ACII 2005

Toward a Rule-Based Synthesis of Emotional Speech on Linguistic Descriptions of Perception

Chun-Fang Huang and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi-shi, Ishikawa, Japan
{chuang, akagi}@jaist.ac.jp

Abstract. This paper reports rules for morphing a voice to make it be perceived as containing various primitive features, for example, to make it sound more "bright" or "dark". In a previous work we proposed a three-layered model, which contains emotional speech, primitive features, and acoustic features, for the perception of emotional speech. By experiments and acoustic analysis, we built the relationships between the three layers and reported that such relationships are significant. Then, a bottom-up method was adopted in order to verify the relationships. That is, we morphed (resynthesized) a speech voice by composing acoustic features in the bottommost layer to produce a voice in which listeners could perceive a single or multiple primitive features, which could be further perceived as different categories of emotion. The intermediate results show that the relationships of the model built in previous work are valid.

1. Introduction

Traditionally, human-computer interaction is a sequence of instruction-reaction steps. The computer does what humans tell it to do. This interaction model requires self-adapting of human working habits to meet the responsive style of the computer. This model may be adequate for a computing environment such as a computer terminal. But in the current ever-changing environment and burgeoning information era, a new computer interaction model is required. In order to create such a new interaction model, we could follow one of two paths. One would be to create a computing environment that is designed by considering how human minds work, and making it as easy as possible [1]. The other is to create a computing environment that instead of asking humans to behave in a self-adapting way, programs the computer to respond differently to coordinate with the human's emotional state [2]. With regard to the second situation, in order to create a computer-adapting environment, we must give computers the ability to sense human emotional states. There are many ways that humans express their emotional states intentionally or spontaneously, such as through facial expression [3] or voice [4]. Due to the maturity of voice recognition technology, one of the most effective ways to help computers adapt themselves might be by voice.

Generally, research on expression in speech falls into two categories. One is the perception of emotional states in speech; the other is the production of emotional

speech. There are many different synthesis techniques for producing emotional speech,. Formant synthesis [5][6][7][8] provides a non-pre-recorded-voice approach and uses a set of rules to control different acoustic parameters. The resulting synthesis voice is less natural but has richer degree of control over varying parameters [10]. Conversely, concatenative synthesis concatenates pre-recorded speech segments (mostly triphones and diphones) to form the speech voice. The resulting synthesis voice, when compared to formant synthesis, is more natural but there is less control over varying parameters [10][11][12][13]. Another technique is morphing. It also uses pre-recorded speech of completed sentences and can be used to manipulate recordings [14][15].

For emotional speech, Figure 1 shows a conceptual diagram of the perceptual model proposed by Huang and Akagi [16]. Unlike most other studies that deal with the direct relationship between emotional speech and acoustic features [5][6][7], this model 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, such as bright or dark. The acoustic features are a set of the acoustic features of speech signals. The concept is based on the assumption that humans perceive emotion from speech according to a combination of different primitive features that they give to the utterance they hear. The concept is inspired from an observation that 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". From the observation, we conclude (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, and this human vagueness should be considered; and (3) although human nature is vague, a precise analytical/mathematic approach to deal with that vagueness is needed.

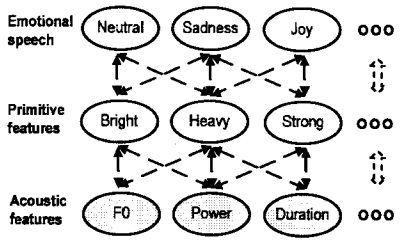

Fig. 1. Conceptual diagram of the perceptual model

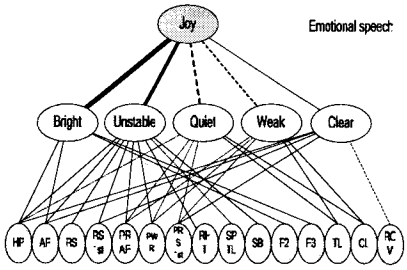

Fig. 2. Resultant perceptual model of emotion Joy. The solid lines indicate the relation is a positive correlation, and the dotted ones indicate a negative correlation. The thicker the line is, the higher the correlation.

A two-phase approach was taken to build the model. 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 our previous work, the first phase was accomplished. We built the perceptual model that includes two relationships. The first relationship, between categories of emotion and primitive features, was built by conducting three experiments and applying fuzzy inference systems. The second relationship, between primitive features and acoustic features, was built by analyzing acoustic features in speech signals. Combining the two relationships, we showed an efficient model for perception of emotional speech. Fig 2 shows one of the resultant perceptual models of the emotion "Joy".

The purpose of our current research is to verify the model, which is the second phase of model building. Our approach is to morph (resynthesize) a speech voice by composing the acoustic features of the bottommost layer to produce speech with the perception of single or multiple primitive features, which can further be perceived as different categories of emotion. To this end, the verification phase was divided into two stages. In the first stage, rules for morphing speech to make it be perceived as various primitive features were established. In the second stage, the rules of primitive feature perception were combined to produce rules for morphing speech to make it be perceived as different emotions. This paper describes how the rules in the first stage were established according to the results of experiments and acoustic analysis and an experiment that was conducted to evaluate the rule performance of relations between acoustic features and primitive features. The second stage is an item of future work.

The remainder of this paper is organized as follows. Section 2 describes the working flow of the current research and the principles of rule establishment. Section 3 gives the results for the rules of primitive features. Section 4 describes our conclusion and future work.

2. Rules of primitive features

In this section, the principles of rule establishment and the working flow of current research are described.

2.1 Principles

There are three mandating principles and one optional principle of rule establishment.

1. *Rules are monotonous.* This means any newly added rule will not weaken the perception of an existing rule
2. *Rules are general.* This means rules are applied to general perception, not any specific personality consideration.
3. *Rules are dynamic.* This means a newly added rule changes the configuration of existing rules but still be monotonous.

One optional principle is *the pursuit of naturalness*. The fact that it is optional does not imply naturalness is unimportant or undesirable to achieve. It is only because the focal point of our research is finding the effectiveness of synthesizing on the basis of primitive features. If we try to pursue naturalness along with the other principles, the focus of the research will be lost.

2.2 Working flow of the second phase

Figure 3 shows the working flow of our current research. For each stage of the verification phase, there are two steps, production and evaluation.

With regard to the first stage to create morphed voice for primitive feature perception, two types of rules for primitive features, atomic rules and compound rules (see Fig. 4), are developed that are designed for resynthsizing the speech voice so that it will be perceived as having different kinds of primitive features. It is an iterative process that includes:

1. Develop atomic rules for resynthesizing speech so that it might be perceived as one or more primitive features. More details about how the rules are developed based on the results of the acoustic analysis and experiments in our previous work will be introduced in the next section.
2. Implement rules using STRAIGHT [14].
3. Conduct perception experiments to verify the rules.
4. Refine rules according to the results of experiments.

Compound rules are for resynthesizing speech to make it be perceived as multiple types of primitive features. The process is identical to that for atomic rules. Compound rules are more complex than atomic rules. Currently only atomic rules have been developed.

In the evaluation step, we evaluate the effectiveness of our rules for primitive features. In this step, atomic rules and compound rules are used to resynthesize the speech voices, to make them be perceived as different types of primitive features and categories of emotional speech, and to conduct perception experiments to verify them. This method is also designed to verify the relationships built by the top-down method. In the following session, the establishment of atomic rules and the results are described.

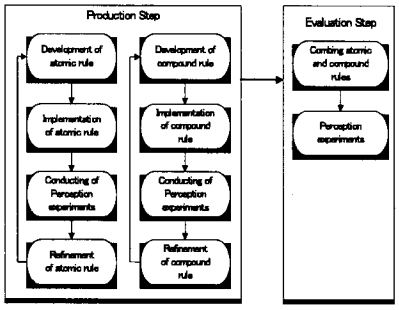

Fig. 3. Working flow of current research

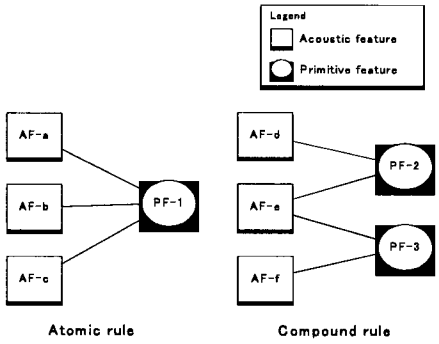

Fig. 4. Atomic rules and compound rules

An atomic rule is defined for each primitive feature. It defines the configurations of acoustic features that participate in it. A *configuration* consists of three types of parameters, the multi-regression correlation coefficient between the primitive feature and all acoustic features, a percentage variation, and ranges of values (maximum and minimum).

3 Methodology

In this section, the method of rule establishment and the intermediate experimental results are described.

3.1 Establishment of rules

There are three steps for creating rules.

Step 1: Analysis of acoustic feature
As was reported in our previous work [16], we 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/[1] was always spoken with pauses in such a way / a ta ra shi i ku ru ma o ka i ma shi ta/, forming 3 accentual phases. Nine acoustic features are measured from the F0 contour, eight from the power envelope, five from the power spectrum, and six from the duration. The acoustic features are the following.

F0 Contour: for the accentual phrase aspect, the features are mean value of rising slope (RS), mean value of rising duration, mean value of falling slope, mean value of falling duration, and mean value of pitch range, where *pitch range* is the band width of the range bounded by the lowest and highest F0 of each phase. For overall utterance aspect, they are average F0 (AP), pitch range, highest F0 (HP), and rising slope of the first accent phrase (RS1st).

Power Envelope: for the accentual phrase aspect, the features are mean value of rising slope (RS), mean value of rising duration, mean value of falling slope, mean value of falling duration, and mean value of power range (PRAP). For overall utterance aspect, they are power range (PWR), rising slope of the first accent phrase (RS1st), and the ratio between the average power in high frequency portion (over 3 kHz) and the average power (RHT).

Spectral: the features are first formant (F1), second formant (F2), third formant (F3), spectral tilt (ST), and spectral balance (SB).

Time Duration: pause length, phoneme length, total length (TL), consonant length (CL), vowel length, ratio between consonant length and vowel length (RCV)

Step 2: Selection of participant acoustic feature
This step was to find out that what acoustic features are most related to each primitive feature. First, values of correlation coefficients between acoustic features and those primitive features that have at least one correlation coefficient over 0.6 were considered significant and were chosen. There were a total of 16 acoustic features, which were briefly described above.

Next, multi-regression analysis was applied; because not all of the 16 acoustic features are related to every primitive feature, it was also necessary to reveal which acoustic features are more related to each primitive feature. The criterion is the absolute value of any acoustic feature that is higher than the average of absolute

[1] In English translation, it means "we bough a new car".

values of all the acoustic features. The selected acoustic features that were used to establish rules are listed in the third column of Table 1.

Step 3: Calculation of percentage variation

This step calculated the percentage variation of acoustic features against a primitive feature. For each primitive feature, 10 utterances with higher perceptual values of primitive features were selected, where perceptual values were rated by perceptual experiments in a previous work. In order to reduce bias in the data, utterances that have the highest differentiated values with the mean of 10 utterances were excluded. Based on the remaining 9 utterances, the percentage variation was calculated by subtracting values of acoustic features from the corresponding values of acoustic features of neutral utterances and then dividing the results by the values of the acoustic features of neutral utterances.

$$\frac{v\_AF - v\_N\_AF}{v\_N\_AF}$$

where v_AF is the value of acoustic features of a primitive feature within selected utterances and v_N_AF is the value of acoustic features of a primitive feature within its corresponding neutral utterance.

Based on the acoustic features selected in step 2 and the percentage variation calculated in step 3, rules of primitives feature were established; they are listed in the third column of Table 1.

3.2 Experiments

The resultant rules were then implemented by STRAIGHT and a perceptual experiment was also conducted. There were 17 voices morphed from one natural utterance, one for each of 17 primitive features. Eight subjects were asked to compare the morphed voices with a neutral voice and choose which voice could be most associated with each primitive feature. The accuracy rates are shown in the second column of Table 1. The result shows that it is possible to make rules for primitive features with our method, and also that the model built in the first phase was successful, because the establishment of rules is based on the results of experiments and acoustic analysis.

4. Conclusion and future work

For emotional speech, a three-layered model, consisting of emotional speech, primitive features, and acoustic features, was proposed in the previous work. A two-phase approach was taken to build the model. In the first phase, the model was built by a top-down method. In this paper, we report the work of the second phase, which was to verify the model by a bottom-up method. A set of rules for primitive features was built from acoustic features and implemented by morphing. They were used to verify the relationships of the model that was built in the first phase. A total of 16 acoustic features have been identified and measured for each primitive feature. For

each acoustic feature of a primitive feature, two types of parameters were calculated, percentage variations and ranges of value of acoustic features. The significance of this research is that the rules are designed in terms of primitive features instead of emotions. They can be used to verify the proposed model. The results of the perception experiments also show that the relationships built in the first phase are appropriate.

With regard to future work, we will first implement the rules in the morphing engine. Then we will conduct perception experiments to verify the reliability of the rules and refine the rules by experimental results.

Acknowledgement

This research is conducted as a program for the "21st Century COE Program" funded by the Ministry of Education, Culture, Sports, Science and Technology. We sincerely thank Fujitsu Laboratory for permission to use their voice database.

References

[1] Raskin, J.: Humane Interface: New Directions for Designing Interactive Systems. Addison-Wesley, Boston (2000).
[2] Picard, R. W.: Affective Computing. MIT, (2000).
[3] Massaro, D.W.: Perceiving Talking Faces. MIT, (1998).
[4] Tatham, M. & Morton, K.: Expression in Speech. Oxford University, (2004)
[5] Cahn, J. E.: Generating Expression in Synthesized Speech, Masters Thesis, MIT, (1989). http://www.media.mit.edu/~cahn/masters-thesis.html
[6] Murray, I. R., & Arnott, J. L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, JASA, 93, p. 1097-1108.
[7] Murray, I. R., & Arnott, J. L.: Implementation and testing of a system for producing emotion-by-rule in synthetic speech, Speech Communication, 16, p. 369-390 (year?).
[8] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S.,Enríquez, E., Aguilera, S., & Pardo, J. M.: Emotional speech synthesis: from speech database to TTS, ICSLP 98, Vol. 3, p. 923-926.
[9] Schröder, M.: Emotional speech synthesis: a review. Proc. Eurospeech 2001
[10] Vroomen, J., Collier, R., & Mozziconacci, S. J. L.: Duration and intonation in emotional speech, Eurospeech 93, Vol. 1, p. 577-580.
[11] Heuft, B., Portele, T., & Rauth, M.: Emotions in time domain synthesis, ICSLP 96.
[12] Edgington, M.: Investigating the limitations of concatenative synthesis, Eurospeech 97. (1997)
[13] Iida, A., Campbell, N., Higuchi, F. & Yasumura, M., A corpus-based speech synthesis system with emotion, Speech Communication, 40, p.161-187. (2003)
[14] Kawahara, H., Masuda-Katsusa, I. & de Cheveign'e, A.: Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds, Speech Communication, 27, p. 187-207, (1999).
[15] Matsui, H. & Kawahara, H.: Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system, Proc. Eurospeech'03, p.2110-16, (2003)

[16] Huang, C-F. & Akagi, M.: A multi-layer fuzzy logical model for emotional speech perception. Proc. EuroSpeech'2005 (accepted)

Table 1. Results of the perception experiment and rules of primitive features.

Primitive Feature	Accuracy Rate	Rules of Percentage Variation (%)
Bright	63%	AP:6.8971, PWR:8.961, PRAP:7.4674, F3:4.2405
Dark	100%	AP: -8.6915, RS1st: -74.618, PRAP:-15.541
High	100%	AP: 6.6863, HP: 8.0809, PWR: 9.327, RRAP: 9.3739, F3: 3.6476, RCV: -6.9915
Low	88%	AP: -9.9074, HP: -8.9125, PWR: -13.191, RRAP: -12.786, TL: 3.2754, RCV: -14.604
Strong	75%	AP: 6.2975, HP: 8.2056, RS: 45.65, PWR: 14.6, RRAP: 16.33, F2: -1.4528, H1A3: -51.204, RCV: 0.60943
Weak	75%	AP: -8.064, HP: -8.7019, PWR: -14.609, RRAP: -16.788, TL: 14.715, CL: 24.708, RCV: -11.72
Calm	75%	AP: -5.5793, RS:-27.871, PWR:-9.484, PRAP:-11.567, F2:-0.0916, TL:6.207
Unstable	50%	AP:6.4277, RS:43.568, PWR:12.953, PRAP:26.326, F2:-1.0208, H1A3:-55.649, RCV:2.095
Well-modulated	63%	AP:6.9762, HP:8.3081, RS:30.157, PWR:11.681, PRAP:20.318, F2:-0.052032, RCV:2.3622
Mono-tonous	75%	AP:2.3465, HP:2.502, RS:0.00046194, RCV:2.8654
Heavy	75%	AP:-9.4597, PWR:10.384, RS1st_A:2.556, PRAP:12.759, H1A3:15.605, RCV:-15.241
Clear	63%	AP:6.3306, RS:33.116, RHT:10.321, RS1st_A:2.2046, PRAP:5.326, RCV:-5.6828
Noisy	100%	AP:6.2975, HP:8.2056, RS:31.105, RS1st:20, PWR:25.486 , PRAP:24.698, F2:-1.4528, H1A3:-51.204
Quiet	63%	HP:-4, RS:-21.978, RS1st:-35.131, PWR:-14.609, F2:0.15395, F3:-1.4341, TL:10.715
Sharp	63%	AP:6.2975, HP:8.2056, RS:16, RS1st:11.003, PWR:13, PRAP:10, F2:-1.4528
Fast	75%	HP:-1.9229, RS:11.953, RS1st:-33.776, PRAP:-5.1078, F2:-1.5146, TL:-10.2257, CL:11.767
Slow	63%	AP:1.4565,F1:-1.7676, TL:12.4309, CL:-4.907, RCV:0.69847