2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)
Honolulu, Hawaii, USA, March 4-7, 2019

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

Zi Wang, Maori Kobayashi and Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 932-1292, Japan
Phone/FAX:+81-80-4024-2594
E-mail: {s1710034, maori-k, akagi}@jaist.ac.jp

Abstract

In order to investigate what acoustic features are important to emotional impressions and how those features relate to emotion perception, we interpolate voices from pairs of typical emotions with a morphing method, collect emotion scores on Arousal-Valence space by a listening test, and analyze how acoustic features relate to the evaluations. The results show that Arousal perception can be stably described by merely using fundamental frequency (F0). In contrast, although this research found that F0 and formants can fit Valence scores, how acoustic features correspond to Valence perception vary with different morphing references. Furthermore, the results show that modification rules of different formant components are necessary for the voice conversion system with better Valence control.

1. Introduction

The analysis and synthesis of emotional voices are interesting research directions. Moreover, a highly sophisticated emotional sound synthesis system can significantly improve the experience of human-computer communication. Some researches considered converting neutral speech to other categories of emotions using Gaussian mixture models (GMM) [1] or deep neural networks (DNN) [2]. Since human emotional voices are mild and not purely categorical, Xue *et al.* [3] proposed a rule-based system that is capable of converting neutral speech to emotional speech to continuous Valence and Arousal (V-A) scale. However, the system has problems in continuous emotion control, especially on Valence scale. There are two reasons for system defects. Firstly, the system is trained by categorical data, which is not continuously distributed on the V-A space and may distort the mapping rules between acoustic features and emotional impression. Secondly, although this study found the spectral sequence is crucial for Valence adjusting, they did not propose a spectral sequence modifying module. Because it is hard to find how the spectral sequence, include multiple formant components, relates to emotion perception using noncontinuous training data. As a result, this system cannot accurately adjust the Valence of voices, as the synthesized hot anger voices are perceived as happy, which two kinds of emotional voices with similar Arousal but different Valence.

This research aims to obtain emotional speech samples continuously spanned on the V-A space by interpolating voices from pairs of typical emotions using morphing techniques. Then we carry out a listening test to collect emotion evaluation of interpolated voices. Finally, we discuss what acoustic features are important to emotional impressions and how those features relate to emotion perception based on the analyses between the emotion evaluation scores and acoustic features of morphed voices.

2. Dimensional Emotion Representation

The dimensional representation method is aiming to map the emotion into a multi-dimensional space, and scale the emotion from low intensity to high intensity continuously. A common dimensional representation is a three dimensions space [4], which includes Arousal (excited-calm), Valence (positive-negative), and Dominance (powerful-weak) axes. Based on the database we have, this research uses Valence-Arousal space to represent positive-negative and excited-calm scales as shown in Fig. 1. On the V-A space, neutral is close to the origin; the 1st, 2nd, and 3rd quadrants of V-A space are corresponding to the happy, angry and sad emotions.

3. Methods

This section explains how we synthesize morphed voices and how to extract acoustic features of voice samples.

3.1 Synthesizing Morphing Voices

To synthesize morphed voices continuously distributed on the V-A space, this study interpolates voices from pairs of

Figure 1: Valence - Arousal space

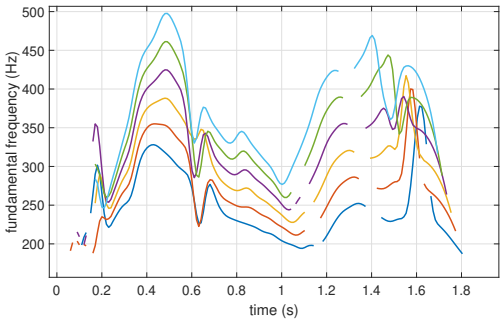


Figure 2: F0 of morphed voices (Neutral-Happy)

typical emotions using TANDEM STRAIGHT [5] and Morphing Toolbox [6], which modifies fundamental frequency (F0), aperiodicity, and STRAIGHT spectrogram respecting to the time and frequency coordinates of reference voices. For example, Fig. 2 illustrates F0 contours variations in Neutral-Happy morphed set. The morphed voice closed to neutral has a lower F0 contour, while F0 counter increases to a higher level as the voices get closer to happy emotion. Also, as morphed voices are gradually transformed from neutral to happy, the duration of voices is gradually shortened. Reference voices for the morphing process are chosen from the Fujitsu database, include neutral, happy, angry, and sad types of emotional voices.

3.2 Acoustic Feature Extraction

This research uses multiple ways to extract several acoustic features related to F0, power, duration, and formants. Also, considering that Japanese is generally regarded as a notion pitch-accent system and importance of accent components for emotion perception is emphasized by previous reports [3], [7], we separate samples according to the criteria of accentual phrases and obverse features in detail. Figure 3 shows an F0 contour and split accentual phrases of a voice sample. The followings are features we used in this study. Accentual

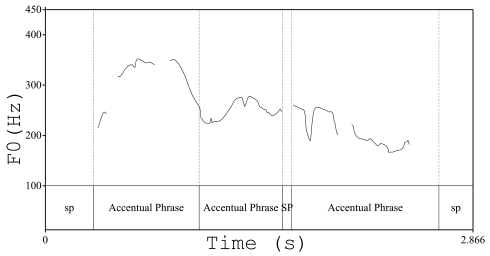


Figure 3: F0 contour and accentual phrases of a voice sample

phrases are split by manual segmentation; the other acoustic features are obtained by multiple estimation methods[5], [6], [8].

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

Further, we normalize features of each voice by the neutral reference which has the same content as follows:

$$F_{\mathrm{Normalized}} = \frac{F_{\mathrm{original}} - F_{\mathrm{Neutral}}}{F_{\mathrm{Neutral}}} \quad (1)$$

F_{Neutral} ,F_{original}, and $F_{\mathrm{Normalized}}$ mean the feature values of neutral reference voices, the feature values of normalizing targets, and the normalized feature values.

4. Listening Evaluation Experiment

To verify whether the Valence and Arousal scores of morphed voices using interpolated acoustic features can be perceived in a continuous way, a listening test was carried out to collect evaluation results of the morphed voices with Valence and Arousal. Ten Japanese listeners with normal-hearing, aged from 22 to 27, participated in the listening test of Valence and Arousal separately for 570 stimuli, including 40 reference and 530 morphed voices. The listeners were asked to evaluate Valence (-2: negative to 2: positive) and Arousal (-2: calm to 2: excite) of the stimuli by a graphic user interface (GUI) in a soundproof chamber.

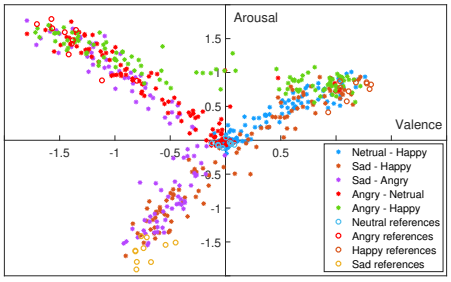


Figure 4: Evaluation scores of all stimuli

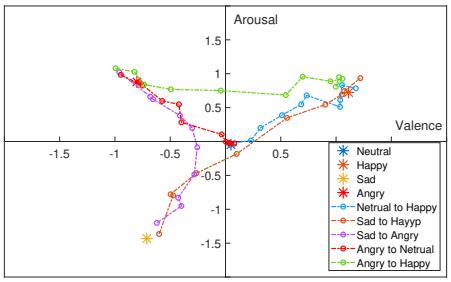
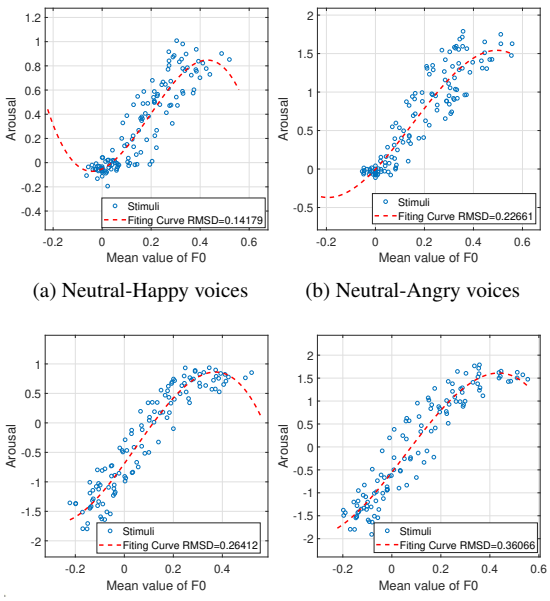


Figure 5: Evaluation scores of one group



(a) Neutral-Happy voices (b) Neutral-Angry voices

(c) Sad-Happy Voices (d) Sad-Angry voices

Figure 6: Fitting Arousal using AP feature

5. Result

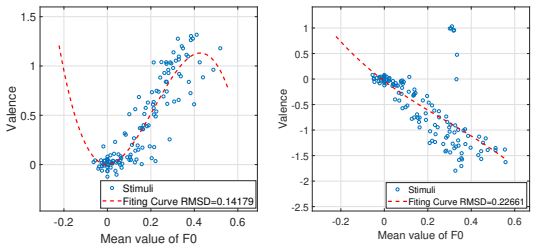
5.1 Evaluations of morphed voices

The mean values of all listeners' evaluated V-A scores are shown in Fig. 4, while Fig. 5 illustrates scores of a set of morphed voices with the same sentence. Those two figures suggest that the morphing techniques successfully synthesize voices between neutral to other emotional categories with continuous distribution. However, the morphed voices between Angry and Happy did not vary smoothly and suddenly changed from the second quadrant to the first quadrant, some morphed voices between Sad and Angry are evaluated close to neutral. Further, although evaluated V-A values do not change as equidistant as the acoustic features, those values generally maintain a monotonous change between reference voices.

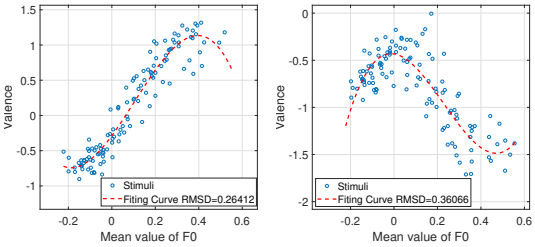
5.2 Analysis of Emotion Evaluations and Features

In order to examine which acoustic features influence on perceptions of emotional speech, we analyzed how acoustic features relate to those emotion evaluations. This research found that Arousal perception can be stably described by merely using F0, regardless of morphing references, like the mean value of F0 (AP). Fig. 6 shows that the feature AP can fit Arousal scale very well under 3-degree polynomial. The reason for using 3-degree polynomial is because the emotional scores remain stable near the stationary points of the cubic fitting function, where reference voices located, and changes rapidly between the stationary points. This pattern is identical to the shape characteristics of cubic functions and can get a relatively low root-mean-square deviation (RMSD). Since the Angry-Happy voices have almost the same scores on Arousal axis, which those green symbols and lines in Fig. 4 and 5, so this group cannot be well fitted.

However, relations between acoustic features and Valence perception vary in different morphing groups. Figure 7 shows how AP feature relates to Valence perception. For Neutral-Happy and Sad-Happy voices (left side), increasing F0 makes stimuli sound positive. In contrast, increasing F0 gives stimuli sound negative for Neutral-Angry (upper right). A remarkable phenomenon is for Sad-Angry voices (bottom right), a certain level of F0 gives stimuli a neutral feeling, but F0 above or below this level makes stimuli sound negative, this result explains why some morphed sounds between Sad-Happy are evaluated as neutral. Besides F0, more analyses illustrate that Valence perception is at least affected by F0 and formants features simultaneously, and not a single formant feature can stably describe Valence perception. Figure 8 indicates that the mean value of the third formant (AF3) can fit Valence well for Angry-Happy group (upper left) but the mean value of the first formant (AF1) cannot (upper right). However, the Valence scores of Neutral-Angry voices are more significantly influenced by feature AF1 rather than AF3 (lower side).

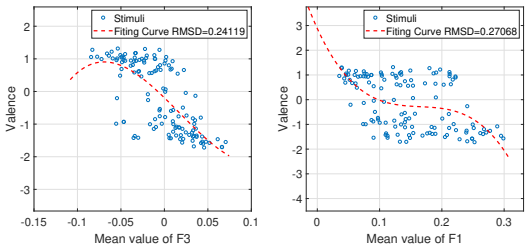


(a) Neutral-Happy voices (b) Neutral-Angry voices

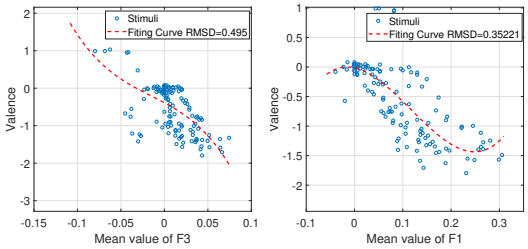


(c) Sad-Happy Voices (d) Sad-Angry voices

Figure 7: Fitting Valence using AP feature



(a) AF3 for Angry-Happy (b) AF1 for Angry-Happy



(c) AF3 for Neutral-Angry (d) AF1 for Neutral-Angry

Figure 8: Fitting Valence using different formant components

6. Conclusions

This research aims to obtain emotional speech samples with continuous distribution, then discuss what acoustic features are important to emotional impressions and how those features relate to emotion perception. From the results mentioned in section 5.1, we generated morphed voices with continuous distribution on the V-A space. From the results discussed in section 5.2, we ascertain that Arousal perception can be stably described by merely using F0 while Valence perception is at least affected by F0 and formants features simultaneously. This result explains why angry voices synthesized by Xue *et al.*'s system, without formats features modification, are evaluated as happy emotion. Also, considering the significances and corresponding relationship of different acoustic features varying in different morphing groups as we mentioned in section 5.2, we have hypothesized that acoustic features, like F0 or different formant components, impact Valence perception together in a non-uniform way on different areas of V-A space. Also, in order to get an emotional sound synthesis system with better Valence control, it is necessary to propose modification rules for different formant components separately depends on different areas of V-A space.

References

[1] R. Aihara, R. Takashima, T. Takiguchi, Y. Ariki: GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features, American Journal of Signal Processing, Vol. 2 No. 5, pp. 134-138, 2012.

[2] Z. Luo, T. Takiguchi, Y. Ariki: Emotional Voice Conversion Using Deep Neural Networks with MCC and F0 Features, Computer and Information Science, 2016 IEEE/ACIS 15th International Conference on. IEEE, pp. 1-5, 2016.

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

[4] H. Schlosberg: Three dimensions of emotion. Psychological Review, Vol. 61, No. 2, pp. 81-88, 1954

[5] H. Kawahara and M. Morise: Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework, Sadhana, Vol. 36, pp. 713-727, 2011

[6] H. Kawahara, T. Takahashi, M. Morise and H. Banno: Development of exploratory research tools based on TANDEM-STRAIGHT, Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, pp. 111-120, 2009

[7] H. Fujisaki: Information, prosody, and modeling-with emphasis on tonal features of speech, Speech Prosody 2004, International Conference, pp. 1-10 2004.

[8] L.R. Rabiner and R.W. Schafer: Theory and Applications of Digital Speech Processing, Upper Saddle River, NJ: Pearson, 2010.