Abstract

Voice conversion method is applied to synthesizing emotional speech from standard reading (neutral) speech. Pairs of neutral speech and emotional speech are used for conversion rule training.

The conversion adopts GMM (Gaussian Mixture Model) with DFW (Dynamic Frequency Warping). We also adopt STRAIGHT, the high-quality speech analysis-synthesis algorithm. As conversion target emotions, (Hot) anger, (cold) sadness and (hot) happiness are used.

The converted speech is evaluated objectively first using mel cepstrum distortion as a criterion. The result confirms the GMM-based voice conversion can reduce distortion between target speech and neutral speech.

A subjective test is also carried out to investigate perceptual effect. From the viewpoint of influence of prosody, two kinds of prosody are used to synthesis. One is natural prosody extracted from neutral speech and the other is from emotional speech. The result shows that prosody mainly contributes to emotion and spectrum conversion can reinforce it.

1. Introduction

In recent years, expressive speech synthesis has become an important subject due to requests to realize more familiar human interface of a spoken dialogue system, more advanced output of TTS (Text-to-Speech) system and so on. Emotional speech synthesis is of particular interest.

Analysis of acoustic features of emotional speech and its synthesis rules have been researched[1] especially from the viewpoint of prosody. In the other hand, corpus-based approach is realized by waveform unit selection speech synthesis method with large size emotional speech corpus[2]. However this method also has difficulty on its database design, recording and labeling.

Based on these research, we attempt to synthesize acoustic parameters of emotional speech automatically. In this paper, we describe a new emotional speech synthesis method that adopts voice conversion system based on GMM (Gaussian Mixture Model) and DFW (Dynamic Frequency Warping).[3]. The conversion process is applied to generate emotional speech parameters from those of standard reading style (neutral) speech. Parameters used for conversion are mel cepstrum coefficients extracted by STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectrum), the vocoder type high-quality analysis-synthesis method[4].

As speech material, we recorded six kinds of emotional speech and neutral speech from two female professional narrators. As conversion target emotions, we choose (hot) anger, (cold) sadness and (hot) happiness.

Objective tests that compares spectrum converted from neutral speech with that of emotional natural speech are conducted using mel cepstrum distortion as a criterion. Listening tests are also carried out to clarify effect of the conversion for perceptual emotional expression.

In the following sections, design policy and recording condition of emotional speech database, an outline of the conversion method are described in sections 2 and 3, respectively. The following section describe experimental results. After discussion, we conclude this paper.

2. Speech Database

2.1. Emotions

As conversion target emotions, we use angry, sad and happy. The selection is based on concept that they are primary emotions and they can be expressed continuously and individually in speech. However, various speaking styles are observed within same emotion, e.g. according to emotional magnitude[5]. It is necessary that each speech database have consistent characteristics for to apply voice conversion technique, we recorded two kinds of speech for each emotion from the viewpoint of emotional amplitude: hot and cold.

2.2. Texts

Texts for conversion rule training speech data are compiled from Web bulletin boards, on which emotional experiences are contributed by anonymous participants. We collected 688, 523 and 587 Japanese sentences for angry, sad and happy speech recording, respectively.

From the text corpus, 61, 64 and 66 sentences are selected for training data of angry, sad and happy conversion rules. Sentence selection is conducted to include all of the CV units appearing in Japanese. As a result, several rare CVs, e.g. /nya/, /ba/ are not included.

As texts of speech data for evaluation, we use first 50 sentences from the ATR Japanese phonetically-balanced text set (503 sentences in total)[6]. This 50 sentence set (A-set) is designed to include minimum phone set of Japanese. It is desirable that evaluation texts have semantically neutral characteristics because they’re used for an emotion discrimination perceptual test. Although the ATR text set was not designed to confirm semantic neutrality, we use it because it is supposed that they are relatively free from semantic bias.

2.3. Recording Conditions

Two female professional narrators (Speaker FUM and FON, hereafter) were asked to read the text set with emotion. They’re
also asked to use consistent emotional representation within same emotion. All texts are also read with their standard reading style, the speech is used for neutral speech data. The recording conditions are illustrated in Table 1. The speech from desktop microphone (L-ch) is downsampled to 16k[Hz] and used for experiments in the following section.

<table>
<thead>
<tr>
<th>Sampling rate</th>
<th>48k [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantizing bit</td>
<td>16 [bit]</td>
</tr>
<tr>
<td>DAT recorder</td>
<td>SONY PCM-R700</td>
</tr>
<tr>
<td>Microphones</td>
<td>SONY C-38B (Desktop condenser type for L-ch.) Amcron CM311A (Headset condenser type for R-ch.)</td>
</tr>
<tr>
<td>Preamplifier</td>
<td>TASCAM MX-4</td>
</tr>
<tr>
<td>Environment</td>
<td>Soundproof chamber</td>
</tr>
</tbody>
</table>

Table 1: Recording conditions

3. Voice Conversion

3.1. Method

The voice conversion algorithm based on the GMM (Gaussian Mixture Model) was proposed by Stylianou et al.[7]. GMM continuously models acoustic space of a set of speech data, Toda et al.[3] reported an application of the voice conversion method in conjunction with the high quality vocoder-type speech analysis-synthesis method STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectrum [4]). The STRAIGHT method is based on pitch adaptive spectrogram smoothing and fundamental frequency extraction using TEMPO (Time-Domain Excitation extractor using Minimum Perturbation Operator) [8]. The STRAIGHT allows flexible manipulation of speech parameters such as vocal tract length, pitch and speaking rate while maintaining high speech quality. In addition, the DFW (Dynamic Frequency Warping) method is also introduced to generate converted speech for the subjective test (described at 4.2), because it is confirmed that this improves naturalness of the converted speech.

As acoustic features for conversion, we employ mel cepstrum coefficients derived from the smooth spectrum extracted by the STRAIGHT method. The cepstral order was set to 40, and the coefficients are used to map between source (neutral speech) and target (emotional speech) using DP matching based on cepstral distance. The number of GMM class is 64.

3.2. Target Emotions

We use hot angry, cold sad and hot happy speech data for conversion targets. The selection is based on analytical result of average mel-cepstral distance among emotional speech databases. Hereafter we simply describe these target emotions as angry, sad and happy.

4. Evaluation

To evaluate effect of conversion, we conducted objective and subjective experiments. The objective experiment is conducted to measure the distance between target speech (emotion speech) and synthesized speech converted from neutral speech using mel cepstrum distance as a criterion. On the other hand, the subjective experiments are conducted by listening tests with 10 adult listeners as subjects. They are asked to discriminate emotion in the synthetic speech. They’re carried out as a forced-choice (angry, sad, happy or neutral) test.

4.1. Objective Experiment

The objective experiments is conducted to confirm effect of the voice conversion between emotions within a same speaker.

We employed mel cepstrum distortion (Mel CD) to determine acoustic distance. Distortion between neutral speech and target emotional speech (before), and between converted speech and target emotional speech (after) are calculated using the following equation:

$$\text{MelCD} = \frac{20}{\ln 10} \sqrt{\frac{1}{49} \sum_{i=1}^{40} (mc_i^{(\text{conv})} - mc_i^{(\text{tar})})^2},$$

where $mc_i^{(\text{conv})}$ and $mc_i^{(\text{tar})}$ denote the Mel CD coefficients of the original/converted voice and the target voice, respectively. A procedure of calculating distance between converted spectrum and target emotional spectrum are illustrated in Figure 1.

4.2. Subjective Experiment

The experimental results are illustrated in Figs. 2 and 3 on speaker FUM and FON, respectively. It is observed that mel
cepstrum distortion is reduced properly after conversion for all emotions. As a rough tendency, it is shown that distortion between sad speech and neutral speech is smaller than the other combinations on both speakers.

4.2. Subjective Experiment

A subjective test is also carried out to investigate perceptual effect of the method for emotion control on synthetic speech. As we describe before, DFW process is introduced to the spectrum conversion procedure. In addition, from the viewpoint of prosodic effect, two kinds of prosodic features are given to each synthetic speech sample. One is natural prosody extracted from neutral speech and the other is those of emotional speech. We control $F_0$, speech length and speech intensity as prosodic features. Consequently, four kinds of synthetic speech described below, are used for the listening test. Analysis-synthesized speech of neutral and emotional speech is also presented for comparison.

(1) **analysis-synthesized** Analysis-synthesized speech of neutral and emotional speech.
(2) **emotional spectrum with neutral prosody** Neutral prosody speech is given to natural emotional speech.
(3) **neutral spectrum with emotional prosody** Emotional prosody is given to neutral speech.
(4) **converted spectrum with neutral prosody** Spectrum is converted to that of target emotion from neutral speech.
(5) **converted spectrum with emotional prosody** Spectrum is converted to that of target emotion from neutral speech and emotional prosody is given to it.

Four sentences for each speaker from evaluational database are used for evaluation. The synthetic speech are presented from headphones in a soundproof room using a laptop PC. The evaluation method is a forced-choice test. Subjects were asked to discriminate emotion from neutral, angry, sad and happy exclusively. Speech playback was allowed any number of times.

The result of the subjective test on speaker FUM is shown in Fig. 4. It illustrates that prosodic features mainly contribute to emotion in speech. But it is also observed that converted spectrum reinforce the emotional expression. Same tendency are observed in the result on speaker FON.

5. Discussion

As the result of subjective experiment clearly shows, prosodic features mainly dominate emotional expression. Although syn-
subjective perceptual test is also conducted. The experimental result shows that prosodic features dominate emotion in speech, but the spectrum conversion process enforce the emotional expression.

7. Acknowledgements
The authors would like to thank JST/CREST in Japan. This work was partly supported by them.

8. References