Layer adaptation for transfer of expressivity in speech synthesis

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Layer adaptation for transfer of expressivity in speech synthesis

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{ajinkya.kulkarni, vincent.colotte, denis.jouvet}@loria.fr

Abstract

Expressive speech synthesis using parametric approaches is constrained by the style of the speech corpus used. In this paper, we present the development of an expressive speech synthesis for a new speaker voice without requiring a specific recording of expressive speech by new speaker. We propose deep neural network based layer adaptation framework for transferring the expressive characteristics of speech to a new speaker's voice for which only neutral speech data is available. The focus of the work is on investigating transfer learning mechanism, which will accelerate the efforts towards exploiting existing expressive speech corpus. Experiments using expressive Caroline speech corpus and neutral Lisa speech corpus shows layer adaptation technique is able to transfer expressive characteristics while keeping the speaker's style characteristics.

Keywords: expressive speech synthesis, deep learning, transfer learning, domain adaptation, emotion

1. Introduction

Every sentence spoken by a human inherently possesses expressiveness, which is conditioned on the pragmatics behind the content of the speech (McCready, 2014). Current state-of-the-art speech synthesis system uses deep neural network based architectures and heavily depends on the speech corpus used to train the deep neural networks (Schröder, 2009; Wang et al., 2017a; Zen et al., 2009). Thus, to build expressive speech synthesis for a new speaker, one has to create a speech corpus with various emotions. It is inconvenient to record a speech corpus every time we want to build an expressive speech synthesis system for a new speaker's voice. In order to reuse the existing data on different users, we need expressivity adaptation techniques, or more precisely expressivity transfer techniques. In this paper, we present the development of an expressive speech synthesis system for a new speaker without requiring explicit recording of expressive speech data from this speaker.

Several approaches have been proposed for artistic style transfer in images, while retaining the semantic structure of images (Jing et al., 2017; Li et al., 2017; Gatys et al., 2016). However, few approaches are explored for prosodic style transfer for speech. A significant amount of work has been done on speaker adaptation with small amount of available speaker speech data such as i-vector as a speaker embedding, feature space transformation and Latent Hidden Unit Contribution (LHUC) (Wu et al., 2015; Potard et al., 2015). Parker et al. (2018), proposed expressivity transplantation as an extension to speaker adaptation using LHUC units. An extension to Tacotron was proposed by Wang et al,. (2017b; 2017c) as style tokens, which learn a latent embedding of prosody, derived from a reference acoustic representation containing the desired prosody. Tacotron system requires considerable amount of speech data (~100 hrs) with large computational resources, this creates a bottleneck in the development of speech synthesis systems. For controlling the expressivity in speech synthesis, Akuzawa et al. (2018) incorporated the voiceLoop (Taigman et al., 2017), an autoregressive speech synthesis model with variational autoencoder conditioned on a text. Due to intractability in variational inference, the above approach faces difficulties in fine-tuning to learn a meaningful representation in latent space. With advancement in computational resources and deep learning, large data sets are available in image, text, and speech, but it is still frequent that for a specific application there is not enough data or no data at all. Hence, approaches are investigated to transfer the learned knowledge representation from one task to another related tasks (Pan and Yang, 2009). Recent studies showed that features represented through hidden layers, progress from general to task-specific along the network (Yosinski et al., 2014). Therefore, we proposed domain adaptation technique, in which different layers are adapted with acoustic features from different speakers and different emotions, for transferring the expressivity to a new speaker for which only neutral speech data is available.

In proposed work, we predicted expressive acoustic features for a new speaker from layer adapted model as well as from baseline speech synthesis model. Our experiments shows that predicting components of acoustic features from different models yields significant improvement in transferring the expressivity while retaining the neutral speaker's characteristics. This paper is structured as follows: Section 2 presents our methods and in section 3 describes the experimentation details including speech corpus and experimentation setup. Section 4 obtained results through experimentation are discussed using subjective evaluation and objective evaluation.

2. Method

2.1. Baseline speech synthesis

For the baseline speech synthesis system, we use feedforward neural networks for modeling duration, fundamental frequency, aperiodicity, and spectral features, as detailed in (Zen, 2013; Wu et al., 2016). For preprocessing, the text is first converted into a sequence of contextual labels, which corresponds to linguistic, phonetic and prosodic information about phoneme. The duration model is explicitly trained along with the acoustic model to predict the exact number of acoustic frames per speech segment.

In the training phase, the duration model is trained using the context label features for each phoneme to predict duration information. Then, contextual label information along with duration information is used as input to the acoustic model for predicting output acoustic features namely spectral features, aperiodicity, log of fundamental frequency and voiced-unvoiced flag. Afterward, denormalization is performed on generated acoustic features with mean and pre-computed variances from training data. Finally, the predicted acoustic features are given to vocoder for synthesizing speech waveform.

2.2. Layer adaptation framework

Yosinski et al., (2014), reported that first layers (i.e., close to input) in deep neural networks are able to capture a global representation of input data distribution which is not a task specific. As we move towards higher layers, hidden representation of layers tends to focus on task-specific features. We proposed a layer adaptation approach in which we adapt a limited set of the layers on neutral acoustic features corresponding to the new speaker's voice, and another set of layers on the expressive acoustic features (of another speaker).

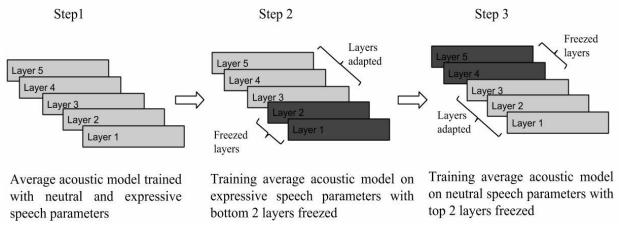

Fig. 1: Layer adaptation technique for transfer of speaker's style characteristics and prosody information.

While updating the weights in adaptation step 2 and step 3 in Fig. 1, no modifications are applied to the frozen layers. This is a popular technique used in transfer learning to freeze part of base model and thus to adapt subset of model parameter to target data distribution (Fatemeh and Ali, 2014).

In the proposed layer adaptation framework, we describe the source domains $D_{A,\,expressive}$ as an expressive speech domain for speaker A and $D_{B,\,neutral}$ as neutral speech synthesis domain for speaker B. The target domain is described by $D_{B,\,expressive}$ as an expressive speech domain for speaker B for which expressive speech is not available. For source domains $D_{A,\,expressive}$ and $D_{B,\,neutral}$, we build baseline speech synthesis model, $M_{A,\,expressive}$ and $M_{B,\,neutral}$, trained using contextual labels and acoustic features represented as $\{x_A, y_{A,\,expressive}\}$ and $\{x_B, y_{B,\,neutral}\}$. The main objective of this work is to build a model, $M_{B,\,expressive}$, which will generate acoustic features, $y_{B,\,expressive}$ for given contextual label features, x_B. For minimization of discrepancy, we extracted the knowledge regards to speaker's identity and expressive information by adapting layers on training data $\{x_A, y_{A,\,expressive}\}$ and $\{x_B, y_{B,\,neutral}\}$ selectively.

The main goal of domain adaptation approach is to allow models to reduce the domain discrepancy between source and target as well as adapt the models to target domain data distribution. In the proposed work, focus is on acoustic features for layer adaptation. First, we jointly trained average duration model and average acoustic model, $M_{average}$ with $\{x_A, y_{A,\,expressive}\}$ and $\{x_B, y_{B,\,neutral}\}$ from both source domains $D_{A,\,expressive}$ and $D_{B,\,neutral}$. Second, top n layers are trained with $\{x_B, y_{B,\,neutral}\}$ contextual labels from speaker B. Third, bottom n layers trained with $\{x_A, y_{A,\,expressive}\}$ contextual labels from speaker A.

In synthesis phase, we predict output contextual labels, $y_{B,\,expressive}$ for target domain, $D_{B,\,expressive}$ using layer adapted acoustic model $M_{B,\,expressive}$, and duration model $M_{average}$. Thereafter, we also proposed to generate acoustic features $y_{B,\,expressive}$ by predicting spectral features (mel generalized cepstrum coefficient, mgc) from layer adapted model $M_{B,\,expressive}$, while the other acoustic features (fundamental frequency, aperiodicity and voiced-unvoiced) are predicted from baseline expressive speech synthesis model, $M_{A,\,expressive}$, as shown in Table 3. In the next section, we will discuss the details of various experimentation conducted with layers adapted and prediction of acoustic features for expressive speech with speaker's style of speaker B.

3. Experimentation

3.1. Speech datasets

We worked with two speech corpora, namely Lisa neutral speech corpus and Caroline expressive speech corpus in the French language recorded from two female speakers. Lisa corpus is a neutral speech corpus with 1815 utterances approximating 3 hrs of recording. Caroline expressive speech corpus consists of several emotions, namely joy, surprise, fear, anger, sadness and disgust. In this paper, we used only anger emotion as expressive

speech, which is approximating 1 hr of recording with 500 utterances. All the speech signals were used at a sampling rate of 16 kHz.

In this paper, we parameterized speech using WORLD vocoder (Morise et al., 2016) with 187 acoustic features every 5 milliseconds time frame, namely 180 spectral features as mel generalized cepstrum coefficient (mgc), 3 log fundamental frequencies (lf0), 3 band-aperiodicities (bap) and 1 value for voiced-unvoiced information (vuv). Based on the mean and standard deviation values, the acoustic features extracted from the WORLD vocoder were z-normalized. Each speech corpus is divided into train, validation and test sets in the ratio of 80%, 10%, 10% respectively. To show the transfer of expressivity, we used Caroline anger speech as speaker A and Lisa neutral speech as speaker B.

Method	MCD	F0 RMSE	VUV	MOS
Merlin neutral	5.625	27.41	16.83	3.18 ± 0.50
DNN neutral	5.376	21.50	23.05	2.50 ± 0.53
Merlin anger	5.426	41.63	9.23	2.42 ± 0.45
DNN anger	5.894	23.90	8.33	2.68 ± 0.49

Table 1: Average measure of distortion for different acoustic features, MCD (dB), F0 RMSE (Hz), VUV (%) and MOS scoring of quality of test utterances.

Method	Adapted layers	Expressive MOS	Speaker MOS
model I	L1, L2: anger, L4, L5: neutral	1.53 ± 0.41	2.69 ± 0.54
model II	L1, L2: neutral, L4, L5: anger	1.63 ± 0.44	2.68 ± 0.48

Table 2: Subjective evaluation for layer adapted models using MOS scoring of expressivity and of speaker identity for anger emotion in Lisa voice.

Method	mgc model	lf0 model	Expressive MOS	Speaker MOS
model III	model I	DNN anger	2.01 ± 0.50	1.63 ± 0.42
model IV	model II	DNN anger	1.98 ± 0.56	1.71 ± 0.40

Table 3: Subjective evaluation for acoustic features predicted through layer adapted model and baseline model.

3.2. Experimental setup

Using a deep learning framework, we modeled the duration and acoustic features of the baseline speech synthesis system. We used a feedforward neural network with 5 hidden layers, each having 512 units with hyperbolic tangent as an activation function. In this work, we used Adam optimizer with a learning rate of 0.001 (Kingma and Ba, 2014). We trained each model till 25 iterations, mini-batch size of 64 frames and a dropout rate of 0.1. We compared our implementation of baseline speech synthesis system with Merlin (Wu et al., 2016), an open source speech synthesis toolkit trained with the model configuration as mentioned above.

For performing layer adaptation, first, we created an average acoustic model trained with utterances from Caroline anger and Lisa neutral. We used the same configuration as of the baseline speech synthesis for an average acoustic model. Afterward, we adapted the acoustic model using neutral and anger acoustic features consecutively on top-n and bottom-n layers, as depicted in step 2 and step 3 of Fig. 1. We experimented with various number of adapted layer (n = 2, 3, 4). After an informal listening test on synthesized speech samples from layer adapted models with variations of number of adapted layers. We selected models in which 2 layers are adapted on acoustic features. Thereafter, we built model I, model II using combinations of neutral (Lisa) acoustic features and anger (Carolina) acoustic features adapted on top and bottom layers respectively (step 2 and step 3 in Fig 1.), as stated in Table 2. In second proposed work, we predicted mgc features from layer adapted acoustic model (model I and model II), while lf0, bap, vuv features were predicted from anger baseline model denoted as DNN anger. We presented the experimentation with different adapted models in Table 2 and Table 3.

4. Results

4.1. Objective evaluation

We conducted error measures of distortions on test data between reference acoustic features and those generated by baseline acoustic models Merlin neutral, Merlin anger and DNN neutral, DNN anger, where DNN is referred to our baseline feedforward neural network implementation. For objective evaluation, we used mel cepstrum distortion (MCD), root mean square error (RMSE) of fundamental frequency, voiced-unvoiced prediction accuracy as shown in Table 1.

We compared our implementation of baseline speech synthesis system with Merlin, an open source speech synthesis toolkit trained with the same model configuration, refer Table 1. These results suggest that the quality of the implementation of the baseline model is comparable to the Merlin based speech synthesis on both anger and neutral voice.

As no reference acoustic features are present for expressive speech in Lisa's voice to conduct objective evaluation, we carried out a subjective evaluation for expressive (anger) speech synthesis in Lisa's voice, which we will discuss in next section 4.2.

4.2. Subjective evaluation

We conducted Mean Opinion Score (MOS) listening test for subjective evaluation of baseline speech synthesis (Streijl et al., 2016). In this listening test, participants

have to rate the stimuli on a scale of 1 to 5, where 1 is bad

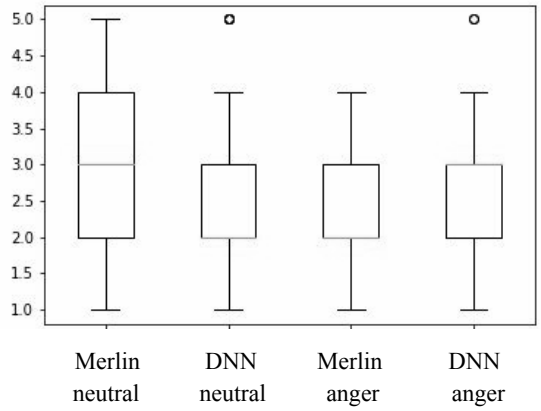

Fig. 2: Boxplot of MOS score for baseline speech synthesis

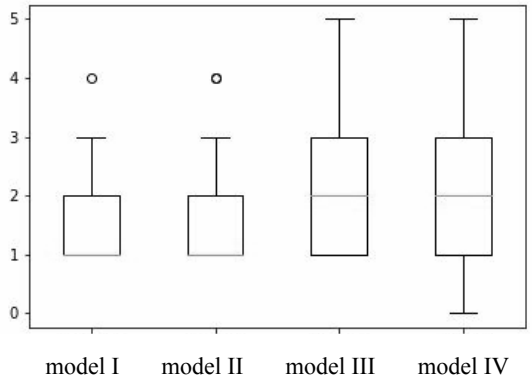

Fig 3: Boxplot of expressive MOS for layer adapted models

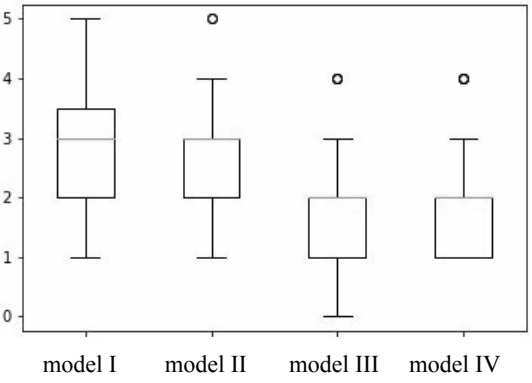

Fig 4: Boxplot of speaker MOS for layer adapted models

and 5 is excellent. We performed a subjective evaluation for baseline speech synthesis developed with neutral (Lisa) voice and anger (Caroline) voice. We compared our baseline implementation of speech synthesis with Merlin based speech synthesis trained on same experimental setup. In this test, there were 7 French listeners who scored 10 sets of stimuli for each model, randomly selected from the test set. Results of MOS scores together with their 95% confidence interval (CIs) are stated in Table 1. and box plots in Fig 2. It is observed that DNN anger model is slightly better than Merlin based speech synthesis model. For a neutral speech, participants preferred the Merlin based baseline than our DNN neutral model implemented in Pytorch framework (Paszke et al., 2017).

For subjective evaluation of layer adapted acoustic models (model I, model II, model III and model IV), we used two mean opinion scores, first expressive mean opinion score in which participants were asked to evaluate how much anger emotion characteristics of Caroline's voice are transferred in the synthesized anger speech on the scale of 1 being bad and 5 the best stimuli having anger expressive characteristics. Second, speaker mean opinion score, where participants were asked to provide a score based on how much Lisa's voice quality is present in stimuli on the scale of 1 (bad) to 5 (best).

The participants were instructed to give two scores from 1 to 5, first to evaluate anger characteristics in stimuli compared to reference anger stimuli, and second to evaluate speaker characteristics in stimuli compared reference speaker stimuli from Lisa speech samples. Addition to this, in the listening test, reference anger stimuli (for expressive MOS) and reference Lisa speaker stimuli also presented during subjective evaluation. We conducted above listening test with same setup as used in baseline MOS evaluation test with a set of stimuli from model I, model II, model III and model IV (models to be evaluated).

Box plots for expressive MOS and speaker MOS are shown in Fig 3, Fig 4, and expressive MOS and speaker MOS scores are given in Table 2, Table 3. Expressive MOS scores of model III and model IV are comparably better than model I and model II, which indicates predicting prosodic features from differently trained models helps the adaptation process. On other hand, model I and model II were able to achieve the comparably better speaker MOS scores compared to model III and model IV. This indicates that there is a trade-off between two proposed approaches about the way acoustic features are being predicted. Results of subjective evaluation for expressive MOS and speaker MOS reflects that predicting components of acoustic features from layer adapted model and DNN anger model, improved the anger characteristics in synthesized speech with Lisa's voice. Results of subjective evaluation for expressive MOS and speaker MOS reflects that predicting components of acoustic features from layer adapted model and DNN anger model, improved the anger characteristics in synthesized speech with Lisa's voice.

5. Conclusion

We presented layer adaptation approach which is similar to domain adaptation, where layers are adapted to emotional and neutral speech corpus iteratively. The subjective results showed that the usage of acoustic features predicted from different models (DNN anger and layer adapted model) enhance the performance of

transferring the expressive characteristics than using single layer adapted model.

There is a trade-off between knowledge transfer of expressivity and retaining the speaker's identity in synthesized speech. It is hard to identify which parameters of the neural network represent the attributes of speaker characteristics and of expressivity. Moreover, expressivity and speaker characteristics are bounded aspects of prosodic features. The work presented in this paper is a preliminary work and it laid the groundwork for future implementations for transferring the prosody information.

References

Akuzawa, K., Yusuke, I., and Yutaka, M., (2018). Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder. 10.21437/Interspeech.2018-1113, Hyderabad, India.

Fatemeh, D., and Ali, G., (2014). Minimizing the Discrepancy Between Source and Target Domains by Learning Adapting Components. Journal of Computer Science and Technology, Volume 29, pp. 105-115.

Gatys L. A., Ecker, A. and Bethge, S., M., (2016). Image Style Transfer Using Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 2414-2423.

Kingma, D. P., Ba, J., (2014). Adam: A Method for Stochastic Optimization. Journal: CoRR, volume: abs/1412.6980.

Jing, Y., Yang, Y., Feng, Z., Ye, J., Song, M.,, (2017). Neural Style Transfer: A Review. Journal: CoRR, volume: abs/1705.04058.

Li, Y., Wang, N., Liu, J., Hou X., (2017). Demystifying Neural Style Transfer. Journal: CoRR, volume: abs/1701.01036.

McCready, (2014). Expressives and Expressivity. Open Linguistics, Volume 1-1.

Morise, Masanori, Yolomori, F., and Ozawa, K., (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Transactions on Information and Systems E99.D.7, pp. 1877–1884.

Pan, S. J., Yang, Q. (2009). A Survey on Transfer Learning. IEEE Transaction on Knowledge and Data Engineering, pp. 1345-1359.

Parker, J., Stylianou, Y., Cipolla, R., (2018). Adaptation of an Expressive Single Speaker Deep Neural Network Speech Synthesis System. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5309-5313.

Paszke, A., Gross, S., Chintala, S., Chanan, G. Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., (2017). Automatic differentiation in PyTorch. NIPS-W 2017.

Potard, B., Motlicek, P., Imseng, D., (2015). Preliminary work on speaker adaptation for DNN based speech synthesis. Idiap-RR-02-2015.

Schröder, M., (2009). Expressive Speech Synthesis: Past, Present, and Possible Futures. In: Tao J., Tan T. (eds) Affective Information Processing pp 111-126, Springer, London.

Streijl, Robert, C., Winkler, S., Hands, D. S., (2016). Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. Multimedia System. Volume 22.2, pp. 213–227.

Taigman, Y., Wolf, L., Polyak, A., Nachmani, E., (2017). VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. Journal: CoRR, volume: abs/1707.06588.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., Saurous, R. A., (2017a). Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. Journal: CoRR, volume: abs/1703.10135.

Wang, Y., Skerry-Ryan, R. J., Xiao, Y., Stanton, D., Shor, J., Battenberg, E., Clark, R., Saurous, R. A., (2017b). Uncovering Latent Style Factors for Expressive Speech Synthesis. Journal: CoRR, volume: abs/1711.00520.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg E., Shor J., Xiao, Y., Ren, F., Jia, Y., Rif, A., Saurous, (2017c). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. Journal: CoRR, volume: abs/1803.09017.

Wu, Z., Swietojanski, P., Veaux, C., Renals, S., King, S., (2015). A study of speaker adaptation for DNN-based speech synthesis. In Proceedings of Interspeech 2015.

Wu, Z., Watts, O., King, S., (2016). Merlin: An Open Source Neural Network Speech Synthesis System. In Proc. 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, USA.

Yosinski, J., Clune, J., Bengio, Y., Lipson H., (2014). How transferable are features in deep neural networks? Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, pages 3320--3328 Montreal, Canada.

Zen, H., Senior, A., Schuster, M., (2013). Statistical Parametric Speech Synthesis Using Deep Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, DOI: 10.1109/ICASSP.2013.6639215, Vancouver, BC, Canada.

Zen, H., Tokuda, K. and Black A., (2009). Review: Statistical parametric speech synthesis. Speech Communication archive, Volume 51, Issue 11, pp 1039-1064.