Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Speech communication: Paper ICA2016-300

Voice Conversion of emotional speech using hidden Markov model-based speech recognition and synthesis

Tetsuo Kosaka[(a)], Yoshiaki Nakagawa[(b)] and Masaharu Kato[(c)]

[(a)] Yamagata university, Japan, tkosaka@yz.yamagata-u.ac.jp
[(b)] Yamagata university, Japan, tda55485@st.yamagata-u.ac.jp
[(c)] Yamagata university, Japan, katoh@yz.yamagata-u.ac.jp

Abstract

This paper describes a many-to-one voice conversion (VC) technique that does not require a parallel training set of the source and target speakers. A VC method consisting of decoding and synthesis parts that do not require a parallel training set has already been proposed. The basic idea of this system is that an input utterance is recognized utilizing the hidden Markov model (HMM) for speech recognition and the recognized phoneme sequences are used as labels for speech synthesis. Performance of this type of system depends on phoneme recognition accuracy. The aim of this work is to improve the performance of this type of VC system using a highly accurate acoustic model for phoneme recognition. In particular, we focus on VC of emotional speech. In order to achieve this aim, emotional speech uttered by any speaker needs to be recognized. Speech recognition of emotional speech is a challenging task due to vast changes of acoustic features compared with normal speech. In order to solve the problem, we propose utilizing a deep neural network (DNN) as an acoustic model on the assumption that it is able to recognize emotional speech at a high level of accuracy. In order to evaluate the proposed system, subjective speech intelligibility tests were conducted on 8 subjects. The intelligibility was measured at the phoneme level. For the strongest emotion condition, the intelligibility scores were 78.7% and 72.9% with DNN-HMM and Gaussian mixture model (GMM)-HMM, respectively. The experimental results showed that the use of DNN-HMM contributed to the improvement of intelligibility scores in both normal and strong emotion conditions.

Keywords: voice conversion, speech recognition, speech synthesis, deep neural network

1 Introduction

This paper describes a many-to-one voice conversion (VC) technique that does not require a parallel training set of the source and target speakers. Gaussian mixture model (GMM)-based VC is a popular technique with which source speech can be converted into target speech by a GMM that consists of joint probability densities of the source and target acoustic features [1]. The GMM is trained using parallel data sets that consist of utterance pairs from the source and target speakers. However, collection of a large amount of parallel data requires considerable effort. To solve this problem, some techniques have been proposed. For example, Nose et al. proposed a new VC method that consists of a decoding step and synthesis step [2]. The basic premise of this system is that an input utterance is first recognized using hidden Markov models (HMMs) for speech recognition, and then, the recognized phoneme sequence is used as a label for speech synthesis. The performance of this type of a system depends on phoneme recognition accuracy. In a one-to-one VC, speaker-dependent (SD) HMMs can be used for recognition, while speaker-independent (SI) HMMs are needed for a many-to-one VC system. However, the SI models often do not work properly because the speaker individuality of the model does not usually match the source speaker. For this reason, the SI models were used only for phoneme alignment, and the phoneme sequence itself was given in advance in [3].

The aim of this work is to improve the performance of this type of a VC system by using highly accurate acoustic models for phoneme recognition. In particular, we focus on the VC of emotional speech. To achieve this aim, emotional speech uttered by any speaker needs to be recognized. Speech recognition of emotional speech is a difficult task owing to vast variation in acoustic features compared with normal speech. To solve this problem, we propose using a deep learning-based acoustic model. Recently, deep neural network (DNN)-based speech recognition has received considerable attention for its performance in speech recognition tasks. We expect that it will recognize emotional speech with high accuracy.

2 Voice conversion system

2.1 Overview

The proposed VC system consists of two steps, a decoding step and synthesis step. In the decoding step, an input utterance from the source speaker is analyzed; F0, energy, aperiodicity index (AP), and the phoneme sequence with duration of each phoneme are extracted. In the synthesis step, spectral parameters are generated from the phoneme sequence and the energy using HMMs. After that, a converted utterance for the target speaker is generated by a vocoder with the F0, AP, and spectral parameters. Those steps are discussed in more detail below.

2.2 Decoding step

Figure 1 shows a block diagram of the proposed VC system. For extracting the F0, energy, and AP parameters, we employ the STRAIGHT analysis-synthesis system [4]. The AP parameters are averaged to form five-dimensional data. To obtain the phoneme sequence, we utilized an automatic speech recognition system (ASR). This ASR needs to work in SI mode for realizing

many-to-one VC. In the ASR system, DNN-HMMs are used as acoustic models. In these experiments, as a comparison, we also used GMM-HMMs. The DNN-HMMs are trained using a large amount of academic presentation speech in the Corpus of Spontaneous Japanese (CSJ) [5]. We expect that using emotional speech for training the models will lead to better recognition performance; however, emotional speech was not used for training in our experiments owing to lack of data. We plan to use emotional data for adapting DNNs in the near future.

2.3 Synthesis step

In this step, we employed a statistical speech synthesis method based on HMMs [6]. Generally, only text information is used for speech synthesis. In our system, we use the F0 parameter, AP features, and phoneme sequence with duration derived from the source speaker for synthesis because we want to retain prosodic information from the source speaker. In other words, only spectral features of the target speaker are obtained using the HMM. We use mel-frequency cepstral coefficients (MFCCs) as spectral parameters. Subsequently, the STRAIGHT vocoder generates an utterance of the target speaker using the F0, AP, and spectral features.

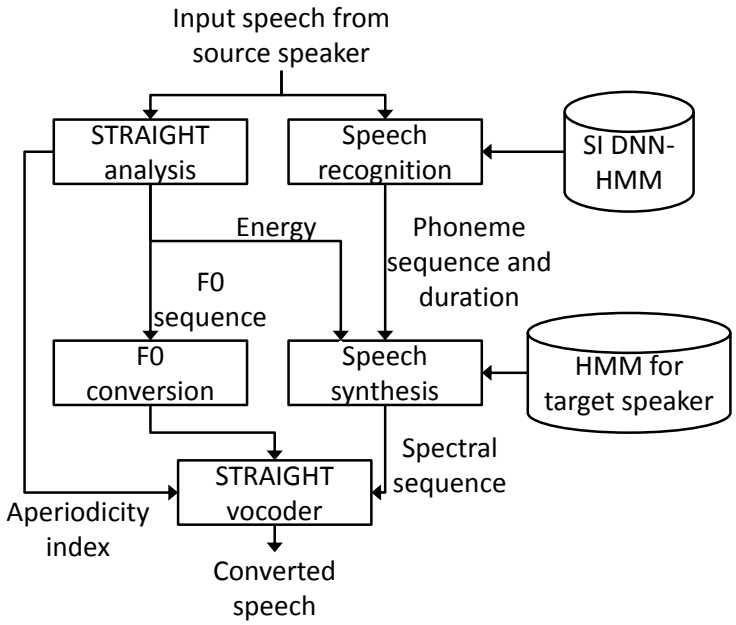

Figure 1: Block diagram of the proposed system

3 Voice conversion experiments

3.1 Experimental conditions

To decode an input utterance from the source speaker, we used DNN-HMMs. The total speech length for training is approximately 203 h. The input layer of the DNN uses 75 coefficients with a temporal context of 11 frames. The DNN has seven hidden layers with 2048 hidden units in each layer. The final output layer has 3003 units, corresponding to the total number of HMM

states. A 25-dimensional feature, which consists of the log mel-filter bank (FBANK) features and log power, is derived for each frame. Additionally, the delta and delta-delta features are calculated from the 25-dimensional feature, so the total number of dimensions is 75 per frame. For comparison, we also used GMM-HMMs. The GMM-HMMs are trained with 447 h of speech data from the CSJ. It is a set of shared-state triphones that has 3000 tied states with 32 mixtures of diagonal covariance Gaussians per state. We use five-state, left-to-right SD-HMMs for speech synthesis. Each state consists of a single Gaussian. The SD-HMMs are trained with 450 sentences per speaker from the ATR Japanese speech database. The number of target speakers is two (one male and one female). For synthesis, a 46-dimensional feature that consists of MFCCs, F0, and five-dimensional AP features is derived for each frame. In addition, the delta and delta-delta features are calculated from the 46-dimensional feature.

We use speech data from the Online Gaming Voice Chat Corpus with Emotional Label (OGVC) as the input speech from source speakers [7]; four speakers (2 males and 2 females) with emotional intensity labels are used. The emotional intensity labels consist of a four-point scale from 0 (neutral) to 3 (strong emotion).

3.2 Results of emotional speech recognition

First, we evaluated the phoneme recognition accuracy of emotional speech uttered by four speakers. For comparison, both DNN-HMMs and GMM-HMMs were used as acoustic models. Figure 2 shows the results of the experiments. Phoneme recognition accuracies of 90.1%, 85.6%, 82.4%, and 78.1% were obtained for emotional intensities of 0, 1, 2, and 3, respectively, using the DNN-HMM. The recognition performance decreased with increasing emotional intensity because the acoustic features of emotional speech vastly change with emotional intensity. In the experiments using the GMM-HMMs, accuracies of 85.4%, 78.6%, 74.2%, and 68.4% were obtained for emotional intensities of 0, 1, 2, and 3, respectively. The results show that the performance of the DNN-HMMs surpasses that of the GMM-HMM.

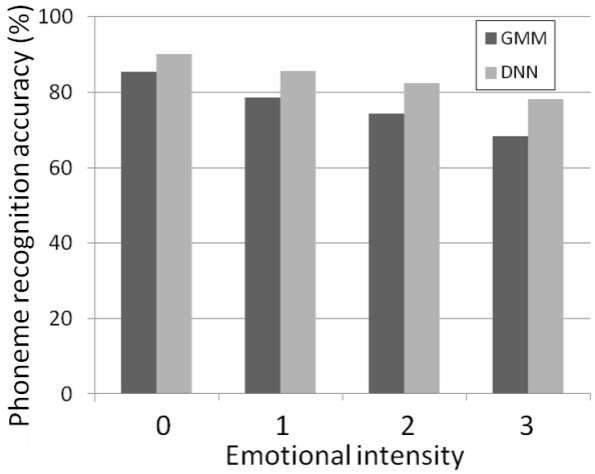

Figure 2: Phoneme recognition accuracy

3.3 Results of intelligibility evaluation

For subjective evaluation of the generated speech, we conducted subjective speech intelligibility tests with eight subjects. Intelligibility was measured at the phoneme level. Figure 3 shows the results with error bars that reflect 95% confidence interval (CI). To evaluate the effect of phoneme recognition error, we used the correct phoneme sequence in the decoding step for comparison (*oracle*). This experiment simulated the condition in which phoneme recognition accuracy was 100%. For comparison, experiments without prosody control were also conducted. In these cases, a text-to-speech (TTS) system was used where prosody information was obtained from acoustic models.

For speech with an emotional intensity label of 0, the intelligibility scores were 92.5% and 90.2% with the DNN-HMM and GMM-HMM, respectively; for speech with an emotional intensity label of 3, they were 78.7% and 72.9% with the DNN-HMM and GMM-HMM, respectively. The results suggest that the phoneme recognition accuracy of the decoding step affects the results of intelligibility tests. In addition, the intelligibility scores decrease in conditions of strong emotional intensity. In the case of *oracle*, the intelligibility score was 96.8% for speech with an emotional intensity label of 0, and 95.7% for speech with an emotional intensity label of 3. These figures indicated the upper limit of performance for the system. In the comparison between the conditions with and without prosody control, prosody-controlled output showed better intelligibility performance. The context and prosody of the input speech match well in the proposed method, while the prosody obtained from acoustic models sometimes does not match the context obtained using the TTS system. We consider that prosody information affects intelligibility score. In the case of speech with an emotional intensity label of 0, the intelligibility scores were 88.3% without prosody control and 92.5% with prosody control using the DNN-HMMs, while phoneme accuracy using the phoneme recognition unit was 90.1%. This indicates that the prosody information covers the phoneme recognition error.

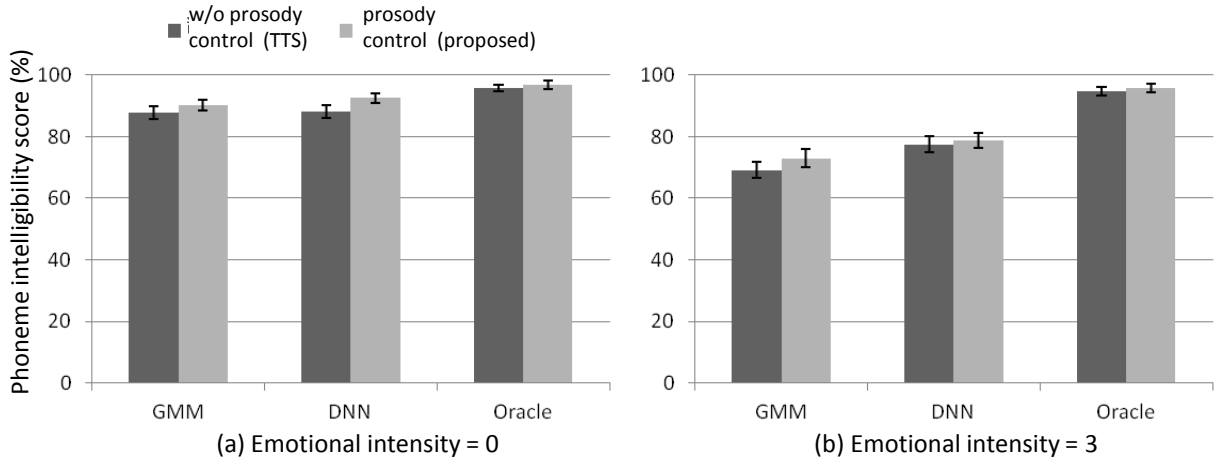

Figure 3: Phoneme intelligibility score

3.4 Subjective evaluation of emotional intensity

We subjectively evaluated how much emotional intensity was reflected in converted speech using eight subjects and 56 test sentences per subject. The subjective evaluation was conducted on a five-point scale from 1 (lowest intensity) to 5 (highest intensity). For comparison, target speech sentences were used in addition to converted speech sentences. Figure 4 shows the results of this evaluation.

Note that the subjective score of target speech is 2.7 when the emotional intensity label is 0. This means that the speaker intended to utter in an emotionless tone; however, subjects judged the utterance as emotional speech. The utterances without prosody control (i.e., those generated by the TTS system) show low emotional intensity scores in all cases. In contrast, the subjective scores vary a good deal depending on emotional intensities in prosody-controlled speech. This suggests that prosody control is very important for expression of emotion. The difference between the GMM-HMMs and the DNN-HMMs is small in these experiments. This is because the results of speech recognition mainly affect not the expression of emotion, but the intelligibility of the speech.

4 Conclusions

We investigated a many-to-one VC technique that does not require parallel training sets for the source and target speakers. The proposed system consists of decoding and synthesis steps. To improve the performance of the decoding step, we utilized DNN-HMMs for recognizing input utterance of the source speaker. The experimental results showed that the use of the DNN-HMMs contributed to the improvement of intelligibility scores in both normal and strong emotion conditions.

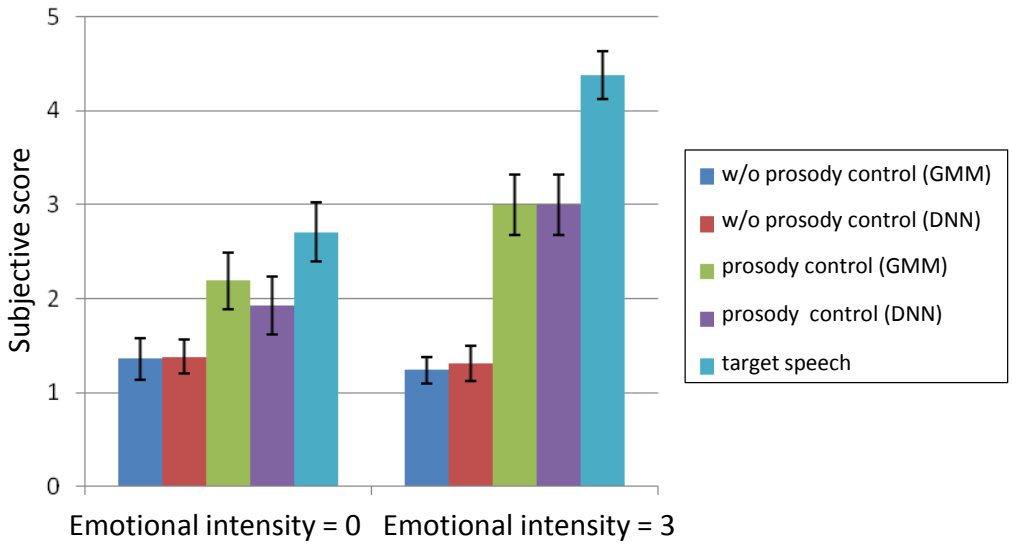

Figure 4: Subjective score of emotional intensity

Acknowledgments

We thank Dr. Takashi Nose of Tohoku University for useful advice about speech synthesis.

References

[1] Kain, A.; Macon, M.W. Spectral voice conversion or text-to-speech synthesis, *Proceedings of ICASSP'98*, Seattle, WA (USA), May 12-15, 1998, pp 285-288.

[2] Nose, T.; Ota, Y.; Kobayashi, T. HMM-based voice conversion using quantized F0 contest, *IEICE Trans. Information and Systems*, Vol E93-D (9), 2010, pp 2483-2490.

[3] Nose, T.; Kobayashi, T. Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency, *Speech communication*, Vol 53 (7), 2011, pp 973-985.

[4] Kawahara, H.; Masuda-Katsuse, I.; Cheveign´e, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Communication*, Vol 27 (3-4), 1999, pp 187-207.

[5] Furui, S.; Nakamura, M.; Ichiba, T.; Iwano, K. Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese, *Speech Communicaiotn*, Vol47 (1-2), 2005, pp 208-219.

[6] Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis," *Proceedings of EUROSPEECH'99*, Budapest, Hungary, September 5-9, 1999, pp 2347-2350.

[7] Arimoto, Y.; Kawatsu, H.; Ohno, S.; Iida, H. Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment, *Acoustical Science and Technology*, Vol 33 (6), 2012, pp 359-369.