Investigation of Emotionally Morphed Speech Perception and its Structure using a High Quality Speech Manipulation System

Hisami Matsui† *and Hideki Kawahara*†‡

†Faculty of Systems Engineering, Wakayama University, Japan
‡Human Information Science Laboratory, ATR, Japan
kawahara@sys.wakayama-u.ac.jp

10.21437/Eurospeech.2003-610

Abstract

A series of perceptual experiments using morphed emotional speech sounds was conducted. A high-quality speech modification procedure STRAIGHT [1] extended to enable auditory morphing[2] was used for providing CD quality test stimuli. The test results indicated that naturalness of morphed speech samples were comparable to natural speech samples and resynthesized samples without any modifications when interpolated. It also indicated that the proposed morphing procedure enables to provide stimulus continuum between different emotional expressions. Partial morphing tests were also conducted to evaluate relative contributions and interdependence between spectral, temporal and source parameters.

1. Introduction

An important issue in human computer interaction is versatile and high-quality control of emotional aspects of synthetic speech which needs effective understanding of physical correlates of non-linguistic and paralinguistic information. There are several review papers on emotional speech [3, 4]. However, yet no effective and high-quality control method has been established, partly due to methodological limitations as well as motivations of research. The most substantial problem was the possibility that perception of emotional aspects is fragile against degradations due to speech manipulations. Recent introduction of a very high quality auditory morphing method has changed this situation [2]. This method enables an exemplar-based strategy using ecologically valid (in other words naturally sounding) stimuli, in addition to conventional research strategies (for example, analytical and synthetic) for investigating this indicated problem.

The morphing procedure is implemented using a high-quality speech manipulation system STRAIGHT [1], that is based on a F0 adaptive and interference free time-frequency representation. The time-frequency representation is basically a smoothed version of the conventional sound spectrogram. it is straightforward to understand and control the morphing procedure taking advantage of the rich source of knowledge in speech perception and production.

In this paper, effects of auditory morphing of emotional speech are illustrated in terms of naturalness and trajectories in a perceptual space. Test results for investigating physical correlates of emotional attributes are also described using partially fixed auditory morphing stimuli. These are the first attempts on a research strategy called "systematic downgrading", that was proposed for investigating this difficult problem [2].

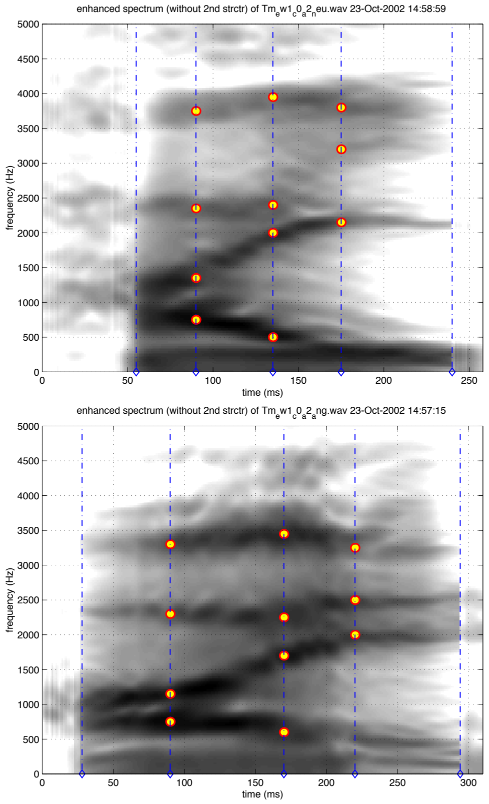

Figure 1: *Smooth time-frequency representations of a Japanese word uttered by a male actor under neutral (upper) and angry (lower) emotional conditions. Anchor points in the time-frequency domain are plotted as open circles and temporal anchors are plotted as vertical dash-dot lines.*

2. Morphing procedure

Morphing is a procedure to regenerate a signal from a representation on a trajectory between anchor points in an abstract parameter space. There are several technical issues on implementing a morphing procedure. Specifically, the coordinate system and the localized distance metric must reflect auditory perceptual characteristics, and the transformation must be as simple as possible. In this article, the time-frequency plane is the coordinate system. The transformation is represented as a simple piecewise bilinear transformation, because, unlike image morphing, the time-frequency coordinate is not isotropic. Practically, only up to 5 anchor points on a frequency axis at one temporal location and up to 4 temporal anchor points for one CV syllable are found to be sufficient. For fundamental frequency, it is relevant to morph the parameter in the log-frequency domain, because the F0 dynamics is represented in terms of a linear dynamical equation in the log-frequency domain [5]. For spectral density, morphing is computed using a dB representation, because it is a relevant approximations of intensity perception. The time-frequency periodicity index [6, 7] is also transformed by the same mapping function as the time-frequency representation.

Figure 1 shows the time-frequency representations of a Japanese word (/hai/ means “yes” in English) spoken by a professional male actor under two different emotional conditions (neutral and anger). It illustrates how anchor points are assigned. The morphed speech sound is generated using the morphed time-frequency representation and morphed F0 and source information.

3. Perceptual evaluations

A series of experiments were conducted initially to verify naturalness of morphed samples and secondly to investigate physical correlates of transformations of perceived emotional states.

3.1. Stimulus preparation

Professional actors (one male and one female) were payed for recording speech samples under seven different emotional expressions (neutral, anger, sadness, fear, happiness, surprize and disguise). Speech samples consist of isolated words, words with preceding contexts, words with following contexts and words in surrounding contexts. The target words were /hai/, /iie/ and /koNnitiwa/ (“yes”, “no” and “hello” in Japanese, respectively). Declarative and interrogative speaking styles were also recorded. Recording was done in a studio with assistance of professional recording engineers. An omnidirectional condenser microphone CU-41 (Sanken) was used and directly recorded digitally using 48 kHz 16bit linear PCM format. Isolated words spoken by a male actor were selected for morphing experiment, based on a screening test to verify that the intended emotional states are perceived correctly.

A preliminary test was conducted to determine the JND (just noticeable difference) of the morphing rate. Based on this test result, a morphing step size of 0.25 was employed in the following experiments.

3.2. Naturalness

Ten subjects (6 male, 4 female) participated in the naturalness evaluation test. Three test words (/hai/, /iie/ and /koNnitiwa/) spoken under four emotional conditions (neutral, anger, happiness, and sadness) were used. Morphing between four pairs of emotional conditions (neutral and anger, neutral and sadness, happiness and anger, sadness and happiness) were synthesized using morphing rate ranging from -0.25 to 1.25 with a step size of 0.25. The following additional stimuli were added to prevent response bias. Additional stimuli; original speech samples, resynthesized samples with 2F0 and 0.6F0 as resynthesis F0s, resynthesized samples with stretched (multiplied by 1.5) and compressed (multiplied by 0.,7) frequency axis. The total number of stimuli was 132 for each session.

Stimuli were randomized and each stimulus was presented twice in a session. Subjects were asked to evaluate naturalness of each stimulus in a 5 step rating (1: highly unnatural, 5: highly natural). Inter stimulus interval was set to 3 seconds.

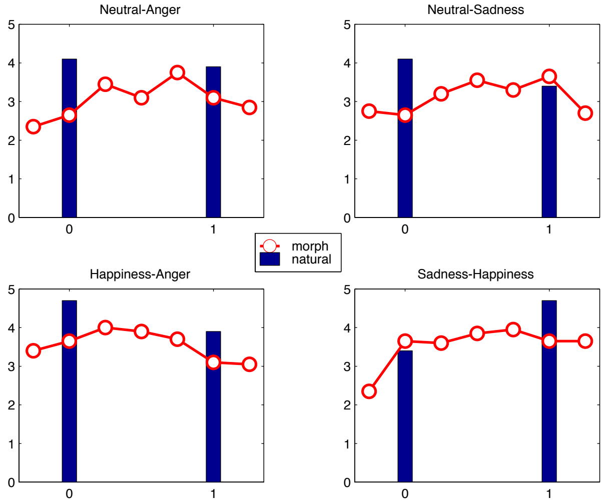

Figure 2: *Naturalness evaluation. Vertical axis represents averaged naturalness rating. Horizontal axis represents morphing rate. Bars represent results for the original speech sample. Circles with connecting lines represent result for morphed speech samples.*

Figure 2 shows results for the Japanese word /hai/. The vertical axis represents a simple numerical average of subjects’ ratings. It is interesting to observe that interpolated morphing sounds are evaluated often more natural than the simple resynthesized sounds. Test results for the other test words also showed a similar pattern. Tukey’s HSD test on the naturalness rating revealed that statistically significant difference in naturalness was only found between extrapolated speech samples and others. This indicates that by using interpolated stimuli only, ecological relevance is not violated.

3.3. Morphing trajectory

Ten subjects (6 male and 4 female) participated in an experiment to perceptually evaluate the emotional attributes of morphed speech samples. A Japanese word /hai/ spoken by a male speaker was selected as the test word. Morphing stimuli were prepared for six possible pairs between four emotional states (neutral, anger, happiness and sadness). Test stimuli also consisted of the twelve original speech samples (four emotional states for all three test words). Subjects were asked to evaluate six emotional attributes (anger, happiness, sadness, fear, surprize and disguise) in a 5 step rating procedure (0: No trace of the attribute was perceived. 4: The attribute was perceived very much.) for each stimulus. Inter stimulus interval was set to 8 seconds.

Principal component analysis was applied to the subjects’ ratings. First two components covered 88.1% of the total variation. Figure 3 shows morphing trajectories in the plane spanned by the first two principal components. The upper plot shows morphing trajectories made from pairs of stereotypical emotions. The lower plot shows morphing trajectories made from pairs of neutral and stereotypical emotions. They illustrated that morphing stimuli by interpolation provide monotonic trajectories between exemplar emotions. They also indicate that morphing stimuli by extrapolation do not extrapolate on this principal component plane.

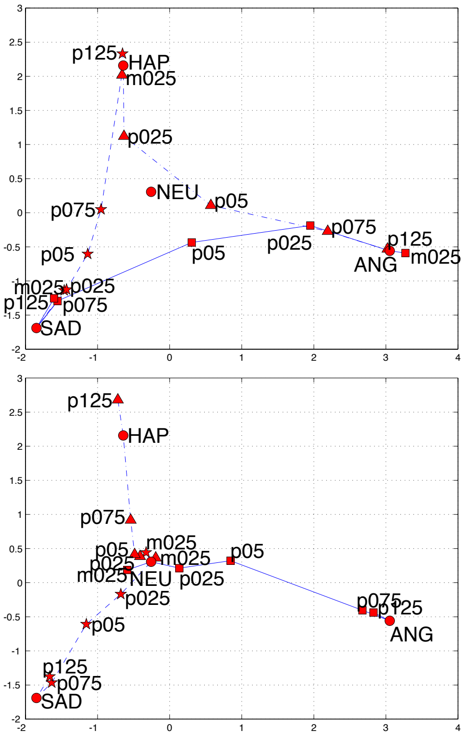

Figure 3: *Morphing trajectories in a perceptual space. Filled circles represent original emotional speech samples and the other marks represent morphed speech samples. Text annotations represent morphing rate. (m025 and p05 represent* -0.25 *and* 0.5 *respectively.)*

3.4. Partial morphing: naturalness

Morphing one or two parameters while keeping other parameters unchanged enables partial morphing. Ten subjects (6 male, 4 female) participated in this naturalness evaluation test. Morphing and partial morphing stimuli were synthesized for three pairs of emotional states; neutral and anger, neutral and happiness, and, neutral and sadness for a word /hai/. Morphing rates used were 0, 0.25 0.5 0.75 and 1. Only interpolating morphing samples were tested based on the previous test results. Four partial morphing conditions were tested; coordinate system morphing, coordinate system and F0 morphing, intensity morphing (morphing intensity of the time-frequency representation), intensity and F0 morphing. The test stimuli also consisted of four original speech samples. The experimental procedures were similar to the naturalness tests described in section 3.2.

Figure 4 shows the test results. Generally, naturalness of the partial morphing stimuli were less natural than that of the full morphing ones. However, partial morphing results with one parameter are not always lesser in naturalness than ones with two parameters.

3.5. Partial morphing: trajectory

Ten subjects (6 male, 4 female) participated in an experiment to test perception of emotional attributes of partially morphed speech samples. The same stimuli used in the naturalness test were also used. Experimental procedures were similar to the trajectory tests described in section 3.3. Principal component analysis was also employed to represent results on the principal component plane. Figure 5 shows the partial morphing trajectories. Partial morphing trajectories with one parameter do not reach the target emotional states, while partial morphing with two parameters results in close trajectories with those of fully morphed sounds.

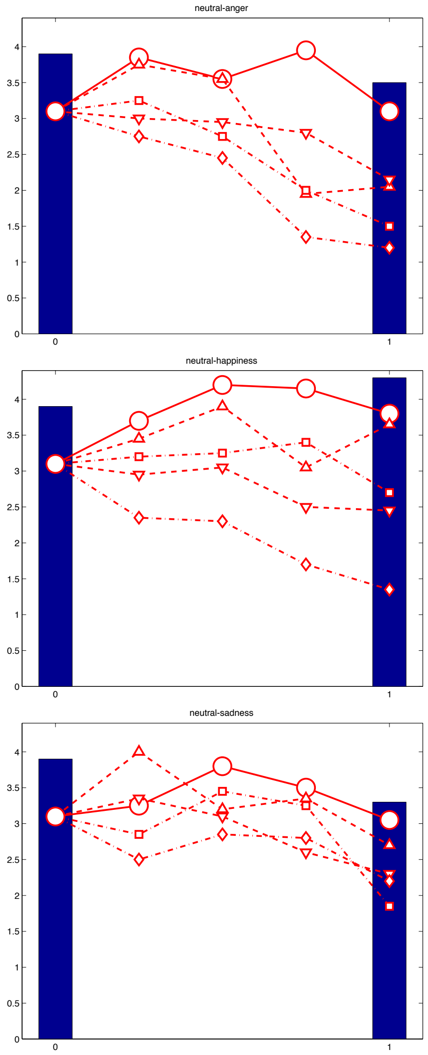

Figure 4: *Naturalness of partial morphing. Bar represents natural speech. Circle represents morphing with all parameters. Upright triangle represents coordinate system morphing. Inverted triangle represents morphing with coordinate system and F0. Diamond represents morphing with intensity. Square represents morphing with intensity and F0. Morphing continuum was made between neutral speech and three emotional speech samples (anger, happiness and sadness from top to bottom respectively)* .

4. Discussion

The morphing procedure used in this article is based on manual assignments of anchor points in the time-frequency representation which resembles conventional sound spectrogram. It is a contrasting difference from the other automatic auditory morphing procedures [8, 9]. This seemingly low-tech feature of the proposed procedure allows fine control of morphing charac-

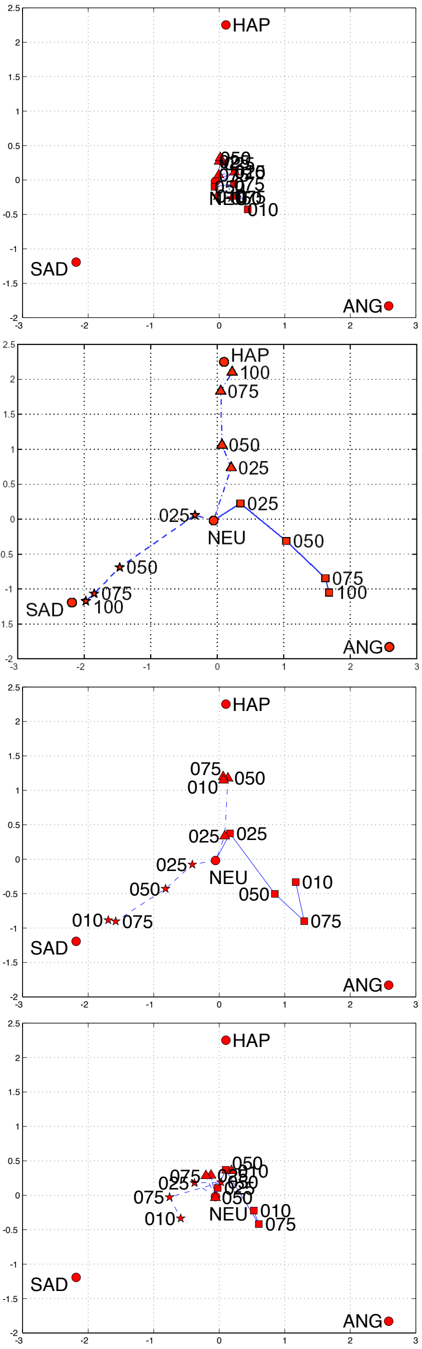

Figure 5: *Morphing trajectories in a perceptual space for partial morphing. Filled circles represent original emotional speech samples and other marks represent partially morphed speech samples. (Form top to bottom; coordinate, intensity and F0, coordinate and F0, intensity were morphed.)*

teristics and is highly desirable as a research tool.

The naturalness test results and the trajectories for partial morphing seems to disagree. This is because better trajectories by partial morphing with two parameters do not always provide higher naturalness. This may suggest that there exists a strong interrelation between F0, intensity and coordinate deformation. This is an interesting research topic.

5. Conclusions

A series of experiments were conducted to evaluate naturalness and perceived emotional state of morphed speech samples using a high-quality speech manipulation system STRAIGHT [1]. The test results indicated that naturalness of morphed speech samples were comparable to natural speech samples and resynthesized samples without any modifications when interpolated. It also indicated that the proposed morphing procedure enables stimulus continuum between different emotional expressions. The test results for partial morphing indicated that spectral intensity and F0 trajectories have relatively higher contributions on perceived emotional state than deformation in the time-frequency coordinate system. It is also important to note that partial morphing significantly degrades naturalness indicating that there is a strong constraint between spectral, temporal and source parameters.

6. Acknowledgements

This work was partly supported by CREST of JST and Grant-in-Aid for Scientific Research (A) 12040241 and (B) 14380165.

7. References

[1] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[2] Hideki Kawahara and Hisami Matsui, “Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,” in *Proc. ICASSP*, 2003, [in print].

[3] I.R. Murray and J.L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[4] M. Schroeder, “Emotional speech synthesis: A review,” in *Proc. EUROSPEECH Scandinavia*, 2001, pp. 561–564.

[5] H. Fujisaki, “A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour,” in *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, O. Fujimura, Ed., New York, 1998, pp. 347–355, Raven Press.

[6] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” in *Proc. Eurospeech’99*, 1999, vol. 6, pp. 2781–2784.

[7] Hideki Kawahara, Jo Estill, and Osamu Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight,” in *Proc. 2nd MAVEBA*, Firenze, Italy, 2001, [CD ROM].

[8] M. Slaney and B. Lassiter, “Automatic auditory morphing,” in *Proc. ICASSP’96*, 1996, vol. 2, pp. 1001–1004.

[9] H. Banno, K. Takeda, K. Shikano, and F. Itakura, “Speech morphing by independent interpolation of speech envelope and source excitation,” *J. of IEICEJ*, vol. J81-A, no. 2, pp. 261–268, 1998, [In Japanese].