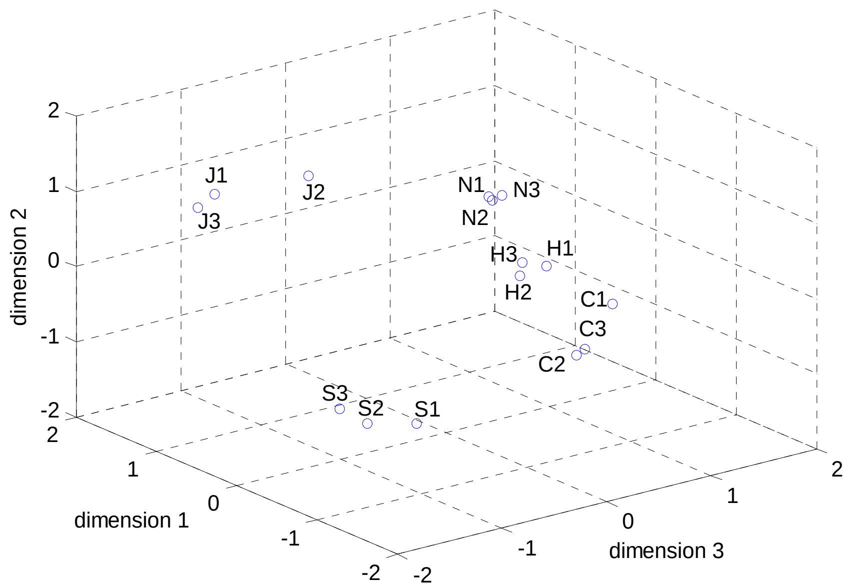


Fig. 2. The resulting perceptual space of utterances in different categories of expressive speech. One circle represents one utterance.

3.1.1.3 Experiment 3

Experiment 3 was conducted to determine suitable semantic primitives for the perceptual model. 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 the following pre-experiment.

- Pre-experiment

  Sixty adjectives were selected as candidates for semantic primitives. 46 of these were from the work by Ueda (1988) (with English glosses), in which he asked 166 listeners to choose adjectives which are often used to describe voice timbre. 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 form 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 were the ones most frequently circled, and were thus chosen for Experiment 3.

Table 3. 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

ID	Adjective (Japanese)	Adjective (English)	Correlation Coefficient
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

- Method of Experiment 3

The stimuli and subjects were the same as in Experiment 2. 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 when they heard each utterance, and to indicate how appropriate the adjective was for describing the utterance they heard (0: very appropriate, 3: not very appropriate). 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. Since the scaling we used in the evaluation of Experiment 3 has two extreme ends (not very appropriate and very appropriate), therefore for each semantic primitive, there exists a situation that people hold a neutral stance with regard to an utterance. Therefore, Equation (1) is the regression equation we used:

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

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.

- Results and discussion

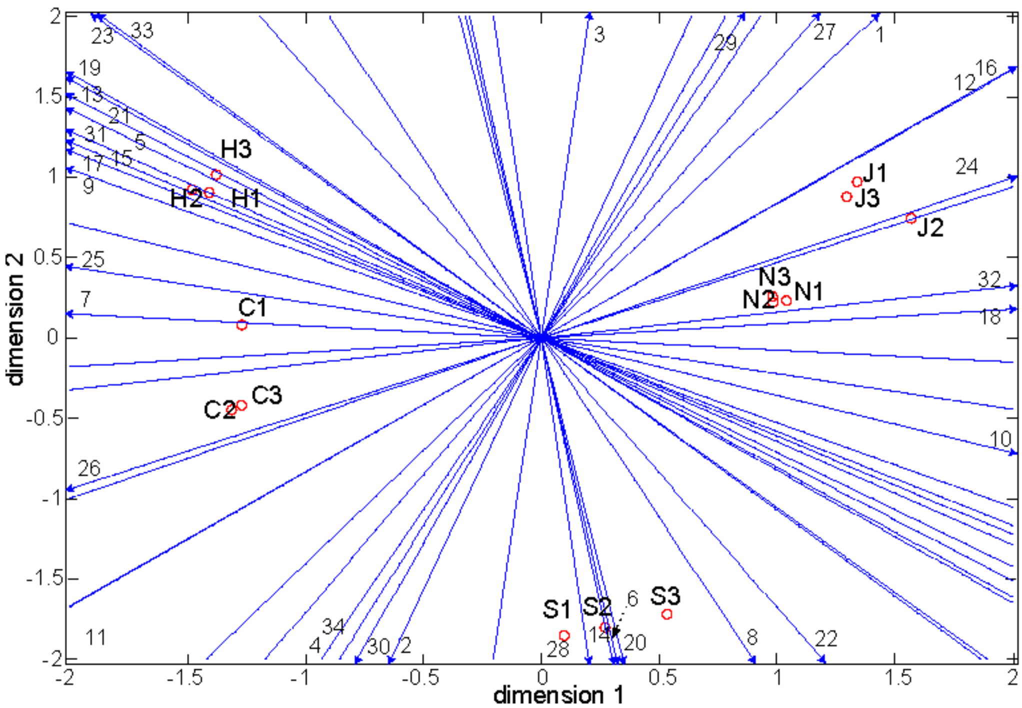
Fig. 3 is the diagram that presents the 34 adjectives plotted in dimension-1 against dimension-2, dimension-1 against dimension-3, and dimension-3 against dimension-2 of the 3D perceptual space. The utterances are represented by the same IDs as in Figure 2 and a line in the plot indicates an adjective 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.

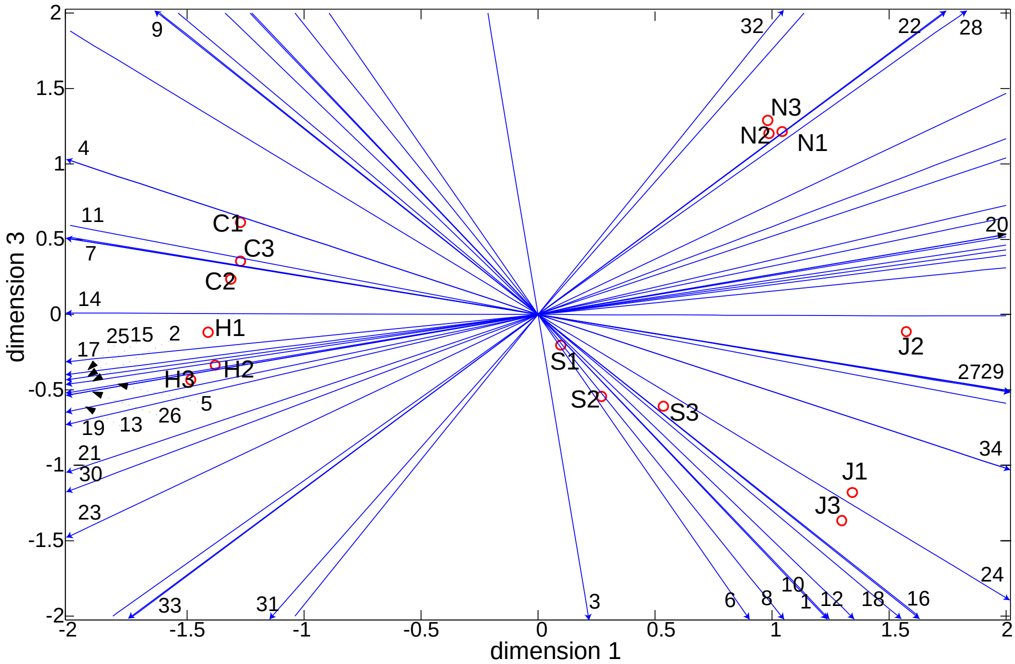
(2) The angle between each pair of adjectives: The smaller the angle is, the more similar the two adjectives.

(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.

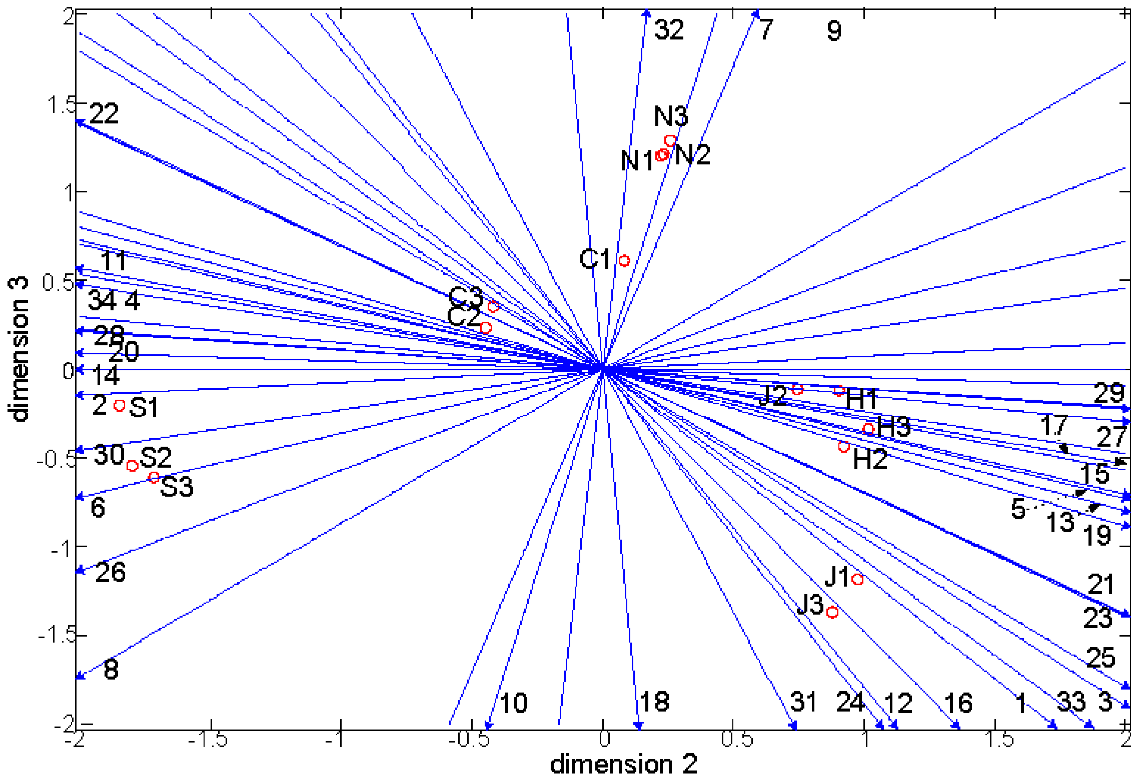
Candidate adjectives were chosen according to criteria (1) and (2). Criterion (3) was used to decide the final list of 17 adjectives which were chosen as semantic primitives: bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast and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



(a)


dimension 1
dimension 3
N3
N2
N1
C1
C3
C2
H1
H2
H3
J2
S1
S2
S3
J1
J3

(b)


dimension 2
dimension 3
N3
N1
N2
C1
C3
C2
J2
H1
H3
H2
S1
S2
S3
J1
J3

(c)

Fig. 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 Fig. 2.

3.1.2 Fuzzy inference system (FIS)

3.1.2.1 Why apply fuzzy logic?

We suggest that the relationship between expressive speech categories and semantic primitives 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. In this sense, traditional statistical methodology may not be appropriate for solving the problem; rather, fuzzy logic appears to be better suited.

The reasons are the following:

(1) fuzzy logic embeds existing structured human knowledge (experience, expertise, heuristics) into workable mathematics (Kecman 2001), and this is exactly what the model proposes to do in dealing with the perception of expressive speech.

(2) fuzzy logic is based on natural language (Jang, Sun, and Mizutani, 1996), and the natural language used in our model is in the form of semantic primitives.

(3) fuzzy logic models nonlinear functions of arbitrary complexity (Wolkenhauer, 2001), 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.

3.1.2.2 Experiment 4

In order to build FIS with highly reliable probability, we conducted another experiment for collecting more utterances as input data.

- Method

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 middle rating of "3", and two of them had the lowest rating of "1", which totaled to 35 utterances as the stimuli for experiment 4.

The subjects were identical to those in the previous experiments. 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 17 adjectives, which were listed in Table 3, 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.

3.1.2.3 Construction

By applying adaptive neuro-fuzzy (Chiu, 1994), 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 3.1.1.1 and 3.1.1.3, 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 utterances of the 50 utterances.

  An initial FIS model, a first-order Sugeno type model (Sugeno, 1985). 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 backpropagation 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 verify the resultant FIS of step 2**

The purpose of Step 3 is to use checking data to help model validation. 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.

3.1.2.4 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, Fig. 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 (N) and the semantic primitive Monotonous on the left and another non-linear relationship between Hot Anger (HA) and Monotonous on the right.

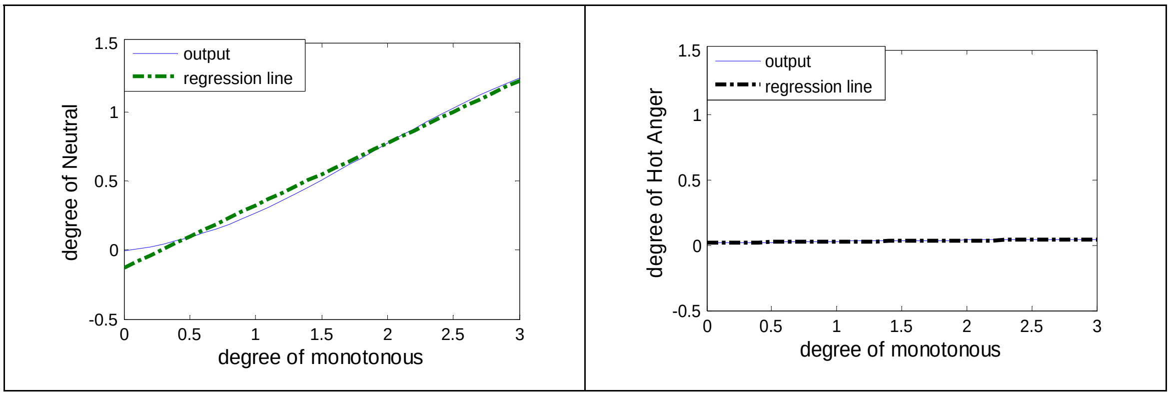


Fig. 4. Slope of regression line. Left graph describes the relationship between monotonous and Neutral (N), right graph describes the relationship between monotonous and Hot Anger (HA).

Table 4: 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 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	well-modulated	0.124
clear	0.127	unstable	0.063	well-modulated	0.091	weak	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
weak	-0.181	weak	-0.036	clear	-0.062	sharp	-0.079	quiet	-0.047

3.1.3 Results

The relationship between semantic primitives and expressive speech categories are characterized in Table 4. We selected the “top five” semantic primitives 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 4; the other FIS results also seem to be good matches with expressive speech utterances.

3.2 The relationship between semantic primitive and acoustic feature

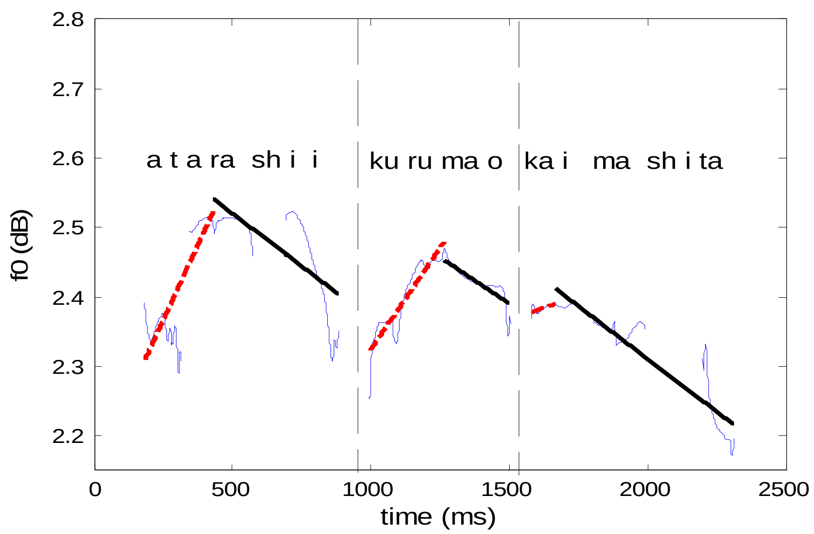
The relationship between semantic primitives and acoustic features was calculated by correlation analysis. The acoustic features measured were F0 contour, power envelope, power spectrum, and time duration.

3.2.1 Acoustic feature analysis

F0 contour, power envelope, and spectrum were calculated by using STRAIGHT (Kawahara et al, 1999). In the analysis, FFT length was 1024 points and frame rate was 1ms. These parameters are default to STRAIGHT, when the sampling frequency was 22050 Hz, which we adopted for sound data

- 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"[2], because each 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 5 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).



---

[2] The use of "accentual phrase" is used generically here to refer to both intonational phrases as well as accentual phrases, without necessarily distinguishing between the two, as is done in the JToBI system (Venditti, 2005)

Fig. 5. 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 high frequency portion (above 3 kHz) and the average power (RHT) were measured.

- Spectrum

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

  - Formants: 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 Nguyen et al., 2003. The first, second, and third formants (F1, F2, and F3) were calculated with LPC-order 12.
  - Spectral tilt (ST): To measure voice quality, spectral tilt was calculated from A1-A3, where A1 is the level in dB of the first formant and A3 is the level of the harmonic whose frequency is closest to the third formant (Maekawa, 2004). Other measures of voice quality are H1-H2 (Menezes and Maekawa, 2006) or OQ (Ni Chasaide and Gobl, 1997; Keating and Esposito, 2006), but only A1-A3 was done in the study.
  - Spectral balance (SB): This is a parameter that serves for the description of acoustic consonant reduction, and was calculated according to the following equation (Kienast and Sendlmeier, 2000):

$$SB = \frac{\sum f_i \cdot E_i}{\sum E_i} \quad (1)$$

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 ratio of consonant duration to vowel duration (RCV).

3.2.2 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 high frequency portion (above 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).

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 5.

Table 5: Correlation coefficients between the semantic primitives and the acoustic features.

PF	bright	dark	high	low	strong	weak	calm	unstable	well modulated	monotonous	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

4. Results and discussion

Based on the results described in Section 3, a perceptual model for each category of expressive speech was built. Figures 6-10 illustrate the perceptual model for Neutral (N), Joy (J), Cold Anger (CA), Sadness (S), and Hot Anger (HA), 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 Fig 7 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 5, are those related to F0 contour and power envelope. From this, 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 5 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.

The resulting perceptual models indicate 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. 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, we adopt the analysis by synthesis method. 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 3.2.2 above. In this way, we are able to assess the validity of the relationship between semantic primitives and acoustic features purported by our perceptual model.

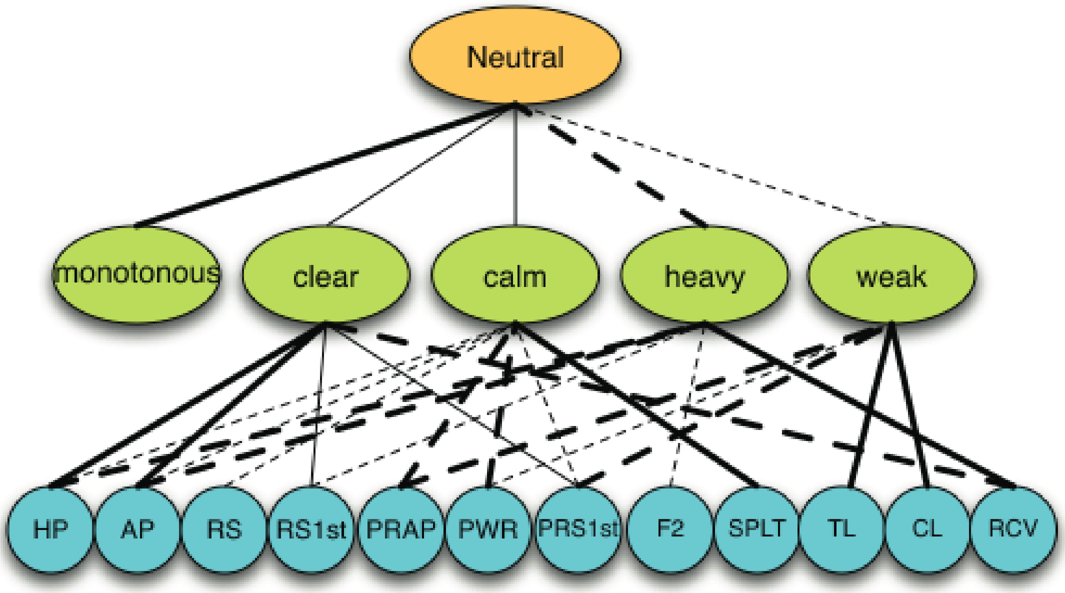


Fig. 6. Resultant perceptual model of Neutral (N).

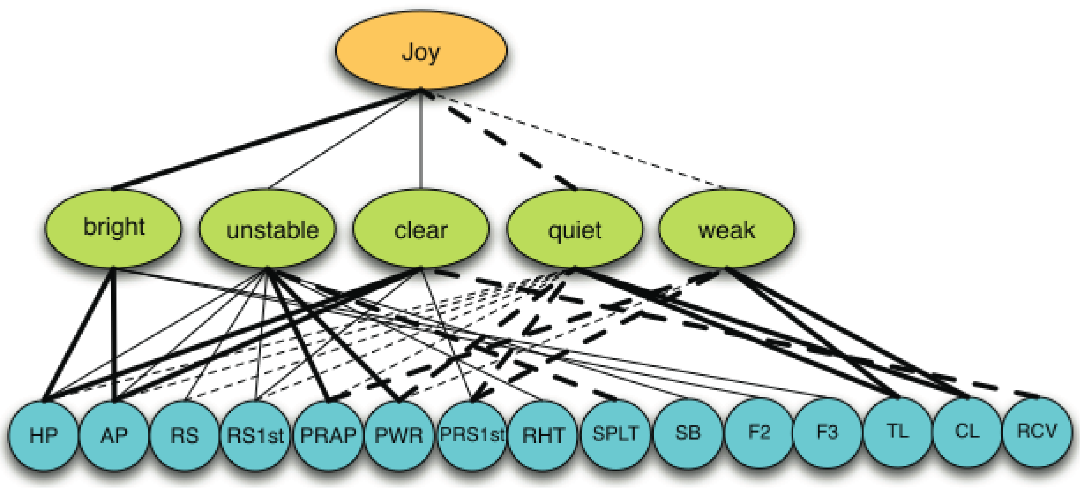


Fig. 7. Resultant perceptual model of Joy (J).

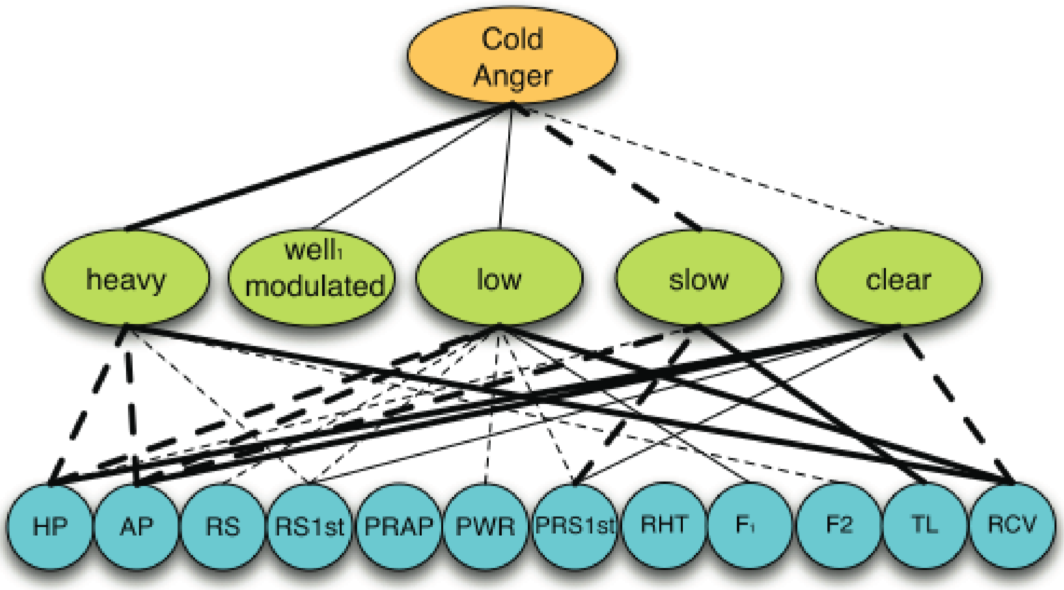


Fig. 8. Resultant perceptual model of Cold Anger (CA).

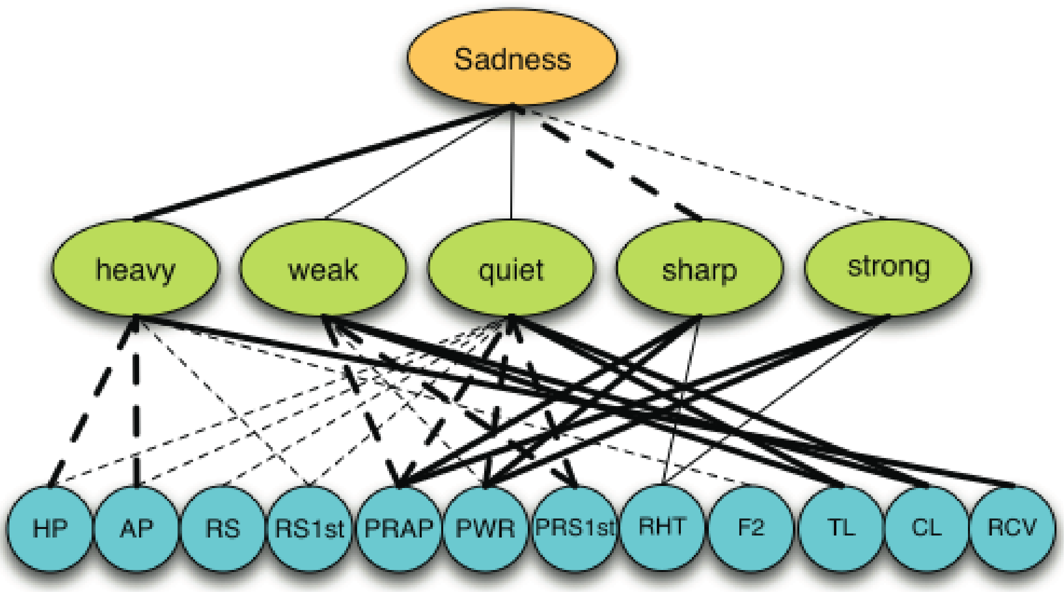


Fig. 9. Resultant perceptual model of Sadness (S).

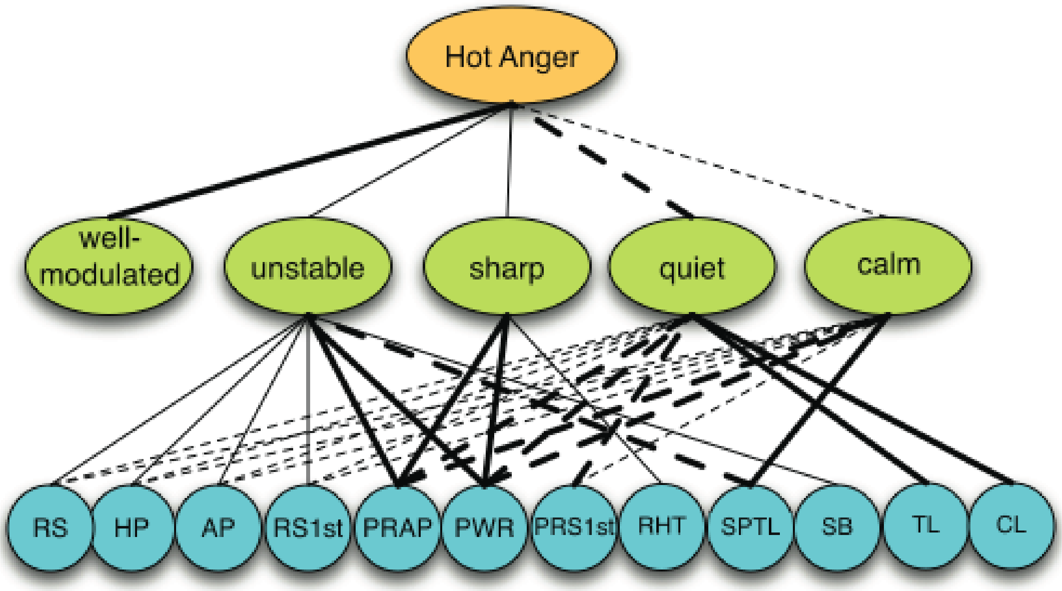


Fig. 10. Resultant perceptual model of Hot Anger (HA).

5. Verification of the emotional perception model

The purpose of the verification process is to validate the effectiveness of the two types of information in the relationships built in the modeling process. 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, in Fig 7, we know 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.

To focus the verification process on the validation of analysis results, a morphing technique is used. By using the morphing technique, we can take a neutral utterance as input, and by systematically manipulating acoustic features of the input utterance, we can 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.

5.1 Rule development

To simplify the manipulation of acoustic features in the morphing process, the models shown in Figs 7 to 10 are represented as rules. Two types of rules (i.e., morphing rules) were developed: (1) base rules and (2) intensity rules. The base rules provide the basis for the verification of the models. They are created 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). The intensity rules are created 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.

For verifying the relationship between acoustic features and semantic primitives, the base rules morphed a neutral utterance into an utterance which could be perceived in terms of one and only one semantic primitive. These rules are called “SR-base” rules, because they morph a “semantic primitive” utterance (i.e., “SU-utterance”) and help assess which acoustic features are involved in creating the percept of each semantic primitive. The intensity rules involved morphing rules which morphed an utterance in such a way that the intensity of the semantic primitive changed. These intensity rules are called “SR-intensity” rules, because they change the intensity of the “SU” and thus help assess how the change in the intensity of the acoustic features changed the intensity levels of semantic primitives.

For verifying the relationship between semantic primitives and expressive speech, the base rules (referred to as “ER-base” rules) involved rules to morph a neutral utterance into an utterance which could be perceived in terms of one expressive speech category (i.e., “EU-utterance). The intensity rules (referred to as “ER-intensity” rules) involved rules to morph an utterance in such a way that the intensity of the expressive speech (EU) category changed.

For both the 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.

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

5.2 The Verification of the relationship between semantic primitives and acoustic features

5.2.1 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, we need (1) 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 5), and (2) 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. 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 3.1.2.2), 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, we calculated the differences between the values of its acoustic features and the values of the acoustic features of the neutral utterance from which it should be morphed. 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, we averaged the percentage variations of each of the 9 utterances to give the values of acoustic features for each semantic primitive. Equation (2) presents the calculation.

$$\frac{\sum_{i=0}^{8} \frac{vaf_i - vnaf_i}{vnaf_i}}{9} \quad (2)$$

Where vaf_i is the value of acoustic features of i th utterance and $vnaf_i$ is the value of the corresponding neutral utterance. One example of the parameters is shown in Table 7. 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.

5.2.2 Rule implementation

A speech morphing process was developed (see Fig. 11) in order to implement the rules. 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 rule. Finally, we used STRAIGHT to re-synthesize the expressive speech utterance using the modified F0 contour, power envelope, spectrum and duration.

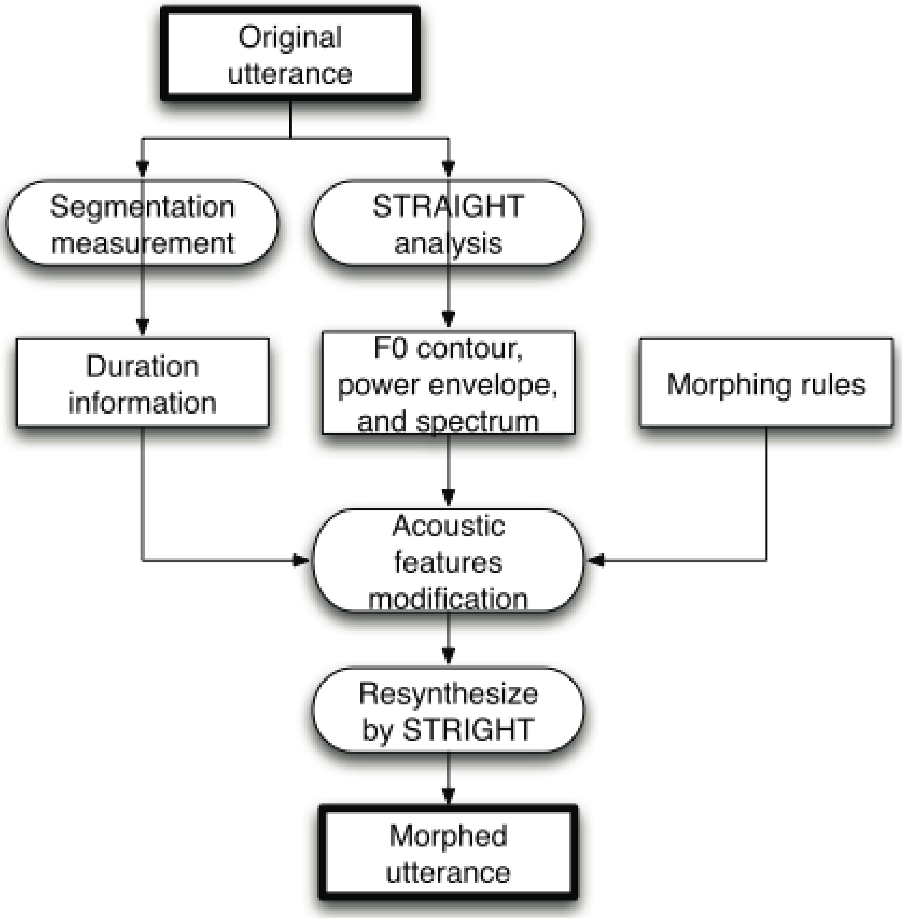


Fig. 11. Process of morphing voices in STRAIGHT

5.2.3 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.

5.2.3.1 Experiment

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, which were different from those subjects of the three experiments described in Section 3, with normal hearing ability who were asked to compare (a) a morphed speech utterance with (b) the neutral speech utterance and to choose which utterance was most associated with a particular semantic primitive. The question was "Is (a) or (b) more 'bright'?" Paired stimuli were presented randomly to each subject through binaural headphones at a comfortable sound pressure level.

5.2.3.2 Results and discussion

The results shown in Table 6 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 understand which of the selected acoustic features significantly influenced the perception of the semantic primitives. These results suggest that the created base rules are effective. 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 6 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%

Semantic Primitive	Accuracy Rates
clear	80%
noisy	100%
quiet	90%
sharp	90%
fast	100%
slow	100%

5.2.4 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 connected lines shown in Figs 7 to 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, we adjusted the parameters of the base rules so that the morphed speech utterance could be perceived as having different levels of intensity of the semantic primitives. We created three intensity rules (SR1, SR2, and SR3). 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 7 the line between Bright and AP (Average Pitch) is a solid thick line. Therefore, the value of the parameter AP was increased from, 106.9% to 110.9% (see Table 7 (a)). However, in Figure 7 the line between Bright and F3 is a solid thin line. Therefore, a smaller value is given to the parameter F3 from 104.2% to 106.2%. The parameters of SR1 come from the base rule of bright, which was calculated from Equation (2).

Another example is shown in Table 7 (b). It shows that the rules of Heavy are created by the same method. However, in Figure 6 the lines between Heavy and AP and between Heavy and HP (highest pitch) are dashed thick lines, therefore, the values of AP and HP were decreased from 96.10% to 92.10% and from 95.40% to 91.40% respectively. In Figure 6, the lines between Heavy and RS1st (rising slope of the first accentual phrase) and between Heavy and F2 are dashed thin lines, therefore, the values of RS1st and F2 were decreased from 96.70% to 94.70% and from 94.30% to 92.30% respectively.

Table 7. Example of Rule Parameters for Bright and Heavy. 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. Neutral utterances are 100%.

(a) Rule of Bright

Acoustic Feature	SR1	SR2	SR3
Highest Pitch (HP)	106.9%	110.9%	114.9%
Average Pitch (AP)	107.5%	111.5%	115.5%
F2	103.3%	105.3%	107.3%
F3	104.2%	106.2%	108.2%

(b) Rule of Heavy

Acoustic Feature	SR1	SR2	SR3
Highest Pitch (HP)	96.10%	92.10%	88.10%
Average Pitch (AP)	95.40%	91.40%	87.40%
RS1st	96.70%	94.70%	92.70%
F2	94.30%	92.30%	90.30%
RCV	114.80%	118.80%	122.80%

5.2.5 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).

5.2.5.1 Experiment

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, which were described above, respectively. 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 Fig. 12.

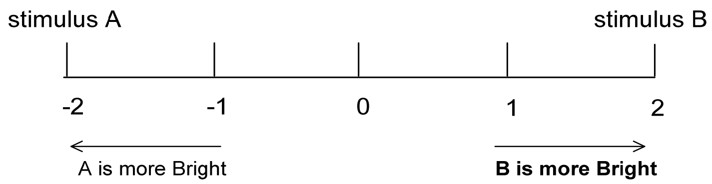
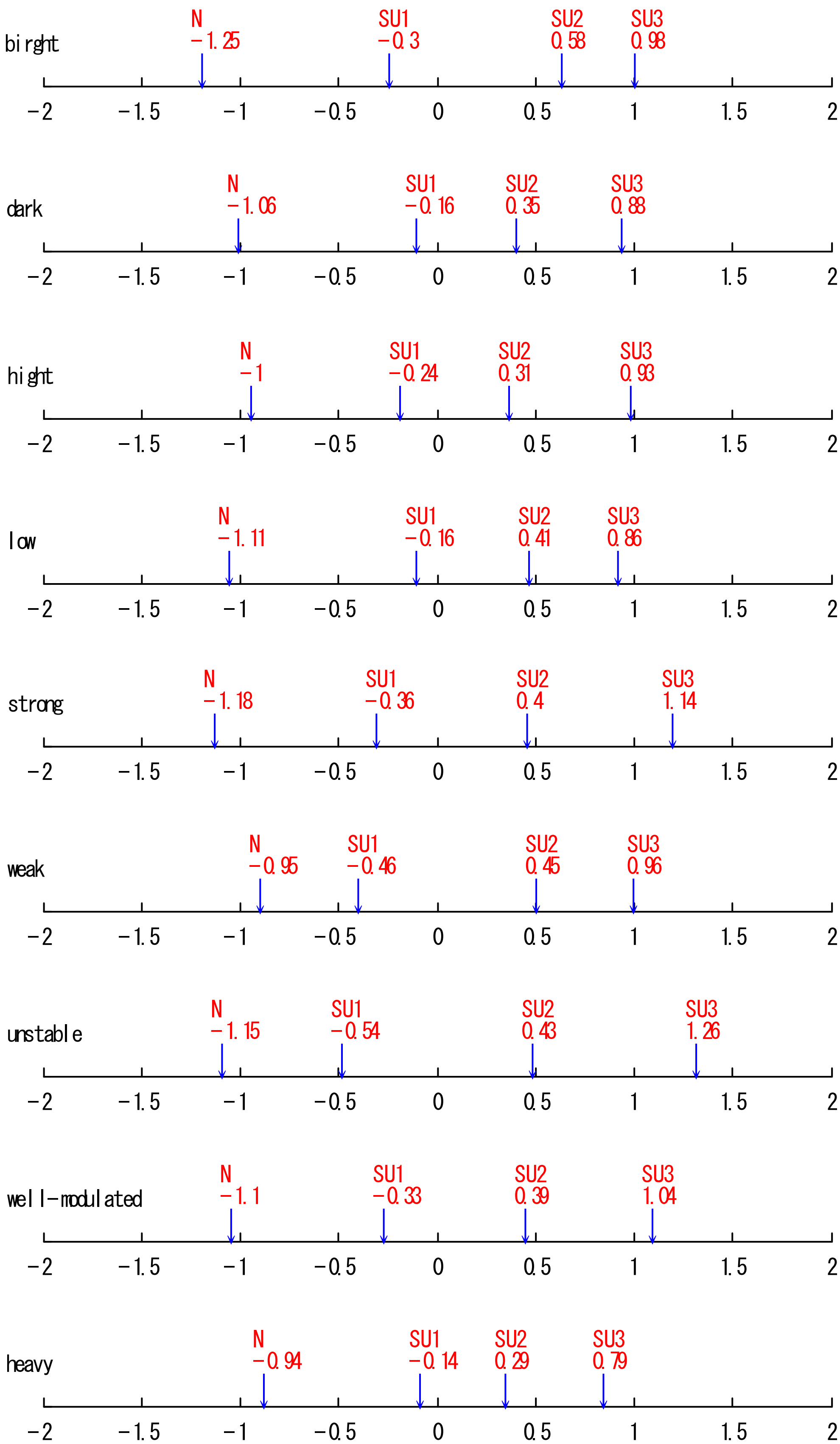


Fig. 12. Template of the questionnaire for assessing intensity of the semantic primitive "bright"

5.2.5.2 Results and discussion

Fig 13 shows the results of how listeners perceived the levels of intensity of each of the 4 utterances, i.e., 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.


birght
N -1.25
SU1 -0.3
SU2 0.58
SU3 0.98
dark
N -1.06
SU1 -0.16
SU2 0.35
SU3 0.88
hight
N -1
SU1 -0.24
SU2 0.31
SU3 0.93
low
N -1.11
SU1 -0.16
SU2 0.41
SU3 0.86
strong
N -1.18
SU1 -0.36
SU2 0.4
SU3 1.14
weak
N -0.95
SU1 -0.46
SU2 0.45
SU3 0.96
unstable
N -1.15
SU1 -0.54
SU2 0.43
SU3 1.26
well-modulated
N -1.1
SU1 -0.33
SU2 0.39
SU3 1.04
heavy
N -0.94
SU1 -0.14
SU2 0.29
SU3 0.79
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2

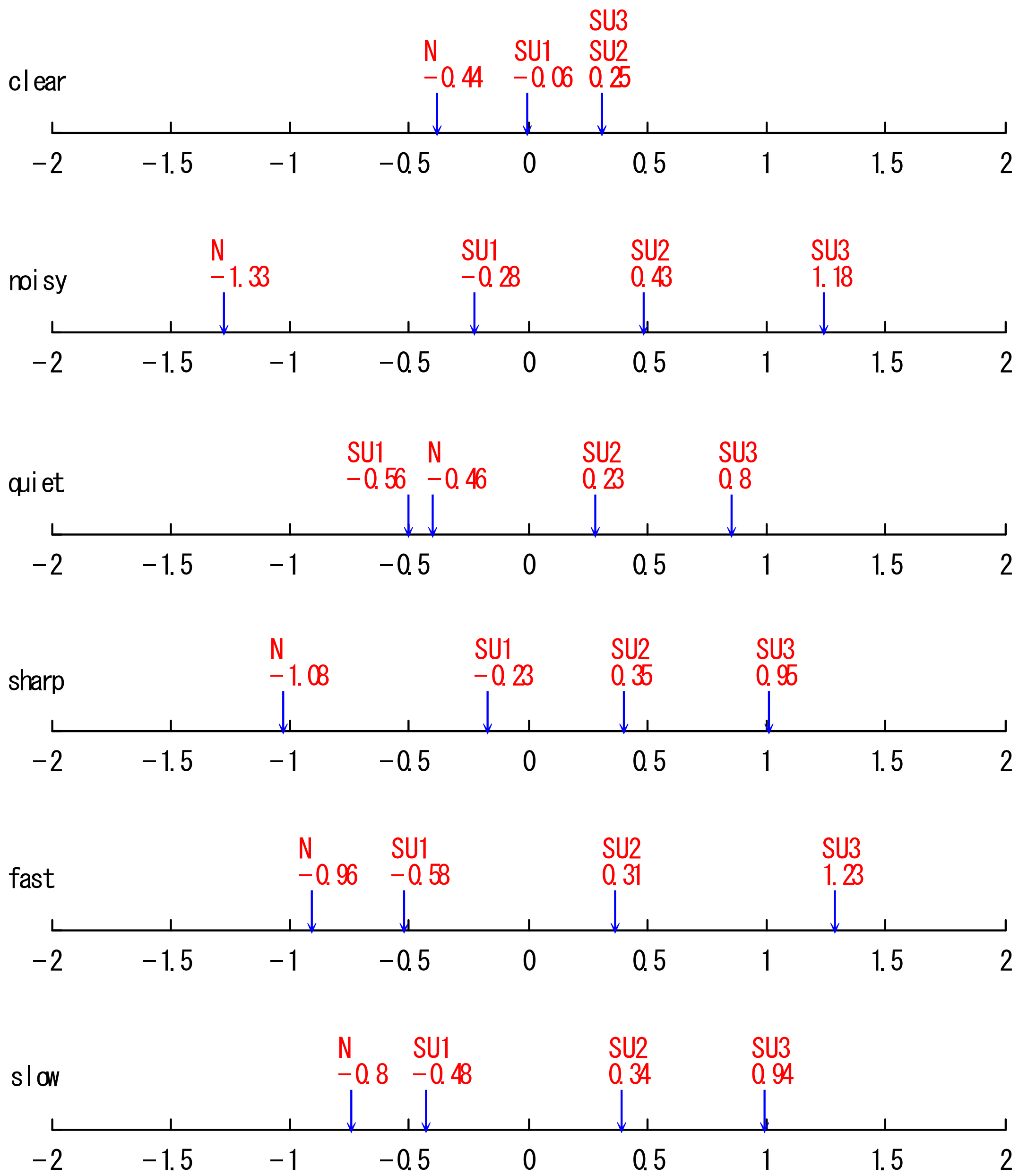


Fig. 13. Experimental results of semantic-primitive rule evaluation. The intended intensity levels are N < SU1 < SU2 < SU3. That is, SU3 should have a higher level of intensity perception than SU2 than SU1 than N. N is the neutral utterance, and SUn represent semantic primitive utterances created by the semantic primitive intensity rules.

5.3 The Verification of the relationship between expressive speech and semantic primitives

5.3.1 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, we need (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 Figs 7 to 10. For (1), only those semantic primitives shown in Figs 7 to 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 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 are shown in Table 4. 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.

5.3.2 Experiment for evaluating ER-base rule efficiency

To examine whether the combination of semantic primitive rules is 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.

5.3.2.1 Experiment

The experimental conditions including subjects were the same as in the previous experiments described in Section 5.2. 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'? "

5.3.2.2 Results and discussion

Table 8 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, even these expressive voices are morphed from neutral utterances.

Table 8. Experiment results of base rule evaluation

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

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

In order to create the intensity rules, we combine the parameters of the semantic-primitive intensity rules so that the morphed speech utterance could be perceived as having different levels of intensity of expressive speech. We created three intensity rules (ER1, ER2, and ER3). 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 9 shows an example of this way of combining intensity rules. As can be seen from Table 4 and Fig. 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. We use the same weight and weight combination values when creating expressive-speech base rules for combining the expressive-speech intensity rules here.

Table 9. 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

5.3.4 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).

5.3.4.1 Experiment

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 than the utterance (EU1) created by ER1, and the utterance (EU3) created by ER3 should have stronger perception 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, which were described above, respectively. 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.

5.3.4.2 Results and discussion

The results in Fig. 14 show the four stimuli of each expressive speech category listed in ascending order of the perception intensity. 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 we propose here in that it substantiates the relationship between semantic primitives and expressive speech categories.

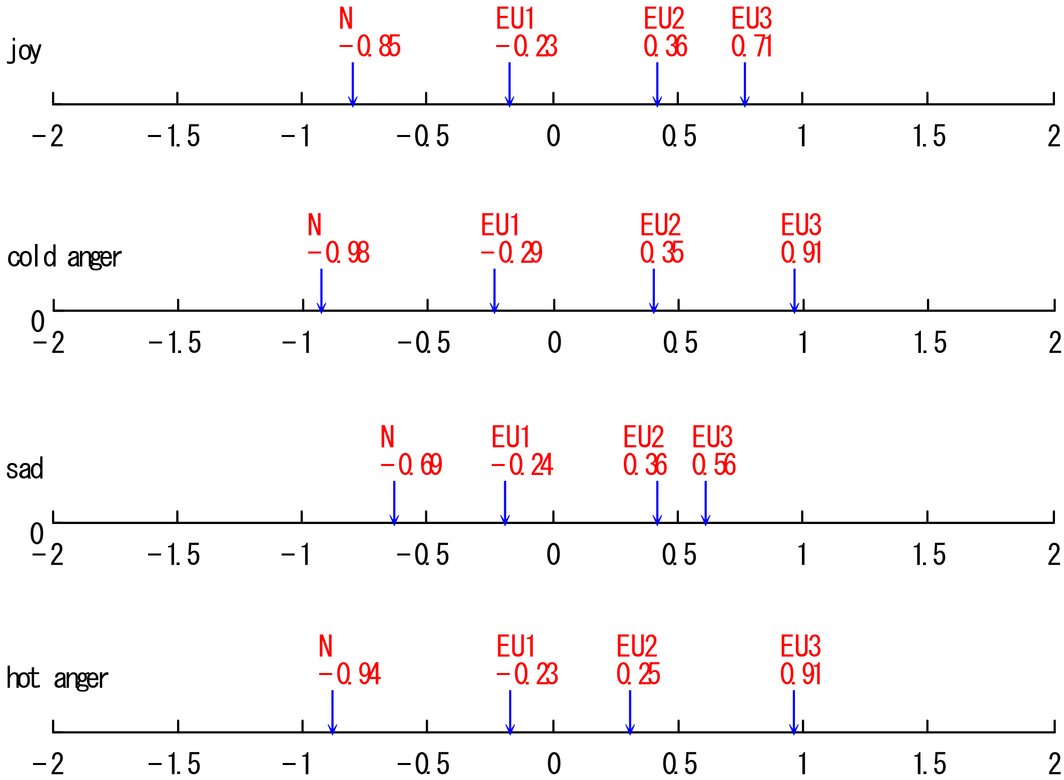


Figure 14. Experimental result of intensity rule evaluation. The intended intensity levels are N < EU1 < EU2 < EU3. That is, EU3 should have a higher level of intensity perception than EU2 than EU1 than N. N is the neutral utterance, and EUn represent expressive speech utterances created by the expressive speech intensity rules.

6. Conclusion

This paper proposes a perceptual model that explains the perception of expressive speech based on the observation that the perception of expressive speech may not be directly from the acoustic signal, but rather from semantic primitives, i.e., adjectives that people use to describe their feelings/attitudes upon hearing expressive speech. To explore this idea, we proposed a three-layer model: categories of expressive speech constitute the top layer, semantic primitives constitute the middle layer, and acoustic features the bottom layer. The construction of such a model should provide concrete support for the proposed ideas. There are two approaches for achieving this goal. One is to construct a general-purpose model which can explain the perception of expressive speech categories, across a number of speakers, based on examination of voices of multiple actors. The other is to construct a specific-purpose model based on the voice of a single actor. The purpose of this model would be to validate the method proposed here, to demonstrate that expressive voices can be simulated and controlled by using a multi-layered model.

Two relationships of the model were constructed by a top-down method. The relationship between expressive speech and semantic primitives was built by conducting three experiments and applying a fuzzy inference system (FIS) to the experimental results. The relationship between semantic primitives and acoustic features was built by analyzing acoustic features measured from the F0 contour, power envelope, power spectrum, and duration.

A bottom-up method in which speech utterances were morphed, based on two types of experimentally-derived rules (base rules and intensity rules), was used in order to verify the model. Based on these experiments with morphed speech, we found a significant relationship between semantic primitives and acoustic features. The ability of the model to successfully morph expressive speech utterances validated the semantic primitives chosen, although modification of the semantic primitives may be necessary given different expressive speech databases. The verification results validate our two assumptions: (1) different expressive speech utterances result in different types and intensity levels of perception of semantic primitives and (2) expressive speech categories can be perceived differently in terms of types and intensity levels as a function of a listener's perception of the types and intensity levels of semantic primitives.

The three-layered model outlined in this research is a new and different point of view for expressive speech perception. New in this approach is the concept of semantic primitives, used as the middle layer in the model. Also, new in this study are the use of a fuzzy inference system to mathematically describe the vague nature of the human interface between perception of semantic primitives and acoustic features in the speech signal.

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 Cowie and Cornelius (2003), Devilliers (2003), and Grimm (2007). 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. Another important future work is to extend the specific-purpose model we constructed here to include a general-purpose model based on multiple speakers to have wider applicability in explaining perception of expressive speech.

It is hoped this model 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. We hope to extend our work to include different speech databases (e.g. spontaneous voices), different listener populations (e.g. native vs. non-native listeners), as well as to include more psychologically-relevant levels in the proposed model structure. We hope that our study will be a stepping stone for future work in the ongoing exploration of expressive speech, and specifically, for examining how expressive speech categories and acoustic features are related to semantic primitives, as well as to what extent semantic primitives might be universal.

Acknowledgments

This research is conducted as a program for the “21st Century COE Program” by Ministry of Education, Culture, Sports, Science and Technology. We sincerely thank Fujitsu Laboratory for permission to use the voice database. This study was also supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan. We also sincerely thank Donna Erickson for her valuable comments.

Reference

Banziger, T., Scherer, K.R., 2005. The role of intonation in emotional expressions. Speech Communication 46, 252-267.

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

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

Chateau, N., Maffiolo V., Blouin C., 2004. Analysis of expressive speech in voice mail messages: the influence of speakers' gender. In: Proc. ICSLP2004, Korea.

Chiu, S., 1994. Fuzzy model identification based on cluster estimation, Journal of Intelligent and Fuzzy Systems, 2(3).

Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Proc. ICSLP96, Philadelphia.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32-80.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40, 5-32.

Darke, G., 2005. Assessment of Timbre Using Verbal Attributes. In: Proc. CIM05, Montreal.

Devillers, L.,., Vidrascu, L.,, Lamel, L., Challenges in real-life emotion annotation and machine learning based detection. Neural Networks, Volume 18, Issue 4, May 2005, Pages 407-422

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

Erickson, D., 2005. Expressive speech: Production, perception and application to speech synthesis. Acoustical Science and Technology 26, 317-325.

Fauconnier, G., 1997. Mappings in Thought and Language, Cambridge University Press.

Friberg, A., 2004. A fuzzy analyzer of emotional expression in music performance and body motion. In: Proc. Music and Music Science, Stockholm, 2004

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

Fujisaki, H. Manifestation of Linguistic, Para-linguistic, non-linguistic Information in the Prosodic Characteristics of Speech, IEICE, 1994-0

Grimm, M., Mower, E., Kroschel, K., and Narayanan, S. Primitives based estimation and evaluation of emotions in speech. Speech Communication, 2007.

Huang, C.F and Akagi, M., 2005. Toward a Rule-Based Synthesis of Emotional Speech on Linguistic Descriptions of Perception. Lecture notes in computer science, Volume 3784/2005, pp. 366-373.

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

Juslin, P. N, 2001. 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.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. 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 27,187-207.

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

Kendon, A. (ed.),1981.Nonverbal Communication, Interaction and Gesture: Selections from Semiotica, Approaches to Semiotics 41). The Hague: Mouton and Co.

Kienast, M., Sendlmeier, W. F., 2000. Acoustical analysis of spectral and temporal changes in expressive speech, ISCA workshop on speech and emotion, Belfast.

Maekawa, K., Kitagawa, N., 2002. How does speech transmit paralinguistic information? Congnitive Studies 9, 46-66.

Maekawa, K., 2004. Production and perception of ‘paralinguistic’ information. In: Proceedings of Speech Prosody, 367-374, Nara

Manning, P. K., 1989. Symbolic Communication: Signifying Calls and the Police Response, The MIT Press

Mehrabian, A., 1972. Nonverbal communication. Aldine-Atherton, Chicago, Illinois.

Murray, I. R., Arnott, J. L., 1993. 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), 1097-1108.

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

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

Osgood, C., May, W. H., & Miron, M. S., 1975. Cross-cultural universals of affective meaning. Urbana: University of Illinois Press.

Raskin, J., 2000. The Human Interface: New Directions for Designing Interactive System, Pearson Education.

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

Scherer, K. R., Ladd, D.R., Silverman, K. A., 1984. Vocal cues to speaker affect: Testing two models. Journal of the Acoustical Society of America 76, 1346-1356.

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

Scherer, K. R., 2003. Vocal communication of emotion: a review of research paradigms.

Speech Communication 40, 227-256.

Schroder, M., 2001. Expressive speech synthesis--a review. In: Proc. Eurospeech 2001, 561-564, Aalborg.

Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S., 2001. Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proc. Eurospeech 2001, 87-90, Denmark.

Sobol Shikler T., Robinson P., 2004. Visualizing dynamic features of expressions in speech, In: Proc. ICSLP2004, Korea.

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

Tolkmitt and Scherer, 1986 F.J. Tolkmitt and K.R. Scherer, Effect of experimentally induced stress on vocal parameters, J. Exp. Psychol. [Hum. Percept.] 12 (1986) (3), pp. 302–313.

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

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

Ueda, K., 1996, A hierarchical structure for adjectives describing timbre, Journal of the Acoustical Society of America.

Ueda, K., 1988. Should we assume a hierarchical structure for adjectives describing timbre? Acoustical Science and Technology 44(2), 102-107 (in Japanese)

Van Bezooijen, 1984 R. Van Bezooijen, The Characteristics and Recognizability of Vocal Expression of Emotions, Foris, Drodrecht, The Netherlands (1984).

Venditti, Jennifer J., 2005. The J_ToBI model of Japanese intonation. In S.-A. Jun (ed.)

Prosodic Typology: The Phonology and Intonation of Phrasing, pp. 172-200. Oxford University Press.

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

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

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

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