Objective evaluation of the DNN-based dialog speech synthesizer with dimensional control of emotion *

☆ Masaki Yokoyama, Tomohiro Nagata and Hiroki Mori (Utsunomiya University)

1 Introduction

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

Most studies on speech synthesis considering emotions are based on the basic emotion 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 the dimensions [2]. With this method emotion can be described in more detail than with emotion categories.

Most, speech corpora used in speech synthesis research is read-style [3]. However, 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 corpora is inadequate for reproducing human communication by speech synthesis, which leads us to investigate speech synthesis using natural dialog speech corpus.

Previously, we studied a dialog 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 of paralinguistic information.

In recent years, on the other hand, neural networks (NN) took place of HMMs for modeling context-dependent acoustic parameters for speech synthesis [5]. Incorporating neural network is expected to improve the quality of synthetic speech, as well as the controllability of paralinguistic information.

In this paper, we propose a method of controlling paralinguistic information in neural network-based dialog speech synthesis. And then, we studied synthetic speech by objective evaluation.

2 Natural dialog speech corpus

The UU Database [6] is a speech corpus for studying linguistic and phonetic phenomena in expressive spoken dialog. The database consists of natural dialogs spoken by seven pairs of college students. The task of the dialogs is "four-frame cartoon sorting." Thanks to the amusing nature of the task, the database is characterized by a wide variety of recorded expressive dialog 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. 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 and 7: extremely pleasant.

3 Paralinguistic information control in DNN speech synthesis

In speech synthesis, the role of neural network is to model the relationship between linguistic features and acoustic parameters of speech. In this paper, we also model the dependency of acoustic parameters on paralinguistic information, as well as linguistic features.

This is achieved by giving paralinguistic information in the form of dimensions into the input layer, as shown in Fig. 1. For example, giving low pleasantness and high arousal to the input layer as paralinguistic information would change the output acoustic parameters 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 dimension. At the time of synthesis, giving arbitrary paralinguistic informa-

*感情の次元制御による DNN 対話音声合成の客観評価 , 横山雅季, 永田智洋, 森大毅 (宇都宮大)

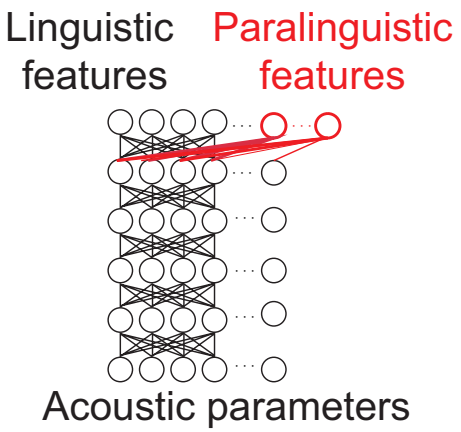

Fig. 1: Network architecture.

tion to the input layer will make the synthesized speech reflect the specified paralinguistic information.

4 Paralinguistic information control

4.1 Model structures

Model training and synthesis was performed using the speech synthesis tool nnmnkwii [7] and PyTorch [8]. 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, a 4-dimension vector that represents the relative position of phonemes in frames is given to the input of the acoustic model. In this study, two additional dimensions are appended as paralinguistic information. Therefore, the input layer of the duration model and the acoustic model has 387 and 391 dimensions, respectively.

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 four fully-connected hidden layers, each of which has 2048 units. In both models, dropout was applied before the output layer to prevent over fitting.

4.2 Model training

For the model training, 559 utterances of one female speaker were used. The sampling frequency was 16 kHz. World [9] was used for extraction of acoustic parameters. 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.

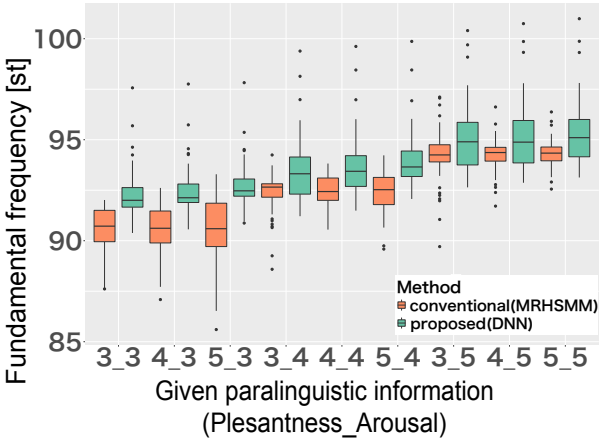

Fig. 2: Changes in the mean value of the fundamental frequency with changes in paralinguistic information.

5 Objective evaluation

The effect of manipulating input paralinguistic information on the output acoustic parameters with the proposed deep neural network (DNN) was investigated. For the test, 94 utterances that do not overlap the training set were synthesized with the conventional MAP-MRHSMM and proposed DNN.

Fig. 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 [10, 11], 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. A two way ANOVA (pleasantness(3) × arousal(3)) for DNN revealed a significant main effect of pleasantness ($F(2, 845) = 138.567$, $p < 0.01$) and arousal ($F(2, 845) = 1066.680$, $p < 0.01$).

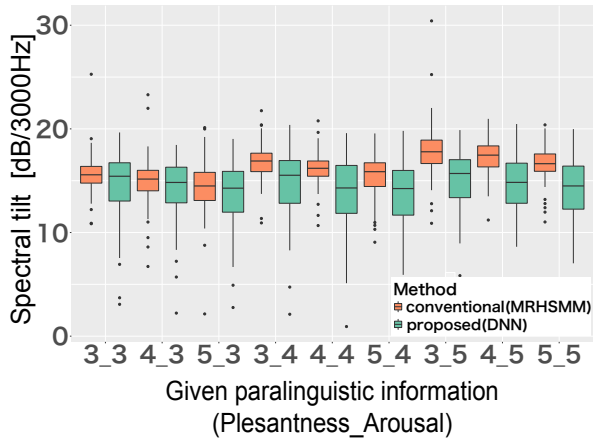

Fig. 3: Changes in spectral tilt at vowel /a/ with changes in paralinguistic information.

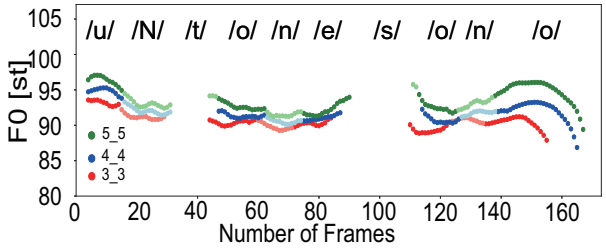

Fig. 4: F0 contours for different combinations of paralinguistic information.

Changes in spectral tilt of vowels /a/ are shown in Fig. 3. The result shows that the spectral tilt gets shallower as the input to the pleasantness dimension is changed to more pleasant. Because steep spectral tilt makes speech softer, it can be said that unpleasantness makes the synthesized speech soft. A two way ANOVA (pleasantness(3) $\times$ arousal(3)) for DNN revealed a significant main effect of pleasantness ($F(2, 845) = 52.785$, $p < 0.01$) and arousal ($F(2, 845) = 7.308$, $p < 0.01$).

An example of DNN-synthesized F0 contours for a test utterance with different paralinguistic information is shown in Fig. 4. The sentence was 'uNtone sono' ("um the ..." in Japanese). Comparing the F0 contours with different paralinguistic features, higher-pitched and longer utterance is synthesized for more pleasant and more aroused input.

Although there is no correlation between pleasantness/arousal and phoneme duration, there does exist a positive correlation between pleasantness/arousal and pause duration. A two way ANOVA (pleasantness(3) $\times$ arousal(3)) for DNN revealed a significant main effect of pleasantness ($F(2, 845) = 63.383$, $p < 0.01$) and arousal ($F(2, 845) = 61.462$, $p < 0.01$). Also, a significant interaction between pleasantness and arousal was found ($F(4, 845) = 59.602$, $p < 0.01$). A simple test revealed that pause duration was influenced by pleasantness, only aroused was low level (3 and 4).

From these results, it can be concluded that the paralinguistic information given to the input layer is reflected in the acoustic parameters of synthesized speech.

6 Subjective Evaluation

6.1 Experimental Conditions

To confirm the effectiveness of the proposed paralinguistic information control method for dialog 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 information as input. Ten utterances were selected 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) $\times$ 10 (test sentences) $\times$ 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 the evaluate perceived paralinguistic information for each stimulus on a 7-point scale, in the same way as evaluating natural utterances in the UU Database [6].

6.2 Experimental Results and Discussion

The distribution of mean opinion score (MOS) for the naturalness is shown in Fig. 5. The average of naturalness for MRHSMM, DNN (conventional), and DNN (proposed) was 1.98, 3.07 and 2.77, respectively. Some utterances synthesized with MRHSMM were perceived as extremery unnatural, typically for some combinations of pleasantness and

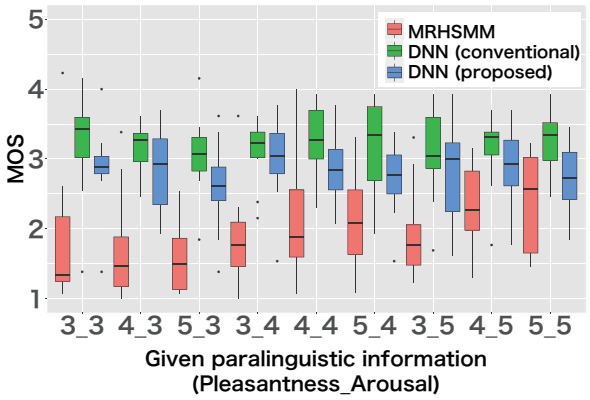

Fig. 5: Results of naturalness evaluation experiment.

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 utterances synthesized with DNN (conventional) has higher naturalness than those synthesized with DNN (proposed). Further investigation revealed that the difference of naturalness mainly came from the duration. We do not think the difference is essential, because this phenomenon disappears with a careful treatment of network initialization.

Fig. 6 shows the distribution between given and perceived values of paralinguistic information, which corresponds to the controllability of paralinguistic information. Pleasantness seems hardly controllable with both methods. On the other hand, regarding the control of arousal, the correlation coefficient with the proposed DNN was almost the same as that of the conventional method. For reference, $R = -0.075$ for the conventional DNN that does not accept paralinguistic information as input.

From these results, it can be concluded that the arousal perceived from synthesized speech can be controlled with the proposed DNN, which at the same time provides the improvement in naturalness.

7 Conclusions

In this paper, we compared the MRHSMM and DNN as paralinguistic information control methods for dialog speech synthesis. The analysis 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.

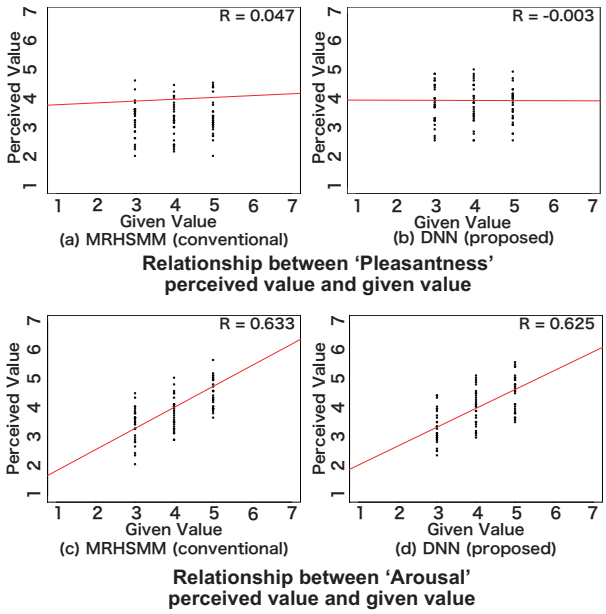

Fig. 6: Perceptual experiment results of paralinguistic information control.

References

[1] P. Ekman, Nebraska Symposium on Motivation, **19**, 207–282, 1972.

[2] J. Russell, J. Pers. Soc. Psychol., **39**, 1161–1178, 1980.

[3] Yamagishi *et al.*, IEEE Transactions on Audio, Speech and Language Processing, **17**, 1208–1230.

[4] Nagata *et al.*, Speech Commun., **88**, 137–148, 2017.

[5] Zen *et al.*, Proc. ICASSP 2013, 7962–7966, 2013.

[6] Mori *et al.*, Speech Commun., **53**, 36–50, 2011.

[7] https://github.com/r9y9/nnmnkwii

[8] https://github.com/pytorch/pytorch

[9] M. Morise, Speech Commun., **84** ,57–65 ,2016.

[10] K. Scherer, Psychological Bulletin, **99**, 143–165, 1986.

[11] Grimm *et al.*, Speech Commun., **49**, 787–800, 2007.