A Multi-Layer Fuzzy Logical Model for Emotional Speech Perception

Chun-Fang Huang and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology, Nomi-shi, Ishikawa, Japan
(chuang, akagi)@jaist.ac.jp

Abstract

This paper proposes a three-layer model for the perception of emotional speech. Much of the earlier work has focused on the relationship between emotional speech and acoustic features that is characterized by statistics. The problem is that there lacks a model which takes the aspects of vagueness and human perception into consideration. In the model proposed here, five categories of emotion constitute the topmost layer of the model. Primitive features, which are the linguistic form of adjectives, constitute the middle layer, which is used to describe the human perceptual aspect. The bottommost layer is acoustic features. Three experiments followed by MDS analysis revealed suitable primitive features. And then, fuzzy inference systems were built to map the vagueness nature of the relationship between emotions and the primitive features. Acoustic features including F0, time duration, power envelope, and power spectrum were analyzed; subsequent regression analysis revealed correlation between the primitive features and the acoustic features. The experimental results and the resulting fuzzy inference systems show a significant relationship between emotions and the primitive features. The results of the analysis also show some acoustic features that have positive or negative correlation with primitive features.

1. Introduction

Both perception and expression are essential to emotional speech in human communication. With technological advances, many applications such as human-computer interaction [1] or customer calling services [2] create demands for understanding the perception of emotions in speech, and further, for more natural speech synthesis. In computer science, most research has measured acoustic features in the speech signal and then used statistical methodology to select the most significant features for the classification or identification of emotions [3] [5]. Although this method has allowed us to discover that certain categories of emotional speech may be characterized by some acoustic features, it fails to give us any way to reproduce those same effects in synthesized speech according to the discovered acoustic features [6]. We consider the problem is due to the lack of a model that takes the aspect of human perception into consideration.

Although many implicit factors affect perception [7] [8], one phenomenon can be noticed. When listening to the voice of a speaker, our first sense is something like “it sounds very bright and slightly fast”, which we then interpret as “the speaker is happy”. But, we never perceive emotion from one voice by saying “this voice sounds as if its fundamental frequency is 300 Hz”. In this study, we call such a “bright-sounding voice” or “fast-sounding voice” a primitive feature. The observation indicates three points of view. (1) The emotion we perceive in speech may depend on what we have sensed, such like “bright”; from the voice. (2) Humans describe their perception of phenomenon with vague linguistic forms not precise values. This human vagueness should be considered. (3) Although human nature is vague, a precise analytical/mathematic approach to deal with the vagueness of humans is needed.

Taking these three points of view into account, we propose a perceptual model. The first point of view results in a three-layered model that is different from traditional two-layer models, which only consider emotional speech and acoustic features. The last two points of view result in the application of fuzzy logic. Linguistic form is vague, uncountable, and also related to human knowledge of communication. The characteristics of linguistic forms correspond to the concepts of fuzzy logic.

The purpose of the study is to build and verify the model. A two-phase approach was taken to achieve the purpose. The first phase builds the model by a top-down method and the second phase verifies the model by a bottom-up method, where the top-down method is analysis and the bottom-up method is synthesis. In this work, we build the perceptual model that includes two relationships by a top-down method. The first relationship, between categories of emotion and primitive features, is built by conducting three experiments and applying fuzzy inference systems. The second relationship, between primitive features and acoustic features, is built by analyzing acoustic features in speech signals. Combining the two relationships, we show a significant model for perception of emotional speech.

The remainder of this paper is organized as follows. In Section 2 we introduce to the proposed model. In Sections 3 and 4, we present how the first and the second relationship are built, respectively. In Section 5 we present the resulting model, and Section 6 concludes with a discussion and suggestions for future work.

2. The three-layer model

![Figure 1: Conceptual diagram of the perceptual model](image)

Figure 1: Conceptual diagram of the perceptual model

417

September, 4-8, Lisbon, Portugal
Figure 1 shows a conceptual diagram of the perceptual model. It consists of three layers, emotional speech, primitive features, and acoustic features, where the emotional speech includes five categories, Neutral (N), Joy (J), Cold Anger (CA), Sadness (S), and Hot Anger (HA). The primitive features are considered as a set of adjectives often used to describe speech. The acoustic features are a set of the acoustic features of speech signals.

3. Relationship between emotional speech and primitive features

In order to build this relationship, three experiments were conducted initially for choosing suitable primitive features. Fuzzy inference system was then applied to build the relationship. The first experiment investigated all utterances in a database in terms of each category, the second experiment constructed the perceptual space of categories of emotion, and the third experiment selected primitive features for the perceptual model. The experimental results were used as training data and checking data for the fuzzy inference system that maps the relationship between emotion and primitive feature.

3.1. Three experiments

Experiment 1 was conducted to examine utterances in terms of emotion. Stimuli were selected from the database produced by Fujitsu Laboratory and recorded by a professional actress. Altogether, 171 Japanese utterances in 19 sentences were selected. Each sentence had one utterance in N and two utterances in each of the other categories, for a total of 9 utterances. A group of 12 graduate students participated in this experiment. Listeners rated the stimuli according to the perceived degree of the emotion. Table 1 shows the confusion matrix of each intended category. Clearly, most utterances can be easily perceived as belonging to their intended categories. CA had the lowest percentage and is easily confused with N. However, the one most confused with N is J. This experiment contributes to the investigation of emotion in the perceptual model; furthermore, the rating results contribute to the selection of appropriate utterances in subsequent experiments.

Table 1: Confusion matrix of 5 intended categories.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>N</th>
<th>J</th>
<th>CA</th>
<th>S</th>
<th>HA</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>0.9751</td>
<td>0.1221</td>
<td>0.0962</td>
<td>0.0541</td>
<td>0.0101</td>
</tr>
<tr>
<td>J</td>
<td>0.0024</td>
<td>0.8705</td>
<td>0.0006</td>
<td>0</td>
<td>0.0025</td>
</tr>
<tr>
<td>CA</td>
<td>0.0199</td>
<td>0.006</td>
<td>0.8599</td>
<td>0.0293</td>
<td>0.0145</td>
</tr>
<tr>
<td>S</td>
<td>0.0026</td>
<td>0.0072</td>
<td>0.038</td>
<td>0.9173</td>
<td>0.0003</td>
</tr>
<tr>
<td>HA</td>
<td>0</td>
<td>0.0002</td>
<td>0.0056</td>
<td>0</td>
<td>0.9896</td>
</tr>
</tbody>
</table>

Experiment 2 was conducted to construct a perceptual space of utterances in different categories. It was followed by an analysis using a multidimensional scaling technique (MDS). The resulting perceptual space was used in the next experiment to select suitable primitive features for the perceptual model. In Experiment 2, 15 utterances were chosen according to the ratings in Experiment 1: for each of the five categories, three utterances were selected that had the highest, middle, and lowest rating. The subjects were identical to Experiment 1. A paired comparison was used so that subjects rated each utterances pair according to how similar they perceived them to be. The rating result was analyzed using MDS. Figure 2 shows the distribution of utterances in the resulting 3-dimensional perceptual space (STRESS value was 0.07). In the figure, one circle represents one utterance and plot symbols like ‘J’ stands for the distribution of the utterances of J. All categories of emotional speech are separated clearly. Also, utterances of the same category are close to each other, which means the distribution in perceptual space can appropriately present the similarity of the utterances and the position of each emotion. Therefore, it is reliable for determining what primitive features are suitable in Experiment 3.

Experiment 3 was conducted to determine suitable primitive features for the perceptual model. The voice data and subjects were the same as in Experiment 2. In this experiment, subjects rated 34 adjectives in terms of how suitably the adjectives described the utterance they heard. In order to clarify how each adjective was related to each category, 34 adjectives were superimposed into the perceptual space with the application of multiple regression analysis. Figure 3 shows one of resultant diagrams in 2 dimensions, where a number indicates an adjective and an arrow indicates the direction of one adjective. Utterances are represented by the same id as in Figure 2.

Subsequently the multiple correlation coefficients between adjectives and utterances were calculated. Primitive features were selected, according to three criteria. (1) The direction of each adjective in the perceptual space. This
indicates which emotion the adjective is most related to. For example, the adjective Clear (24) is more closely related to emotion Joy. (2) The angle between each two adjectives. The smaller the angle is, the more similar the two adjectives are. (3) The multiple correlation coefficient of each adjective.

Finally, 17 adjectives were chosen as primitive features. These are bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast and slow.

3.2. Fuzzy inference system

3.2.1. Construction

To build the relationship between emotion and primitive feature it is first necessary to clarify what the relationship means. We consider the relationship represents how we humans use linguistic forms to describe what we perceive in speech. That expression is vague, not precise. Therefore, traditional statistical methodology may not be appropriate for solving this problem. It is considered that fuzzy logic is well suited for building this relationship because (1) fuzzy logic embeds existing structured human knowledge (experience, expertise, heuristics) into workable mathematics [9]; this corresponds to what the model proposes to deal with, the perception of emotional speech; (2) fuzzy logic is based on natural language [10], which corresponds to primitive features in linguistic form; and (3) fuzzy logic models nonlinear functions of arbitrary complexity [11], which corresponds to the nonlinear and complex relationship between emotions and primitive features.

Thus, the relationship was built using the MATLAB Fuzzy Logic Toolbox to deal with results of those experiments. For each emotion, a fuzzy inference system (FIS) was constructed, where the input is perceptible degrees of primitive features and the output is perceptible degrees of emotions. FIS construction steps are: (1) construct an initial FIS by applying the method of subtractive clustering to analyze the experimental results, (2) train the initial FIS by adaptive neuro-fuzzy methodology[12], (3) refined the resultant FIS of (2) by reducing the number of membership functions, and (4) train the resultant FIS of step 3 by adaptive neuro-fuzzy methodology.

3.2.2. Evaluation

In order to evaluate the relationship built by FIS, we calculated regression lines that describe the relationship between inputs (perceptible degrees of primitive features) and output (perceptible degrees of emotional speech) of each FIS. The absolute values of the slopes indicate how much the primitive features affect the categories of emotional speech. A positive value of slope indicates that the relationship has a positive correlation, and vice versa.

3.3. Results

As Table 2 shows, the relationship built by FIS corresponds to our behavior when we perceive emotional speech. For example, a joyful voice always sounds bright but not quiet. This matches the FIS result for Joy. Therefore, the relationship built by FIS is considered reasonable.

4. Relationship between primitive and acoustic features

The relationship between the primitive and acoustic features was built by analyzing acoustic features in speech signals in terms of F0 contour, power envelope, power spectrum, and time duration.
4.1. Acoustic analysis

F0 contour, power envelope, and power spectrum were calculated by STRAIGHT [13] with a FFT length of 1024 points and a frame rate of 1ms. The sampling frequency was 22050 Hz. We also measured acoustic features on the basis of two aspects – accentual phrase and overall utterance – because most people do not speak continuously. For example, in Japanese the sentence /a ta ra shi i ku ru ma o ka i ma shi ta/ was always spoken with pauses in such a way / a ta ra shi i ku ru ma o ka i ma shi tu/, forming 3 accentual phases. By comparing utterances that are high same sentence but spoken in different categories, the variation of F0 contour and power envelope in both accentual phrases and overall utterances was evident. Taking this factor into account, some acoustic features are measured in each accentual phrase of an utterance, such as the raising slope of F0, and some are in an utterance, like max F0.

Eight acoustic features are measured from the F0 contour, eight from the power envelope, three from the power spectrum, and seven from the duration. Correlation coefficient values between acoustic features and those primitive features that have at least one correlation coefficient over 0.6 are considered significant and are listed in Table 3. The shadowed cells indicate that the correlation coefficients are over 0.6. There are a total of 14 acoustic features. Four are for F0: mean value of rising slope (RS), average pitch (AP) highest pitch (HP) and rising slope of the first accentual phrase (RS1st). Four are for power envelope: mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (RS1st), and the ratio between the average power in high frequency portion (over 3 kHz) and the average power (RHT). Five are for power spectrum: the first, second, and third formants (F1, F2, and F3), spectral tilt (SPTL), and spectral balance (SB). Three are for duration: total length (TL), consonant length (CL), and the ratio between consonant length and vowel length (RCV) [4] [14].

5. Resultant perceptual model

Combining the two relationships described in Section 3 and Section 4, Figure 5 shows the resultant perceptual model of the emotion Joy. The solid lines indicate the relation is a positive correlation, and the dotted ones indicate a negative correlation. The thicker the line is, the higher the correlation.

6. Conclusion and future work

In this paper we describe our proposed three-layer model. Two relationships of the model were constructed by a top-down method. The first relationship was built by conducting three experiments and applying FIS to the experimental results. The second relationship was built by analyzing acoustic features measured from the F0 contour, power envelope, power spectrum, and duration.

The significance of the work is that it provides a different point of view for emotional speech perception. The experimental results show the significance of the relationship between emotions and primitive features. This paper also indicates that fuzzy logic can be used in audible perception. Finally, the analyzed results also show the significance of the relationship between the primitive and the acoustic features. They showed some acoustic features that have positive or negative correlation with primitive features.

In future work, it will be necessary to verify the perceptual model by a bottom-up method, that is, to resynthesize (morph) speech voice on the basis of primitive features rather than directly from acoustic features, and conduct experiments to see if the resynthesized voice can be successfully perceived.

7. Acknowledgements

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.

8. References