An Investigation to Transplant Emotional Expressions in DNN-based TTS Synthesis

Katsuki Inoue*, Sunao Hara*, Masanobu Abe*, Nobukatsu Hojo†, and Yusuke Ijima†
* Graduate School of Natural Science and Technology, Okayama University, Japan
E-mail: k_inoue@a.cs.okayama-u.ac.jp, {hara,abe}@cs.okayama-u.ac.jp
† NTT Media Intelligence Laboratories, NTT Corporation, Japan
E-mail: {hojo.nobukatsu,ijima.yusuke}@lab.ntt.co.jp

***Abstract*—In this paper, we investigate deep neural network (DNN) architectures to transplant emotional expressions to improve the expressiveness of DNN-based text-to-speech (TTS) synthesis. DNN is expected to have potential power in mapping between linguistic information and acoustic features. From multi-speaker and/or multi-language perspectives, several types of DNN architecture have been proposed and have shown good performances. We tried to expand the idea to transplant emotion, constructing shared emotion-dependent mappings. The following three types of DNN architecture are examined; (1) the parallel model (PM) with an output layer consisting of both speaker-dependent layers and emotion-dependent layers, (2) the serial model (SM) with an output layer consisting of emotion-dependent layers preceded by speaker-dependent hidden layers, (3) the auxiliary input model (AIM) with an input layer consisting of emotion and speaker IDs as well as linguistics feature vectors. The DNNs were trained using neutral speech uttered by 24 speakers, and sad speech and joyful speech uttered by 3 speakers from those 24 speakers. In terms of unseen emotional synthesis, subjective evaluation tests showed that the PM performs much better than the SM and slightly better than the AIM. In addition, this test showed that the SM is the best of the three models when training data includes emotional speech uttered by the target speaker.**

I. Introduction

Deep neural network (DNN) has been successfully applied to many areas including speech processing. Recently, DNNs were used to address the limitations of HMM-based text-to-speech (TTS) synthesis. Researchs indicate that DNN-based TTS synthesis can outperform HMM-based TTS synthesis in terms of the quality of synthesized speech [1], [2], [3]. Currently, one of the hottest topics in DNN-based TTS synthesis is to improve flexibility and controllability. For example, in relation to speaker variability, Fan et al. proposed a multitask learning framework [4], Wu et al. utilized speaker adaptation methods [5], and Hojo et al. introduced speaker codes as additional inputs [6]. In addition, Luong et al. attempted to handle gender and age [7], and Li et al. proposed an architecture that can be expanded to easily incorporate new languages [8]. However, few studies exist regarding the divergence of emotional expressions in DNN-based TTS synthesis. To address this, we investigate the divergence of emotional expressions in DNN-based TTS synthesis.

In relation to flexibility and controllability, HMM-based TTS synthesis has shown good performances for variabilities in speaker, speaking styles, and emotional expressions. Yamagishi et al. proposed a method that modeled speaking styles [9], and using that method Tachibana et al. proposed a style interpolation approach to flexibly synthesize various types of speech [10]. Nose et al. proposed a style vector that generated a variety of speaking styles in a multiple-regression manner [11]. However, another approach is available: transplanting a particular speaking style to the neutral speaking style. Kanagawa et al. suggested generating speaker-independent transformation matrices using pairs of neutral and target-style speech and applying these matrices to neutral-style model of a new speaker [12]. Similarly, Trueba et al. [13] proposed to extrapolate the expressiveness of proven speaking-style models into speakers who utter speech in a neutral speaking style using constrained structural maximum a posteriori linear regression (CSMAPLR) algorithm [14]. Ohtani et al. proposed an emotion additive model to transplant emotion into a neutral voice [15]. To address the lack of research on the emotional transplant in DNN-based TTS synthesis, in this paper, we tackle to transplant emotional expressions from the perspective of improving flexibility and controllability.

We investigate DNN architectures suitable for transplanting emotional expressions. In general, to obtain flexibility and controllability, we assume that a DNN architecture should exhibit structures that correspond explicitly to factors such as linguistics, speaker, and emotional information; however, the structure should not be a black box. The separation of speaker and emotional factors appears possible [12], [13]. If a DNN architecture with explicit factors is constructed, emotions can be explicitly controlled. Based on this assumption, we examine three types of DNN architectures: the parallel model (PM), the serial model (SM), and the auxiliary input model (AIM). The evaluation of DNN architectures was based on the availability of emotional speech uttered by a target speaker. In this paper, emotional speech synthesis is divided into trained emotion and transplanted emotion. Trained emotion is defined as a model trained using speech uttered by the target speaker (in the same way as [9], [10], [11]). Transplanted emotion is defined as a model trained without access to the target speaker's emotional speech (in the same way as [12], [13], [15]).

This paper is organized as follows. In Section 2, we describe the three DNN architectures, and our motivation for using each one. In Section 3, the three DNN architectures are evaluated

from the perspective of trained emotion and transplanted emotion. Conclusions and suggestions for future work are presented in Section 4.

II. Three DNN architectures used to transplant emotions

Initially, we describe the features that are regularly used in each of the proposed DNN architectures to control speakers and emotions. Several methods are proposed to control speakers or emotions, e.g., one-hot vector [6], [11], i-vector [5]. Since the one-hot vector is simple and intuitive, we adopt its features to control the speakers and their emotions. One-hot speaker and emotion vectors are used for the speaker ID and emotion ID, respectively. The speaker ID $S^{(i)}$ for the *i*-th speaker is defined as $S^{(i)} = (s_1^{(i)}, s_2^{(i)}, ..., s_N^{(i)})$, where each value $s_n^{(i)}$ is expressed as follows.

$$s_n^{(i)} = \begin{cases} 1 & (n = i) \\ 0 & (n \neq i) \end{cases} \quad (1)$$

where *N* is the dimension of S and equal to the number of speakers in the training data. The emotion ID $E^{(j)}$ for the *j*-th emotion is defined as $E^{(j)} = (e_1^{(j)}, e_2^{(j)}, ..., e_M^{(j)})$, where each value $e_m^{(j)}$ is expressed as follows.

$$e_m^{(j)} = \begin{cases} 1 & (m = j) \\ 0 & (m \neq j) \end{cases} \quad (2)$$

where *M* is the dimension of E and equal to the number of emotions in the training data. The IDs are used to switch layers for training both in the PM and the SM. In the AIM, the IDs are used as auxiliary input features.

The proposed DNN architectures have the structure that learns by switching the combination of speaker and emotion using speaker and emotion IDs. The proposed DNN architectures are able to synthesize the speech with any combination of speaker and emotion simply by selecting speaker and emotion IDs, even when no training data is available for the selected combination.

A. Parallel Model

The Parallel Model (PM) architecture is shown in Fig. 1. We anticipate that the output layers of the PM will handle both the speaker and emotional factor. The PM is newly proposed and is motivated by a multi-speaker DNN [4] and the emotion additive model [15]. In the multi-speaker DNN, hidden layers are regarded as global linguistic feature transformation that are shared by all speakers. Similarly, the PM has an output layer comprised of speaker-dependent layers (Speaker 1, Speaker 2, ... Speaker N, as shown in Fig. 1) and emotion-dependent layers (Emotion 1, Emotion 2, ... Emotion M, also shown in Fig. 1). Therefore, we anticipate that the output speaker-dependent layer has an emotion-independent speaker factor, and the output emotion-dependent layer has a speaker-independent emotional factor. All the speakers and emotions share the hidden layers as the global linguistic feature transformation. Since the PM can output from either the speaker's output layer or the emotion's output layer, the advantage of the PM is that it analyzes the output features with reflection to either the speaker's factor or the emotion's factor.

During the training phase, one speaker and one emotion are simultaneously trained. For example, when a joyful speech of speaker A is supplied, parameters of speaker A's layer and those of the joyful layer are updated. The other speaker- and emotion-dependent layers remain fixed. The selection of speaker and emotion is controlled by the speaker and emotion IDs. The output features from one selected speaker and emotion layer are added as the final output features.

For speech synthesis, only one speaker and one emotion are selected using the speaker and emotion IDs.

B. Serial Model

The Serial Model (SM) architecture is shown in Fig. 2. We anticipate that the hidden layer and the output layer of the SM will handle the speaker factor and the emotional factor, respectively. In other words, we organize the method to add the emotional factor to the speaker factor that is trained from neutral-emotion speech [15]. The SM is newly proposed and is also motivated by a multi-speaker DNN [4]. The SM architecture involves a straightforward expansion to emotion, in which emotion-dependent layers (Emotion 1, Emotion 2, ... Emotion M, as shown in Fig. 2) are simply added over speaker-dependent layers (Speaker 1, Speaker 2, ... Speaker N also shown in Fig. 2). The advantage of the SM is that it clearly separates emotional factors from speaker factors in the preceding layer, thereby filtering out speaker features before passing on the features to the next layer.

During the training phase, one speaker and one emotion are selected from the multi-speaker hidden layer and the multi-emotion output layer, respectively. For example, when the target speech is the joyful speech of speaker A, the speaker A's hidden layer and the joyful output layer are selected using the speaker ID and the emotion ID, respectively. Other speakers in the multi-speaker hidden layer and other emotions in the multi-emotion output layer are not used during the training phase. The output features obtained via one selected speaker hidden layer and one selected emotion output layer are considered as the final output features.

For speech synthesis, only one speaker and one emotion are selected using the speaker and emotion IDs.

C. Auxiliary Input Model

The Auxiliary Input Model (AIM) architecture is shown in Fig. 3. We expect to model the speaker and emotional factors over the entire AIM architecture. The AIM is motivated by speaker codes [6]. In addition to the linguistic features used in conventional DNN-based TTS synthesis, an AIM input layer contains both speaker and emotion IDs. Compared with the PM and the SM, the AIM architecture has no explicit structure that decomposes factors into linguistic, speaker, and emotion information, i.e., all its mappings are distributed implicitly over the entire DNN. The AIM displays the simplest model

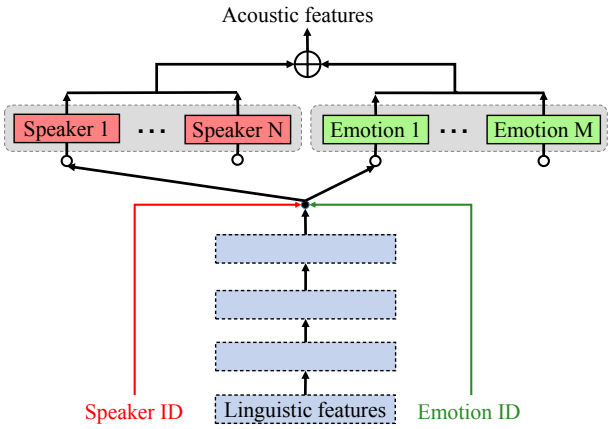

Fig. 1. Parallel Model.

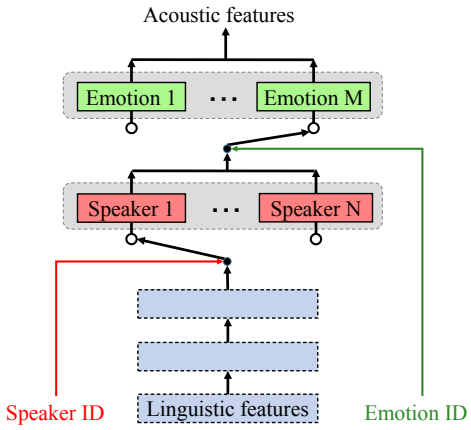

Fig. 2. Serial Model.

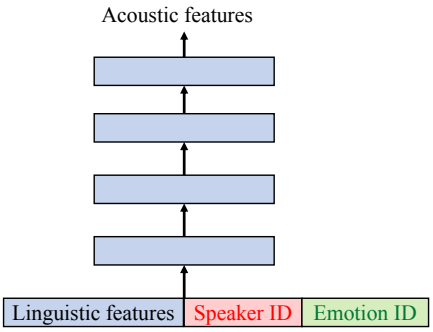

Fig. 3. Auxiliary Input Model.

architecture of the three proposed model architectures. However, one of the disadvantages of the AIM involves difficulty in checking the mapping of speaker and emotion information.

For speech synthesis, we anticipate that the AIM selects the speaker and emotion mapping by using speaker and emotion IDs, respectively. The output features resulting from these mappings are considered as the final output features.

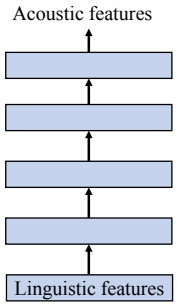

Fig. 4. Speaker- and emotion-Dependent models.

III. Performance Evaluation

A. Experimental procedures

In the experiments, we constructed Japanese speech corpora using two sets of text transcription: α and β. The set α contains 427 sentences, and 16 speakers (comprising 6 male and 10 female speakers) uttered a part of set α with neutral emotion. The average number of sentence is 294, and 134 sentences were uttered in common. The set β contains 915 sentences, and a male and two female speakers uttered a part of set β with neutral, sad and joyful emotions. The average number of sentence is 670, and 508 sentences were uttered in common. Moreover, 5 speakers (comprising 2 male and 3 female speakers) uttered a part of set β with neutral emotion. The average number of sentence is 531. The duration of each emotional speech uttered by each speaker was approximately one hour. The speech signals were sampled at 22.05 kHz at 16 bits. STRAIGHT [16] analysis was used to extract 40-dimensional mel-cepstral coefficients, 10 band aperiodicities, and F0 in log-scale in 5-ms steps.

Tables I and II show the speech data combinations used to train the DNNs for evaluating the transplanted emotion of female speakers A and B, respectively. For instance, DNNs are trained using speech (as shown in Table I), and sad speech is synthesized with the IDs of female speaker A and sad emotion. This is the transplanted sad emotion. The synthesized speech is then evaluated by comparing it to the sad speech uttered by the female speaker A (as shown in Table II).

To evaluate the trained emotion, the speech data combinations, shown in Table III, were used. The trained emotion signifies that the target emotional speech is used for training the DNNs. To act as a reference, the speaker- and emotion-dependent model (SD) is trained. The architecture of the SD is shown in Fig. 4, and it is the same as the conventional DNN. The SD is trained using only the target speaker's target emotional speech.

The input feature vectors are identical for the PM, SM, and SD. There are 294 dimensional binary features of categorical linguistic contexts (e.g., quinphone, the current frame position of the phoneme duration), and 11 dimensional numerical linguistic contexts (e.g., the number of mora in the current word). The dimension of the input feature vectors for the AIM was 332 because both speaker and emotion IDs are

TABLE I
TRAINING DATA TO EVALUATE THE TRANSPLANTED EMOTION OF FEMALE A.

	Female A	Female B	Male	21 speakers
Neutral	◯	◯	◯	◯
Sad	-	◯	◯	-
Joyful	-	◯	◯	-

TABLE II
TRAINING DATA TO EVALUATE THE TRANSPLANTED EMOTION OF FEMALE B.

	Female A	Female B	Male	21 speakers
Neutral	◯	◯	◯	◯
Sad	◯	-	◯	-
Joyful	◯	-	◯	-

TABLE III
TRAINING DATA TO EVALUATE THE TRAINED EMOTION.

	Female A	Female B	Male	21 speakers
Neutral	◯	◯	◯	◯
Sad	◯	◯	◯	-
Joyful	◯	◯	◯	-

added as auxiliary features. The speaker ID was 24 dimensions and the emotion ID was 3 dimensions. The output feature vector contains log F0, 40 mel-cepstral coefficients, 10 band aperiodicities, their delta and delta-delta counterparts, and a voiced/unvoiced flag, which sums up to 154 dimensions. The voiced/unvoiced flag is a binary feature that indicates the voicing of the current frame. It should be noted that 80% of the silent frames are removed from the training data to balance the training data and reduce the computational cost. The output features of the training data are normalized to zero mean and unit variance. In these experiments, phoneme durations extracted from natural speech were used. This is mainly because, as the first step, we would like to model F0 and spectrum parameters. We have a plan to model phoneme durations later.

The number of hidden layers and units per layer was determined experimentally. The AIM and SD each has three hidden layers with 256 neurons in each layer. A sigmoid function was used in the hidden layers followed by a linear activation at the output layer. The PM contains three shared hidden layers with 256 neurons: the speaker output layers (N=24), and the emotion output layers (M=3) using a linear activation function. The SM contains two shared hidden layers with 256 neurons, the speaker hidden layers (N=24) using sigmoid function with 256 neurons and the emotion output layers (M=3) using a linear activation function. The AIM contains 256,410 model parameters; the PM contains 1,278,526 model parameters; and the SM contains 1,841,870 model parameters. For the training process, the weights of all DNN (PM, SM, AIM, and SD) were randomly initialized. The weights are trained using a backpropagation procedure with a minibatch-based MomentumSGD to minimize the mean squared error between the output features of the training data and the predicted values. The initial learning rate of MomentumSGD is 1.28, and the momentum is 0.9. The training data for the minibatch

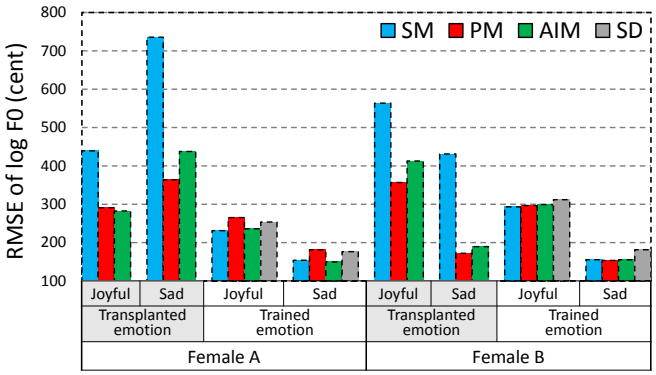

Fig. 5. The objective evaluation results of the root mean squared error (RMSE) of log F0.

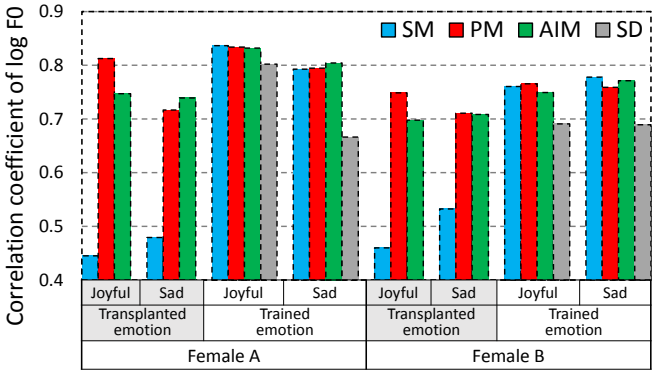

Fig. 6. The objective evaluation results of correlation coefficient of log F0.

was randomly selected, and the minibatch size was 128. The schedule of the training is a similar method to randomly select the data as conventional DNN.

B. Objective evaluation

We evaluated the performance of the three proposed methods using objective measures. Those measures were the root mean squared error (RMSE) of log F0, the correlation coefficient of log F0, and the mel-cepstral distortion. We used 20 utterances, which were not used for training, as test data. Three types of objective measures were calculated frame-synchronously between the parameters extracted from emotional speech uttered by the target speaker and the parameters generated by the DNN-based TTS synthesis.

Fig. 5 shows the results for the RMSE of log F0. In all proposed models, the trained emotion outperforms the transplanted emotion. The result is reasonable because speech uttered by a target speaker is always more preferable for training models than that of a different speaker. Relative to the transplanted emotion, the PM and AIM perform significantly better than the SM, and the PM performs slightly better than the AIM. In contrast, for the trained emotion, no significant difference between the three proposed models is observed. In addition, the SD exhibited poor performance compared with the three proposed models.

Fig. 6 shows the results for the correlation coefficient of log F0. Compared to the RMSE of log F0, there is little difference

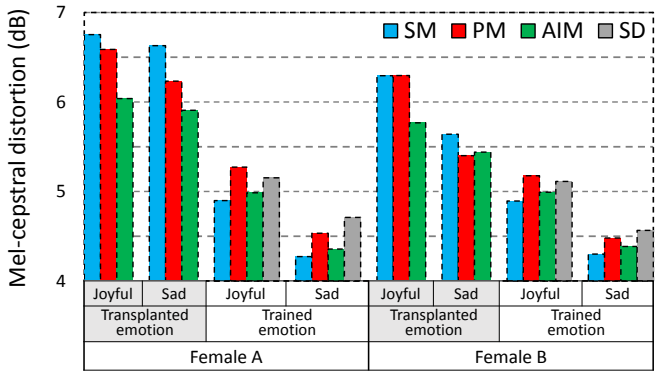

Fig. 7. The objective evaluation results of Mel-cepstral distortion.

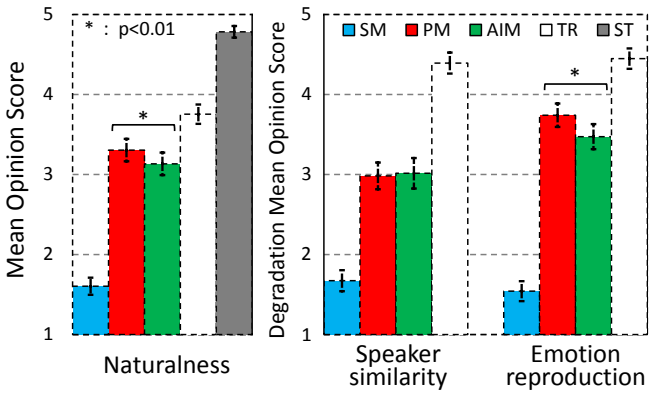

Fig. 8. Naturalness, Speaker similarity, Emotion reproduction test results with their 95% confidence interval.

between the trained emotion and transplanted emotion. However, in terms of the transplanted emotion, the same tendencies as the RMSE of log F0 are observed, i.e., the PM performs slightly better than the AIM, and the SM shows the poorest performance. The SD performed more poorly than the three proposed models in the same manner as the RMSE of log F0.

Fig. 7 shows the mel-cepstral distortion results. The trained emotion again displays a better performance than that of the transplanted emotion for all proposed models. Interestingly, for the trained emotion, the SM exhibits a better performance than the other models; however, the SM produces the worst performance for the transplanted emotion. In terms of the transplanted emotion, the AIM shows better performance than the PM, which is the opposite of the tendency observed for the evaluation results for log F0.

From these results, we can summarize our judgments as follows. In terms of transplanted emotion, the PM and AIM exhibit a better performance for the transplanted emotion than the other models. The PM performs better in the RMSE of log F0 and the correlation coefficient of log F0, whereas the AIM performs better in the mel-cepstral distortion. The SM always exhibits the worst performance in comparison with the other two models. In terms of trained emotion, the SD did not perform better than the proposed models. The main reason for this could be the amount of training data. The SD was trained only using data uttered by a particular speaker, whereas the proposed models were trained using 24 speakers. Finally, the SM performs better in trained emotion, unlike its tendency in the transplanted emotion.

C. Subjective evaluation

To evaluate the performances for transplanted emotion, subjective evaluation was performed for naturalness, speaker similarity, and emotion reproduction. The subjective evaluations included a mean opinion score (MOS) test and two types of degradation mean opinion score (DMOS) test. Thirteen Japanese listeners participated in each subjective test.

For the MOS test, we used synthesized speech generated from three models: the PM, SM, and AIM. In addition, as a reference speech, we used vocoded speech by STRAIGHT (ST) and synthesized speech generated from a trained-emotion model (TR). From the results of our objective evaluation, the SM architecture was adopted as the TR. Sixty sentences (15 sentences covering two emotions from two speakers) were synthesized using each method. The order of the test speeches was randomly selected and remained the same for each listener. A five-point scale (from 1: very unnatural to 5: very natural) was adopted for MOS.

In the DMOS tests, the quality of speech synthesized by the PM, SM, AIM, and TR was compared to the ST in terms of speaker similarity and emotion reproduction. Forty sentences (10 sentences covering two emotions from two speakers) were synthesized for each pair. The order of the test speeches was randomly selected and remained the same for each listener. Five-point scales were used to judge the similarity DMOS and the reproduction DMOS (from 1: very dissimilar to 5: very similar; and from 1: never reproduced to 5: completely reproduced, respectively).

Fig. 8 shows the experimental results for naturalness, speaker similarity, and emotion reproduction. In terms of naturalness, the difference between the TR and ST is approximately 1.0, and the difference between the TR and PM is approximately 0.5. Therefore, the degradation of naturalness is small, although the PM did not use the emotional speech uttered by the target speaker for training. The results for the AIM are slightly poor than for the PM. However, the SM showed the poorest value of approximately 1.7, so we can assert that the naturalness is greatly decreased in the transplanted emotion synthesized by the SM. Judging by the poor performance of the SM, it is evidently difficult to separate the speaker factors from the emotional factors in a layer, i.e., the SM fails to filter out speaker factors before passing features to the next emotion output layer. The architecture of the SM could be too tightly constrained to train. On the other hand, the PM can be more loosely constrained by simply adding speaker factors and emotional factors. The TR gives the highest performance among all the models. This is mainly because the TR is trained by using the target speaker, and therefore, the TR contains not only the emotional expressions regularly used by all speakers, but also those emotional expressions that are specific to the target speaker. We believe that speaker-specific emotional expressions are a different case altogether and are beyond the scope of this study.

In terms of speaker similarity, the TR performs better than the PM and AIM. One possible reason for this is that the

emotional expressions depend on the speaker. The listeners may struggle to recognize the speaker-dependent emotional expressions from the synthesized speech. However, the PM shows a better performance than the SM. This could indicate that the PM architecture can separately handle speaker and emotional factors (to some extent). In other words, explicitly handling speaker and emotional factors might not have any negative effects on the performance of the PM architecture. However, the PM is worst than the TR. Because of this, we will endeavor to investigate the improvement of the PM as a future work.

In terms of emotion reproduction, the PM outperformed the AIM. This result indicates that the PM architecture could successfully learn emotional factors using emotion ID; however, the AIM fails to do so. Judging from the results, we can say emotional factors were dealt with in the output layer of the PM architecture. The SM produced bad evaluation scores with regard to both speaker similarity and emotion reproduction.

In summarizing subjective evaluation, in terms of transplanted emotion, the PM demonstrated the best performance compared with the other two models. The AIM exhibited almost similar performance as the PM in speaker similarity; however, the same was not observed in case of emotion reproduction. This difference could be attributable to both the different structures of the models and the amount of training data that was used. In terms of the speaker, the amount of training data was balanced equally among all speakers. However, in terms of emotion, the amount of neutral emotion was approximately twenty times that of both the sad and joyful emotions. Since the PM consists of different output layers for three emotions, and the layers are selected by the emotion IDs, the PM is explicitly trained for each emotion. Therefore, the PM remains unaffected by the proportions of emotional data received. In contrast, all mappings in the AIM architecture are implicitly distributed over the entire DNN, so the AIM might be affected by the proportions of emotional data received.

IV. Conclusions

In this paper, to synthesize emotional speech, we investigated the performance of three DNN architectures, i.e., parallel, serial, and auxiliary input models. The experimental results showed that, in terms of transplanted emotion, the PM and the AIM demonstrated good performances compared with the SM. In contrast, in terms of trained emotion, the SM demonstrated the best performance compared with the AIM and PM. We believe that, to increase the flexibility and controllability of TTS, the DNN architecture must demonstrate a good performance for transplanted emotion. Therefore, the PM and the AIM are preferable in this regard. In addition, we believe that a DNN architecture should have structures that correspond explicitly to speaker and emotional factors. Therefore, we propose the PM as the best of the three models.

In future, we would like to investigate the potential and possible improvement of the PM, e.g., via interpolations for emotional expressions and speaking styles, or speech morphing to gradually change speech qualities in the time domain. In this paper, we only evaluated data for a female speaker, so we would also like to evaluate transplanted emotion for s male speaker. We would also like to consider the number of hidden layers and units per layer in the proposed model architectures. Besides, the future work will include an investigation of new model architectures (e.g. switching around the place of the speaker and emotion in the SM). Applying the proposed model architecture to duration modeling for emotional speech will also be investigated.

References

[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2013 IEEE International Conference on. IEEE, pp. 7962–7966, 2013.

[2] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on. IEEE, pp. 3829–3833, 2014.

[3] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, pp. 4460–4464, 2015.

[4] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, pp. 4475–4479, 2015.

[5] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. INTERSPEECH*, pp. 879–883, 2015.

[6] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proc. INTERSPEECH*, pp. 2278–2282, 2016.

[7] H. Luong, S. Takaki, G. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2017 IEEE International Conference on. IEEE, pp. 4905–4909, 2017.

[8] B. Li, and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proc. INTERSPEECH*, pp. 2468–2472, 2016.

[9] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," in *IEICE TRANSACTIONS on Information and Systems*, vol. E88-D, no. 3, pp. 502–509, 2005.

[10] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," in *IEICE TRANSACTIONS on Information and Systems*, vol. E88-D, no. 11, pp. 2484–2491, 2005.

[11] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," in *IEICE TRANSACTIONS on Information and Systems*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[12] H. Kanagawa, T. Nose, and T. Kobayashi, "Speaker-independent style conversion for HMM-based expressive speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2013 IEEE International Conference on. IEEE, pp. 7864–7868, 2013.

[13] L. Trueba, R. Chicote, J. Yamagishi, O. Watts, and J. Montero, "Towards speaking style transplantation in speech synthesis," in *8th ISCA Speech Synthesis Workshop*, pp. 159–163, 2013.

[14] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.1, pp.66–83, 2009.

[15] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *Proc. INTERSPEECH*, pp. 274–278, 2015.

[16] H. Kawahara, I. M-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of repetitive structure in sounds," in *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.