Interspeech 2018
2-6 September 2018, Hyderabad

Effects of dimensional input on paralinguistic information perceived from synthesized dialogue speech with neural network

Masaki Yokoyama[1], *Tomohiro Nagata*[1] *and Hiroki Mori*[1]

[1]Graduate School of Engineering, Utsunomiya University, Japan

masaki@speech-lab.org, ken1@speech-lab.org, hiroki@speech-lab.org

Abstract

A novel method of controlling paralinguistic information in neural network-based dialogue speech synthesis is proposed. Controlling paralinguistic information was achieved by feeding emotion dimensions in continuous values into the input layer of the neural networks. Compared to the method using the multiple regression HMM, the naturalness of synthesized speech was improved. The controllability of paralinguistic information was evaluated by examining the shift of the distribution of synthesized parameters. A subjective evaluation test revealed that the correlation between given and perceived paralinguistic information was moderate, though less apparent compared to the multiple regression HMM-based method.

Index Terms: neural network, speech synthesis, spontaneous speech, paralinguistic information, UU Database, dimensions

1. Introduction

To make communication between human and machine closer to one between humans, speech synthesis that can express emotions and attitudes of speakers as paralinguistic information should be realized.

Most studies on speech synthesis considering emotion are based on the basic emotions theory [1]. However, human emotions are not so simple as to be explained by basic emotions alone. One of the description methods of emotion is based on dimensions [2]. A certain emotional state can be represented by a small number of axes such as pleasant-unpleasant and aroused-sleepy. In this method, a complicated emotional state can be represented by a vector having continuous values of each dimension as a component. Therefore, it can be considered that emotion can be described in more detail than expressing emotions by categories.

In addition, speech corpora used in speech synthesis research is usually read-style [3]. However, we usually do not carefully speak word as in read speech corpora. We sometimes hesitate and mispronounce in conversation. Sometimes we emits filler during thinking what to say next. Speech in conversation is also different from read speech in that it conveys speaker's emotion and attitude. For this reason, we believe that read speech corpus is inadequate for reproducing human communication in human-machine communication. Therefore, it is necessary to study, speech synthesis using natural dialogue speech corpus.

Previously, we studied a dialogue speech synthesis based on multiple-regression hidden semi-Markov model (MRHSMM) [4]. Although the MRHSMM enabled to control paralinguistic information in the form of dimensions such as pleasant-unpleasant, aroused-sleepy, etc., synthesized speech tended to have extreme parameters due to badly estimated regression matrices.We have shown that MAP estimation of regression matrices was effective to reduce the overfitting problem [4]. However, the problem still remains for certain combinations of given input as paralinguistic information.

In recent years, neural networks (NN) took place of HMMs for modeling context-dependent acoustic features for speech synthesis [5]. By using speech synthesis based on neural network, it is expected to improve the quality of synthetic speech and the controllability of paralinguistic information.

2. Natural dialogue speech corpus

The UU Database [6] is a speech corpus for studying linguistic and phonetic phenomena in expressive spoken dialogue. The database consists of natural dialogues spoken by seven pairs of college students. The task of the dialogues is "four-frame cartoon sorting." Thanks to the amusing nature of the task, the database is characterized by a wide variety of recorded expressive dialogue speech.

A major feature of the UU Database is that paralinguistic information represented by a six-dimension vector is given for each utterance. The dimensions are pleasantness, arousal, dominance, credibillity, interest and positivity. In the UU Database, paralinguistic information was annotated by three qualified annotators on a 7-point scale for each dimension. For example, for the dimension of pleasantness, 1: extremely unpleasant, 2: very unpleasant, 3: somewhat unpleasant, 4: neutral, 5: somewhat pleasant, 6: very pleasant, 7:extremely pleasant.

3. Paralinguistic information control in neural network speech synthesis

In Speech synthesis, the role of neural network is to model the relationship between linguistic features and acoustic features of speech. In this paper, we also model the dependency of acoustic features on paralinguistic information features, as well as linguistic features. This is achieved by giving paralinguistic information in the form of dimensions into the input layer. For example, giving low pleasantness and high arousal to the input layer as paralinguistic features would change the output acoustic features to those of angry or irritated utterances.

At the time of training, averaged value over three annotators, provided by the UU Database, was fed to an input unit of each paralinguistic feature. At the time of synthesis, by giving arbitrary paralinguistic information in the same way, synthesized speech that reflects the specified paralinguistic information can be obtained.

10.21437/Interspeech.2018-2042

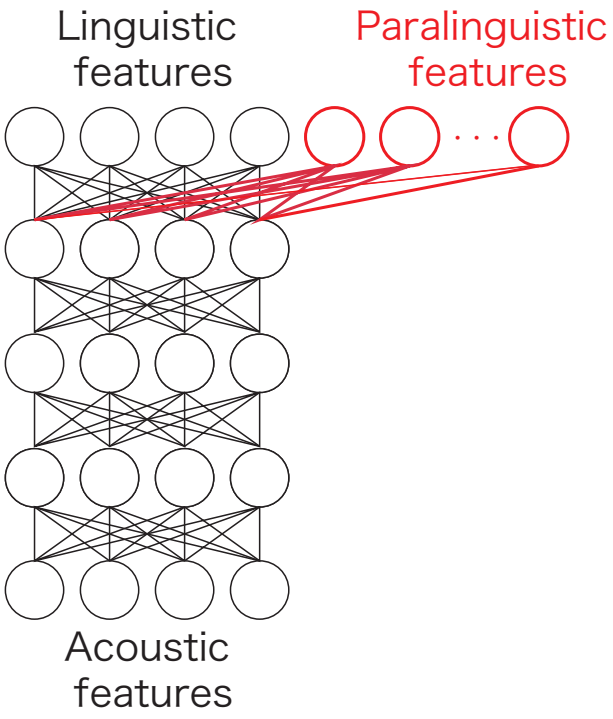

Figure 1: *Image of control method of paralinguistic information by neural network*

4. Paralinguistic information control experiment

4.1. Model structures

Model training and synthesis was performed using the speech synthesis toolkit Merlin [7]. Merlin first trains the duration model and then the acoustic model. The linguistic features given to the input layer are represented by a binary vector expressing the type of phoneme, the accent position, the number of morae, and so on, which has 385 dimensions. In addition to these, 4-dimension vector that represents the relative position of phonemes in frames.

In this study, two additional dimensions are appended as paralinguistic information. They represent the dimensions of pleasantness (pleasant-unpleasant) and arousal (aroused-sleepy), which are the most fundamental emotion dimensions [2]. Therefore, the input layer of the duration model and the acoustic model has 387 and 391 dimensions, respectively.

The output of the duration model is the duration for each phoneme. The output of the acoustic model consists of 112 dimensions, which include logarithmic fundamental frequency, band averaged aperiodicity, 35th-order mel-cepstrum coefficients, voiced/unvoiced, and their corresponding dynamic features.

The duration model and the acoustic model have a common structure with three fully-connected hidden layers, each of which has 2048 units. In both models, dropout was applied before the output layer to prevent over fitting. The activation function is linear for the output layer, and tanh for the hidden layer.

4.2. Synthesis conditions

For the model training, 559 utterances of one female speaker were used. For the test, 94 utterances were synthesized from the context labels of another 94 utterances of the same speaker. The sampling frequency was 16 kHz. WORLD [8] was used for extraction of acoustic features.

In the training, 1% of the utterance was held out for model evaluation. The learning rate and batch size of the phoneme duration model and the acoustic model was set to 0.001 and 256. The dropout was set to 0.20 for the input layer and 0.50 for the others. In both models, Adam was used as a weight optimization method.

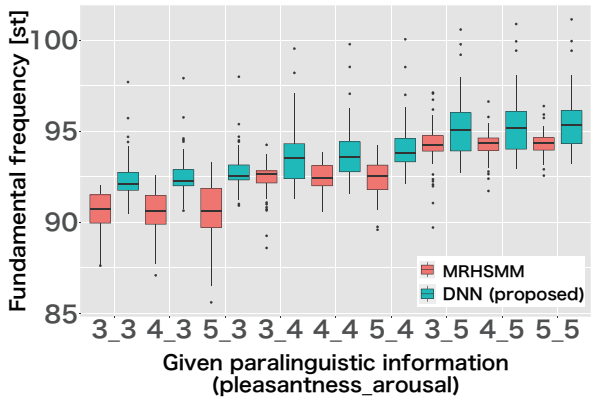

Figure 2: *Changes in the mean value of the fundamental frequency accompanying the change in paralinguistic information*

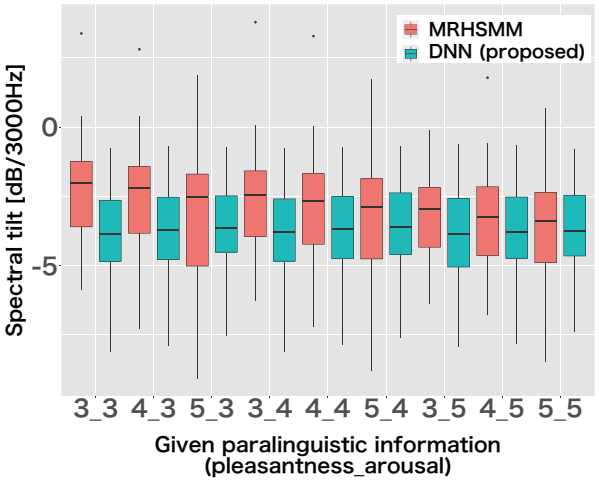

Figure 3: *Changes in spectral tilt with changes in paralinguistic information*

4.3. Analysis of synthesized speech

Controllability of paralinguistic information with the proposed deep neural network (DNN) was investigated by comparison with the method based on the MAP-MRHSMM [4].

Figure 2 shows the change in the distribution of averaged fundamental frequency of synthesized utterances, with given 2-

dimensional values of paralinguistic information. Each pair of values along the horizontal axis indicates the given pleasantness and arousal to the input (4: neutral). The result for MRHSMM shows that the fundamental frequency gets higher as the input to the arousal dimension is changed to more aroused (e.g. 4_3 → 4_4 → 4_5). Considering that the positive correlation between arousal and F0 was repeatedly mentioned in the literature [9], [10], the tendency shown in Fig. 2 is reasonable. The result for DNN also shows similar tendency. However, the correlation for DNN is not as clear as that for MRHSMM. In both methods, changing the pleasantness dimension (e.g. 3_4 → 4_4 → 5_4) does not show apparent effects on F0.

Change in spectral tilt is shown in Fig. 3. The result for MRHSMM shows that the spectral tilt gets shallower as the input to the pleasantness dimension is changed to more pleasant. The result for DNN does not show the tendency.

An example of DNN-synthesized F0 contours for a test utterance with different paralinguistic information is shown in Fig.4. The sentence was 'soodayone' ("That's true, isn't it" in Japanese). Comparing the F0 contours with different paralinguistic features, higher-pitched and longer utterance is synthesized for more pleasant and more aroused input. Also, it can be seen final rise the part of a postpositional particle "ne", which indicates modality such as asking for agreement, is more prominent for more pleasant and more aroused input. As these result show, it was confirmed that the paralinguistic information given to the input layer is reflected in the acoustic features of synthesized speech.

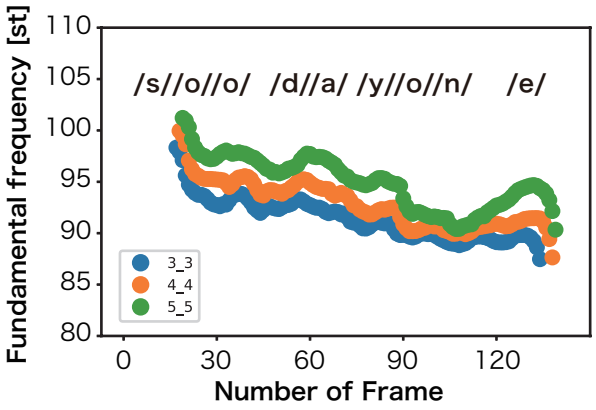

Figure 4: *Changes in input paralinguistic information and fundamental frequency*

4.4. Subjective Evaluation

To confirm the effectiveness of the proposed paralinguistic information control method for dialogue speech synthesis, a subjective evaluation test was conducted. The effectiveness was evaluated from the following two aspects:

1. naturalness
2. controllability of paralinguistic information

The stimulus set was composed of utterances synthesized with MRHSMM and DNN (proposed). The stimuli also include the utterances synthesized with a conventional DNN with the same structure as the proposed DNN except that it does not accept the paralinguistic features. Ten utterances were selected

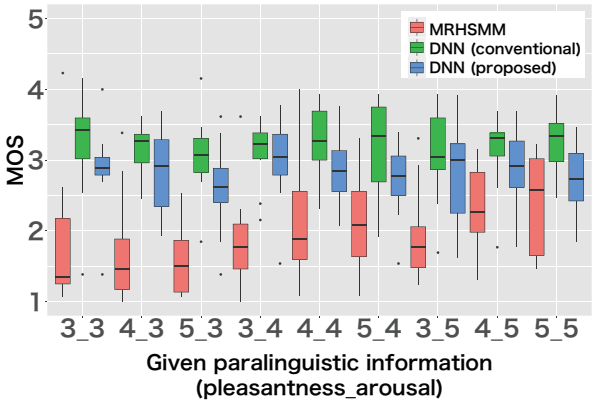

Figure 5: *Results of naturalness evaluation experiment*

from the test set for the evaluation. The set of target paralinguistic information was the nine combinations of pleasantness (3, 4, 5) and arousal (3, 4, 5). Hence, the number of stimuli was 3 (methods) × 10 (test sentences) × 9 (control vectors) = 270. 13 subjects participated in the experiments. Stimuli were presented to each subject through headphones in a quiet laboratory. Each stimulus was presented only once to the subjects.

The subjects were asked first to evaluate the naturalness of each stimulus on a 5-point scale, then to evaluate the perceived paralinguistic information for each stimulus on a 7-point scale, in the same way as evaluating natural utterances in the UU Database [6].

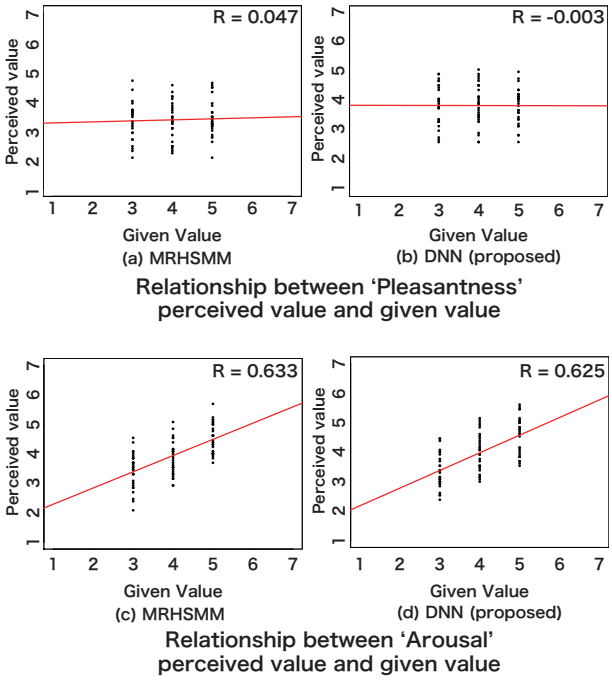

Figure 6: *Perceptual experiment results of paralinguistic information control*

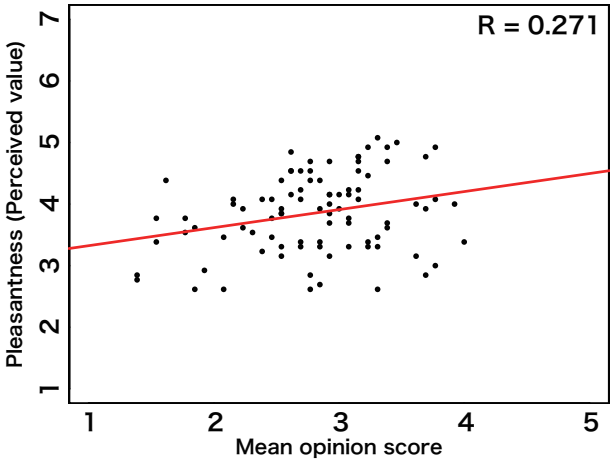

Figure 7: *Relationship between the naturalness and perceived pleasantness*

4.5. Experimental Results and Discussion

The distribution of mean opinion score (MOS) for the naturalness is shown in Fig. 5. The average of the evalution values of naturalness was 1.98, 3.07 and 2.77 in each of MRHSMM (conventional) which is a conventional method of paralinguistic information control, DNN (conventional) which not involving control of paralinguistic information, and DNN (proposed) with control of paralinguistic information it was.

Some utterances synthesized with MRHSMM were perceived as extremery unnatural, typically for some combinations of pleasantness and arousal (3_3, 4_3, 5_3, 3_4, 3_5). On the other hand, utterances synthesized with DNN (conventional) and DNN (proposed) were perceived as relatively natural, regardless of given paralinguistic information. It can be seen that the method based on DNN that does not control paralinguistic information has higher naturalness of synthesized speech than method with control.

Figure 6 shows the distributions of perceived paralinguistic information, where the horizontal axes indicate the given value for (a)(b) pleasantness and (c)(d) arousal, and the vertical axes indicate the evaluated values averaged over the subjects.

A positive correlation ($R = 0.625$) was observed between given and perceived arousal for the DNN-synthesised utterances. The correlation is moderately high, and close to that of MRHSMM. For reference, $R = -0.075$ for the conventional DNN without paralinguistic input features. From this result, it can be concluded that the arousal perceived from synthesized speech can be controlled with the proposed DNN.

On the other hand, no correlation was observed between given and perceived pleasantness.

Fig. 7 illustrates the relationship between the naturalness and perceived pleasantness. The correlation coefficient was 0.27, and utterance with higher naturalness tended to be perceived as more pleasant. This implies that naturalness affected the evaluation of paralinguistic information.

In the case of DNN speech synthesis, the change in acoustic features when paralinguistic features were manipulated was smaller than expected, typically shown in Fig. 2. Although the difference was clearly perceived as shown in Fig. 6, this issue should be addressed by deeper investigation.

5. Conclusions

In this paper, we compared the MRHSMM and DNN as paralinguistic information control methods for dialog speech synthesis. The analyses of synthesized speech revealed that MRHSMM tends to be more sensitive to changes in paralinguistic information. On the other hand, the DNN-based method provides the controllability of paralinguistic information in the form of emotion dimensions, without sacrificing the naturalness.

6. References

[1] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Synposium on Motivation*, vol. 19, pp. 207–282, 1972.

[2] J. Russell, "A circumplex model of affect," vol. 39, no. 6, pp. 1161–1178, 1980.

[3] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive hmm-based text-to-speech synthesis," vol. 17, no. 6, pp. 1208–1230, 2009.

[4] T. Nagata, H. Mori, and T. Nose, "Dimensional paralinguistic information control based on multiple-regression hsmm for spontaneous dialogue speech synthesis with robust parameter estimation," *Speech Communication*, vol. 88, pp. 137–148, 2017.

[5] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," pp. 7962–7966, 2013.

[6] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Commun.*, vol. 53, pp. 36–50, 2011.

[7] S. K. Zhizheng Wu, Oliver Watts, "Merlin: An open source neural network speech synthesis system," *in Procceding. 9th ISCA Speech Synthesis Workshop (SSW9)*, 2016.

[8] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[9] K. Scherer, "Vocal affect expression. a review and a model for future research," *Psychological Bulletin*, vol. 99, no. 2, 1986.

[10] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evalution and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.