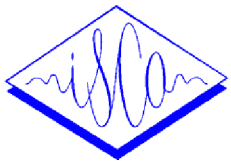
ISCA Archive
http://www.isca-speech.org/archive



7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

High Level Emotional Speech Morphing Using STRAIGHT

Dong-Yan Huang, Susanto Rahardja and Ee Ping Ong

Institute for Infocomm Research, A*STAR
1, Fusionopolis Way
#21 − 01 Connexis
Singapore 138632
{huang, rsusanto, epong@i2r.a-star.edu.sg}

Abstract

This paper presents high-level strategies for controlling emotional speech morphing algorithms. Emotion morphing is realized by representing the acoustic features in their time-frequency plan that is warped and modified to generate natural morphed emotional speech. These acoustic features are desirable to be decomposed into multidimensional space and to be orthogonal. After matching these acoustic features of speech, a morph smoothly interpolates their variations not only in time domain but also their amplitudes in frequency domain to describe a new emotional speech in the same perceptual space. Finally, these descriptors are synthesized to produce morphed speech waveform. This paper describes representations of acoustic features, techniques for matching, and algorithms for interpolating and morphing acoustic features such as duration, spectral envelope and pitch contour using STRAIGHT [1] as an example. The subjective listen test will be showed for emotional speech morphing of which the quality and naturalness were comparable to natural speech samples.

Index Terms: emotional speech morphing, acoustic features, warping, matching, interpolation.

1. Introduction

Audio morphing is well known in the Computer Music community [2, 3, 4, 5, 6]. The core process in an audio morphing system is to preserve the shared characteristics of the starting and final signals, while generating a smooth new sound with an intermediate timbre. Most of these methods are based on the interpolation of sound parameterizations resulting from analysis/synthesis techniques, such as the Short-time Fourier Transform (STFT), Linear Predictive Coding (LPC) or Sinusoidal Models synthesis (SMS).

In human computer interaction, high quality control of emotional aspects of synthetic speech is an important issue. Because of high cost of emotional database collection, most of the universities and research centers collect only a classic set of basic emotions like neutral, anger, happiness, sadness, disguise, surprise, and fear [7, 8, 9]. However, in daily life, spontaneous speech includes all real emotions. All continuum emotional states are desirable to be stimulated between any different basic emotional states. Morphing is one of the feasible ways to provide the continuum emotional state. Recently, high-quality speech manipulation system STRAIGHT has been proposed for emotion morphing with manual placement of anchor points in the reference and the target speech representation [10, 11]. Therefore, there are two goals to be realized in this paper. One goal is to realize an automatic emotion morphing. The second goal is to investigate the manipulation of high-level features accordingly for emotion morphing. The term ”high-level” refers to any feature which does not share close, obvious relationship to samples of waveform. These features includes pitch contour, spectrum, formants, ...etc.

To realize ”high-level” emotional speech morphing, a speech model based on analysis-by-synthesis method is necessary and is able to separate a sound into salient feature dimensions, where each dimension represents a unique quality of a speech. Hence, the transformations are limited along these specific dimensions with a high preciseness of control to the perceptual quality of the final morphed signal.

In our preliminary study, high-quality speech manipulation system STRAIGHT can deliver the best quality of synthesized speech among the analysis-by-synthesis methods such as Harmonic-plus-Noise Model (HNM) [12], Linear Predictive Coding (LPC) or ARX-LF [13]. The problem for the latter three algorithms is that the degradations are considerable due to large speech manipulations. Therefore, the automatic morphing procedure is implemented by using STRAIGHT [1], which represents speech using a F0 adaptive and interference free time-frequency spectrum, an instantaneous-frequency-based F0 and ”high-level” features morphing and interpolation.

The paper is organized as follows: Section 2 describes a control structure for high-level features during the evolution of the morph and explains which features are selected for morphing. Section 3 uses STRAIGHT to extract acoustic features such as fundamental frequency (F0) contour, spectrum, aperiodicity index, and formants, how to set break points in time and frequency domains. Section 4 focus on the specifics of temporal alignment and time warping. Also, techniques for matching, interpolating and morphing are described. Section 5 presents the results of informal listening tests. Finally, the future work will be pointed out.

2. Emotional Speech Morphing System

Figure 1 shows the general block diagram of the emotional morphing system. The analysis/synthesis technique is STRAIGHT [1]. Each input is decomposed by STRAIGHT into three high-level acoustic features - F0 contour, spectral envelope, and aperiodicity index (AP). The F0 contour information can be used to detect voiced/unvoiced break points and the duration of each voiced or unvoiced part. Then all the features are dynamically aligned along the time dimension by interpolating with similarly derived features from the corresponding input to create an overall control structure. Then using this high-level control structure, all the features are warped in time using the

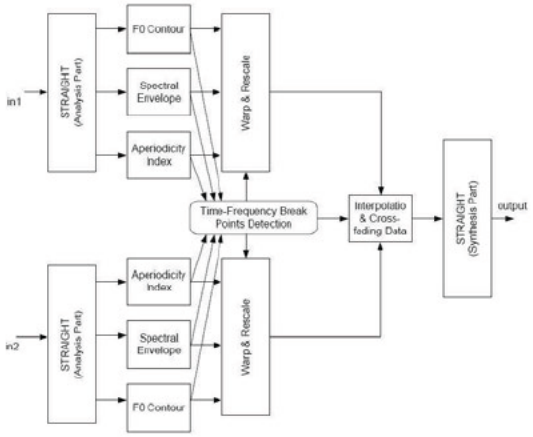


Figure 1: Block diagram of system

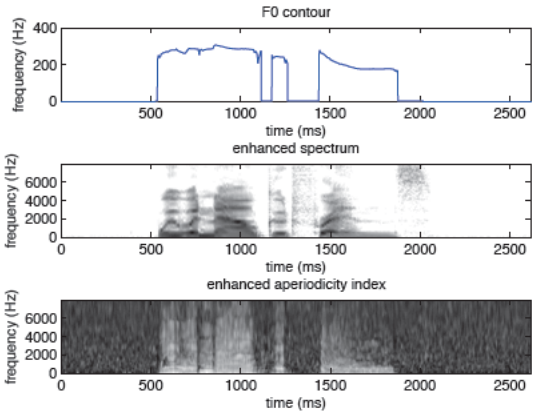


Figure 2: Upper panel: fundamental frequency trajectory; middle panel: enhanced spectrum; bottom panel: aperiodicity index of neutral speech 'we are good friend'

information of break points, and pre-processed to re-scale the values along a particular dimension. The voiced spectrum and envelop are interpolated based on breakpoints of frequency. The processed high-level features from each input is finally interpolated to average out any deviations or variations left in the data, and which give unique emotional characteristics.

The process requires managing a large amount of data, yet provides a flexibility in the type of control. High-level features from one input may be used to govern entirely the evolution of the other speech. There are many possibilities to generate morphed output. The resulting morphed output is, in general, intuitively expected depending on the features used, and thus simple and instinctive for any user to use.

3. Feature Extraction and Time-Frequency Break Point Detection

Extracting features aims at trying to make sense of the speech sequence in a holistic manner, the similar way that a human does. The challenge is that the computer is not able to examine the data holistically to discern pitch or timbre, whereas by contrast, a human listener can easily identify these features. We must also make contingencies to deal with the possibility that a particular feature may not always be present, which depends on the algorithm behave.

Generally, only four distinguishing characteristics are used in a sound's description. They are: pitch, intensity, timbre, and duration. The timbre is perceptual measure including spectrum and envelope.

3.1. F0 contour, spectrum and aperiodicity index

The analysis/synthesis system STRAIGHT can extract pitch, aperiodicity index and a smoothed version of the conventional sound spectrogram frame by frame. However, the characteristic relevant to spectral component such as formants is not extracted. This feature has more of a pronounced effect of the sound. The morphing of this feature results in perceptual impact for our morphed speech. As STRAIGHT represents the pitch contour, aperiodicity index, and smoothed spectrogram in time-frequency domain in a synchronized way, there are no occurrence of conflicting warping, alignment or pre-processing direction. The feature like formants is extracted only from voiced part of speech. The voiced/unvoiced detection should be done prior to extraction of formants. Figure 3 shows fundamental frequency trajectory, spectrum and aperiodicity index of neutral speech "we are good friend" extracted using STRAIGHT.

3.1.1. Detecting voiced/unvoiced segments

Since STRAIGHT is able to extract F0 contour, we use the difference between the current pitch value and the previous pitch value to detect where the voiced section occurs.

$$DF0_i = F0_i - F0_{i-1} \tag{1}$$

The voiced section can be described as two series of breakpoints with positive and negative values by seeking local maxima and local minima, respectively

$$F0_+ = \{(t_{p,1}, DF0_{p,1}), (t_{p,2}, DF0_{p,2}), \cdots, (t_{p,i}, DF0_{p,i})\} \tag{2}$$

and

$$F0_- = \{(t_{n,1}, DF0_{n,1})), (t_{n,2}, DF0_{n,2}, \cdots, (t_{n,i}, DF0_{n,i})\} \tag{3}$$

where $t_{p,i}/t_{n,i}$ is the location of local maxima/minima - break point of time and $DF0_{p,i}/DF0_{n,i}$ is the corresponding value of local maxima/minima at ith frame, respectively.

3.1.2. Duration estimation of voiced/unvoiced segments

The durations of the start section and the end section of sentence can be determined based on the information of voiced/unvoiced detection.

The duration of voiced section in the middle of the sentence can be estimated by

$$Dur_{v,i} = |t_{p,i} - t_{n,i}| \tag{4}$$

and the unvoiced section can be determined by

$$Dur_{uv,i} = |t_{p,i+1} - t_{n,i}| \tag{5}$$

The unvoiced section can be further estimated the duration of pause and of unvoiced consonant via a pre-defined threshold. Figure 3 shows an example of the break points of time and durations of voiced/unvoiced of neutral and angry speech "we are good friend" after the outliers and errant values are removed.

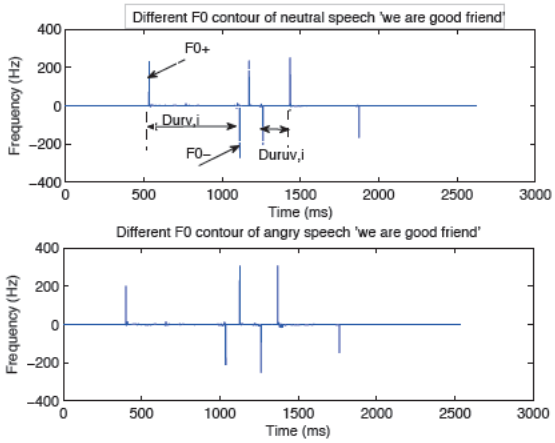


Figure 3: Upper panel: different fundamental frequency of neutral speech; bottom panel: different fundamental frequency of angry speech 'we are good friend'

3.2. Formant extraction

As high-level features extracted by STRAIGHT are all in time-frequency representation, the voiced/unvoiced detection allows us to align all acoustic features of two inputs in time domain. However, the characteristics of sound depends on the spectral shape. In speech morphing, as modifications to spectral envelope of a speech take place, one way to maintain some measure of coherence of two sounds is to maintain the similarity of the spectral shape. Human listener can easily determine the changes in the voiced part of speech. We extract the spectral envelope only from voiced parts and then modify them. The envelope can be represented as a series of breakpoints. In frequency domain, each breakpoint is determined by the amplitude and frequency of each component. This is a very tedious representation, given by the set

$$S = \{(f_1, a_1), (f_2, a_2), \cdots, (f_i, a_i)\} \tag{6}$$

where f_i and a_i are the frequency and amplitude, respectively, of the $i$the component. The formant can be determined by the local maximum for each given section of the spectrum. If the spectrum is divided into k segments stretching from 0 Hz to $f_s/2$ Hz, and let I_k denote the set of within each segment, then

$$S = \{\max(I_1), \max(I_2), ..., \max(I_k)\} \tag{7}$$

where $\max(I_k)$ denotes the component with the maximum amplitude for the kth segment. The number of segments used will obviously determine the accuracy with which the spectral shape is represented. We choose 24 segments to give rough approximation.

4. Interpolation and Morphing

The morphing can be easier understood through face image morphing, in which the in-between images all show one face smoothly changing its shape and texture until it turns into the target face. Similar to image morphing, one speech signal should smoothly change into another, keeping the shared characteristics of the starting change and ending signals but smoothly changing the other properties. The major properties of concern as far as a speech is concerned are its pitch and spectral envelope information.

The core process in a voice morphing system is the transformation of the spectral envelope and the prosodic parameters of the source speaker to match those of the target speaker and linear transformations estimated from time-aligned parallel training data are commonly used to achieve this. Emotional speech morphing system shares not only the characteristic of spectral envelop but also the non-linguistic and para-linguistic information. Once the input data has been analyzed and modeled, and all relevant features have been extracted, we are able to start interpolating the data to generate an intermediate, morphed output.

The component of the morphing system is conceptually straightforward. The high-level features are used for morphing. Our morphing system includes the following components: 1) Weighting function; 2) Temporal aspects of acoustic feature alignment; 3) Rescaling; 4) Matching; 5) Feature smoothing; 6) Interpolation; 7) Resynthesis.

4.1. Weighting function

The weighting function α is key component that unifies all elements for morph. It is used to dictate how much or how little of each input should be present in the output morph, and varies with respect to time. This weighting function comprised of a series of breakpoints that simply describe the evolution of the morph, and the weight of each input. It is bounded from 0 (representing the first input) to 1 (the second input). That is, if α is equal to zero, then the output will consist entirely of the first input or parameter; likewise, if α is one, then only second input or parameter will emerge.

Prior to any interpolation whatsoever, though, all the high-level features of two inputs should be first warped and aligned. For these two inputs, each frame must remain in correspondence with it's counterpart in the other signal. We could add, repeat, or move frames in order to stretch or contract the time axis. Complex interpolation schemes could be employed to facilitate this, however, this would quickly become unmanageable and confusing. Thus, the data will be warped first to ensure that onsets and durations align, and then resampled at equal intervals to ensure that frames correspond.

4.2. Temporal interpolation

In most of high-level morphing systems, extracted high-level features will be used to warp or contort low-level data along the specified feature-dimension, after which the low-level data from both inputs will be combined and interpolated to average-out any deviations that may be present. This top-down approach first utilizes the main features as specified by the user, and then focus on fitting the low-level data into place in accordance with what the interpolated feature dictates. However, our algorithm uses the high-level features for alignment, interpolation, matching, morphing and re-synthesis to generate the morphed speech.

4.2.1. Temporal alignment

Time alignment is general employed so that onsets of each input sound occur simultaneously. Most of the speech or audio morphing system employs time warping to two low-level input data for voiced/unvoiced segments. We must take care to ensure that stretching or compressing a speech does not smear or distort the segments, otherwise it's use as a defining characteristic in the identification of timbre will be severely undermined. The dynamic programming method is firstly used for aligning the voiced/unvoiced segments of two inputs and time warping.

Then the morphed speech is synthesized using STRAIGHT. The quality of morphed speech are very poor. It may be due to poor onsets and attacks detection by the method used and the characteristic in the identification of timbre is severely damaged. As high-level features extracted by STRAIGHT are all synchronized in time-frequency plane, the information of when pitched and un-pitched segments occur is employed to aid in aligning attacks and other inharmonic sections, and also to equalize the duration of features in time-frequency plane. This time axis refers to the frame instance. As the human ear can detect differences of about 5 ms, however, as long as each onset coincides, large changes in time, which affect intermediary data, are acceptable. In time-frequency plane, the detection of all onsets from both inputs is not as strict as detection of onsets of low-level data in time domain, which generally requires precise time instances of onsets.

The time breakpoints of two inputs can be obtained by Eqs 2 and 3. For simplicity, we can align the first onset of two inputs by adding the first a few frames for the input, of which the first onset occurs earlier than another input. A set of target times is generated by interpolating onsets from both inputs in time-frequency plane, for onsets falling within a user-set limit. For this, the weighting function, $\alpha(t)$ is employed:

$$t_\alpha = \alpha(t)(t_2 - t_1) + t_1 \tag{8}$$

where t_1 and t_2 are the detected onset times, and t_α is the interpolated time. Then the data of all features between $\alpha(t)(t_2 - t_1)$ are re-sampled at regularly spaced intervals. This is ideal time warping process by enabling temporally-equivalent events, or frames.

4.2.2. Rescaling

As high-level features in time-frequency domain are used for morphing, we may warp each feature in the time axis to ensure that the features are properly aligned. Then the corresponding features of two inputs, and the new, interpolated value are used to determine the amount of scaling required for the morphed features.

The extracted pitch of two inputs are properly aligned. The distance from the original breakpoints to the new breakpoints must be determined by Eq. 8, and its pitch scaled by the value of the envelope, which is calculated linearly between points as shown:

$$F0 = (\frac{F0_2 - F0_1}{t_2 - t_1}) \cdot (t - t_1) + F0_1 \tag{9}$$

where $F0_2$ and $F0_1$ are the values of the pitch contour for segments t_2 and t_1, and $t - t_1$ is the distance in time axe from the breakpoint t_1. In general, the matrix sizes of spectrograms of two input are different. If the length of spectrogram is less than that of another one, then the spectrogram will be contracted and there will be a discontinuity in the higher time range. To remedy this, the last frame in the spectrum is simply repeated. Next, we could add frames for the stable part of spectrogram. These frames must be re-calculated according to what the new vector dictates at that time point. Specifically, distance of a frame from breakpoint must be determined, and it's amplitude scaled by the value of the twm frames, which is calculated linearly between points as shown:

$$S = (\frac{S_2 - S_1}{t_2 - t_1}) \cdot (t - t_1) + S_1 \tag{10}$$

where S_2 and S_1 are the vectors of the spectrograms of two inputs for time breakpoints t_2 and t_1, and $t - t_1$ is the distance in time axis from the breakpoint t_1.

According to our experience, for two ends of features of the morphed speech, the features at the two ends of original speech are repeated for the morphed features. The pitch contour for the stable pitch contour part is replaced by the value of the pitch contour at the frames in question determined by Eq. 9. The added frame vector of extended spectrogram are replaced by the vector at the time breakpoint t determined by Eq. 10.

4.3. Interpolation in frequency domain

4.3.1. Matching

For morphing high-level features such like spectral and aperiodicity index, matching is a pre-interpolation stage used to form correspondences between harmonics in each input so that the final stage will be relatively straightforward. In our system, the matching method involves looking for the harmonics closest in frequency, harmonic pairings using a morph-matrix and seeking the best matching harmonics frame-by-frame.

Since the best matches is determined on a frame-by-frame basis, that might cause these correspondences to change radically from one instant to the next, causing harmonics to "jump" and be interpolated with differing components at each instant in time. This would have a detrimental effect on the sound, and leads to the instability of the harmonic components. Thus, some degree of control is required in order to ensure that the harmonics are not changing allegiances between frames.

Sliding window algorithm can be used to solve this problem. A single harmonic is considered for three frames at a time. If the same harmonic pairing of the first frame emerges in either the second and third frame, the algorithm sets all frames to share that same pairing if it is possible. Then, the window slides one frame forward and three new frames are taken under consideration, and the process repeats. Hence, the algorithm tries to establish a consistency with each initial match in that every harmonic will be paired with it's original match for as long as possible.

4.3.2. Interpolation for formants

As the characteristics of sound depends on formants, the interpolation of formants will affect the quality of morphed speech.

For the formants, it is desirable to identify and maintain formant information. For this, all peaks in each envelope must be identified by determining maxima. Next, any detected peaks residing nearby in frequency shall be interpolated. This could be accomplished by employing quadratic interpolation to determine the amplitude and frequency location of each peak, and then, with the weighting function, determining where the intermediate formant should reside. For the parabola $y = ax^2 + bx + c$ used to describe a shape of formant, each coefficient may be derived by:

$$\begin{aligned} c &= S_i \\ a &= 0.5 \cdot (S_{i-1} + S_{i+1} - c) \\ b &= 0.5 \cdot (S_{i-1} - S_{i+1}) \end{aligned} \tag{11}$$

where S is the amplitude at the frequency location of the detected maximum, i. As the spectrum has been divided into 24 segments for each formant, i will vary from between 1 and 24. Next, where the peak's maximum occurs (relative to i) is determined:

$$x_{max} = \frac{b}{-2.0a} + i \tag{12}$$

The two extracted formants are interpolated using α, and the new formants is stored as a series of spectral breakpoints. Specifically, the location of the new maximum, its bandwidth, and the values of neighboring points are derived. This process repeats for any formants that fall within a certain frequency range of each other.

For the remainder of the breakpoints left uninterpolated, a simple one-to-one cross-fade is sufficient. The mix of each input's envelope is governed by the weighting function and is determined linearly, as shown

$$\begin{aligned} S_n &= (1-\alpha)S_{1,n} + \alpha S_{2,n} \\ &\text{for all remaining } S_{i,n} \in S_i, \text{for } i = 1, 2 \end{aligned} \tag{13}$$

for each set of breakpoints K remaining in S. This results in an intermediate spectral shape, which can now be applied to the amplitudes of all harmonics.

4.4. Feature smoothing

The pitch detection algorithm, while robust, still introduces errant values and outliers. Therefore, the interpolated pitch also reflects these values, which can have a damaging effect when used to re-synthesis. Two approaches have been used to deal with this problem. One approach is used to remove outliers of F0 using the confidence rating to decide whether that pitch value should be used or not. Another approach is to apply a median filter to each input's extracted pitch to smear out sharp discontinuities of data, and to approximately follow polynomials. This aids in removing errant values, and can also help in providing a stable evolution of F0 contour:

$$m_n = median(F0_{n-\frac{N-1}{2}}, \cdots, F0_{n+\frac{N-1}{2}}) \tag{14}$$

where m is the median pitch value, and $N+1$ is the number of points used in filtering and is usually odd. In our algorithm, 3 to 5 points were used in the filtering.

For spectral envelop, the formants are re-scaled for the voiced segment. In order to prevent any detrimental effects on the data. A moving average (low-pass) filter is used to smooth the corresponding feature segments:

$$\begin{aligned} S_n &= \tfrac{1}{N}(S_{n-\frac{N}{2}} + \cdots + S_{n+\frac{N}{2}-1}) \quad \text{for } N \text{ even} \\ S_n &= \tfrac{1}{N}(S_{n-\frac{N-1}{2}} + \cdots + S_{n+\frac{N-1}{2}}) \quad \text{for } N \text{ odd} \end{aligned} \tag{15}$$

where N is the number of values used in the smoothing process. Usually, good results are obtained using 3 to 5 points. The feature smoothing is to remove the discontinuities of data because any resulting discontinuities are noticeable, and the quality of the overall output suffers.

4.5. Morphing

All the spectral data and aperiodicity data have been formatted, analyzed, manipulated, and is in a state where we can finally start performing the actual interpolation. Ironically, at this point, most of the work in the morphing process has already been done. All that remains is to interpolate between the high-level data comprising the two signals. Thanks to the matching process, we know which harmonics need to be interpolated.

There will be one morphed harmonic generated for every two contributing harmonics, where the contributing harmonics are corresponding pairs as determined in the matching stage. Each morphed harmonic will contain the interpolated values of the contributing harmonics, and will be what are actually synthesized in the end. To interpolate between two harmonics, the weighting function is again employed. For high-level features like spectral envelop and aperiodicity index, the amplitude and frequency must be interpolated according to its own metric. Amplitude varies according to a logarithmic scale in regards to perception; however, the amplitude as it is represented within is always linear. Therefore, a simple weighting such as:

$$S_{morphed,f,t} = (1-\alpha(t))S_{1,f,t} + \alpha(t)S_{2,f,t} \tag{16}$$

The spectral envelop, aperiodicity index and their frequency are interpolated by Eqs. 16 and 17, respectively.

The frequency of each harmonic is easy and straightforward to interpolate. In the same vein as pitch was interpolated, the frequency is determined by:

$$f_{morphed} = f_1(\frac{f_2}{f_1})^{\alpha(t)} \tag{17}$$

where α is the value of the weighting function, and f is the frequency value from each contribution harmonic.

To interpolate between both input's F0 contour, the following formula is used:

$$F0_{morphed} = F0_1(\frac{F0_2}{F0_1})^{\alpha(t)} \tag{18}$$

where $\alpha(t)$ is the value of the weighting function, $F0_1$ and $F0_2$ are the pitch values from each input, The interpolated pitch and spectral centroid will be reflected in the final, morphed output.

4.6. Resynthesis

The final step is to re-synthesize all the interpolated high-level features to generate the morphed sound. The quality of the results is discussed in the next section. Figure 4 shows time-frequency representations of F0 contour, smooth spectrogram of neutral, angry and emotionally morphed speech for the word ("Hello"). The morphed speech was generated by decomposing two inputs into smooth spectrogram, pitch spectrogram and aperiodicity index. The pitch information is used to align these three acoustic features in time axis. The frequency breakpoints are used to interpolated new formant. All the aligned acoustic features are cross-faded. In the results shown in Figure 4, the smooth spectrograms are cross-faded. The interpolated spectrograms are combined and re-synthesize to generate the morphed sounds.

5. Experimental Results

Experiments were carried out to evaluate the naturalness and to investigate the acoustic correlation of the transformation of perceived emotional states.

5.1. Stimulus preparation

Institute for Infocomm Research (I^2R) collected emotional data acted by professional actors under seven different emotional expression (neutral, anger, sadness, fear, happiness, surprise and disgust). There are isolated words, numbers and sentences in speech samples. Recording was done in I^2R audio recording studio with assistance of professional recording engineers. An omnidirectional condenser microphone U 87 Ai (Neumann) was used and recorded using 48 kHz 16 bit linear PCM format. Isolated words spoken by a male actor were selected and resampled to 8 kHz for morphing experiment , based on a subjective

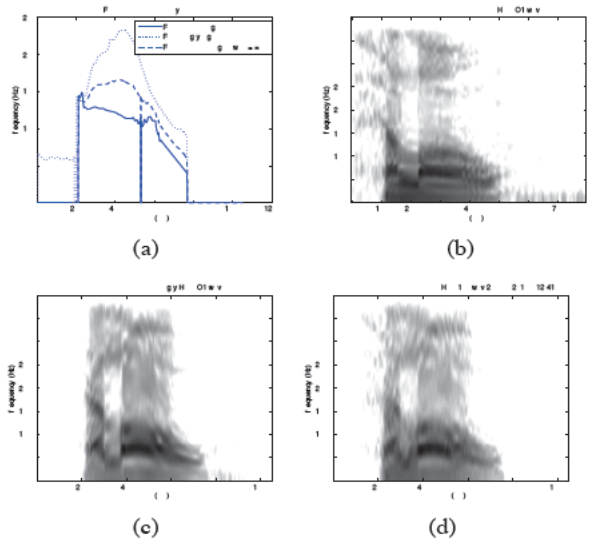


Figure 4: The word "Hello" is shown in time-frequency representations of acoustic features. a) F0 contour of neutral speech, angry speech and morphed speech; b) Spectrogram of neutral speech; c) Spectrogram of angry speech; d) Spectrogram of morphed speech with $\alpha = 0.5$

Table 1: Naturalness Evaluation for emotionally morphed speech

Emotions	Naturalness
Neutral → Anger	3.3
Neutral → Happiness	3.5
Neutral → Fear	3.3

listening test to verify that the intended emotional states are perceived correctly. A morphing step size of 0.1 was employed in the following experiments.

5.2. Naturalness

Nine subjects (5 male, 4 female) evaluated the naturalness of emotionally morphed speech. Two test words (/east/ and /hello/) spoken under four emotional conditions (neutral, anger, happiness, fear) were used. The morphing speech between three pairs of emotional conditions (neutral and anger, neutral and happiness, neutral and sadness) were synthesized using morphing rate ranging from 0 to 1 with a step of 0.1.

The synthesized speech were randomized and each stimulus was presented twice in a session. Subjects were ask to rate the naturalness using the scales $1 \sim 5$. The rating of 1 is the worst and 5 is the best. Table 1 shows the rating of naturalness for all morphed emotional stimuli.

6. Conclusion

In this paper, we developed an automatic high-level feature morphing technique for emotional speech morphing using high-quality speech manipulation system STRAIGHT. The subjective listening test showed that the naturalness of morphed speech were comparable to natural speech samples. We shall further study the role of the temporal, spectral and source parameters and their interaction for relative contributions to emotions and other emotion morphing in the future.

7. References

[1] H. Kawahara. "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pages 1303-1306, Munich, Germany, 1997.

[2] Serra, X., "Sound hybridization techniques based on a deterministic plus stochastic decomposition model", Proceedings of the ICMC 1994.

[3] Tellman, E., L. Haken, B. Holloway, "Timbre Morphing of Sounds with Unequal Number of Features", J. Audio Eng. Soc., 43:9 1995.

[4] Osaka, N., "Timbre Interpolation of sounds using a sinusoidal model", Proceedings of the ICMC 1995.

[5] Slaney, M., M. Covell, B. Lassiter. 1996. Automatic audio morphing, Proc. IEEE Int. Conf. Acosut. Speech Signal Process. 2, 1001-1004 (1996).

[6] Settel, Z., C. Lippe, Real-Time Audio Morphing, 7th International Symposium on Electronic Art, 1996.

[7] Dellaert, F., Polzin, T., Waibel, A., Recognizing emotion in speech, In: Proceedings of ICSLP, Philadelpha, USA, 1996.

[8] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., A database of German emotional speech, In: Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[9] Schiel, F., Steininger, S., Türk, U., The SmartKom multimodal corpus at BAS, In: Proceedings of the 3rd Language Resources & Evaluation Conference (LREC), 2002, Las Palmas, Gran Canaria, Spain, pp. 200-206 (2002).

[10] Kawahara, H., Matsui, H., "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," In Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Vol. I, pp. 256-259, Hong Kong, 2003.

[11] Schiel, F., Steininger, S., Türk, U., The SmartKom multimodal corpus at BAS, In: Proceedings of the 3rd Language Resources & Evaluation Conference (LREC), 2002, Las Palmas, Gran Canaria, Spain, pp. 200-206 (2002).

[12] Y. Stylianou, J. Laroche, and E. Moulines. "High-Quality Speech Modification based on a Harmonic + Noise Model," Proc. EUROSPEECH, 1995.

[13] D. Vincent and O. Rosec. "A new method for speech synthesis and transformation based on a ARX-LF source-filter decomposition and HNM modeling," in ICASSP, 2007.