ADAPTATION OF AN EXPRESSIVE SINGLE SPEAKER DEEP NEURAL NETWORK SPEECH SYNTHESIS SYSTEM

Jonathan Parker[1,2]*, Yannis Stylianou*[2]*, Roberto Cipolla*[1,2]

[1]Department of Engineering, University of Cambridge, UK
[2]Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK

ABSTRACT

One of the advantages of statistical parametric speech synthesis is the ability to alter some of the characteristics of the speech e.g. change the speaker, expression etc. In this paper we present a technique to adapt an expressive single speaker deep neural network (DNN) speech synthesis model to a new speaker, allowing for both neutral and expressive speech in the new speaker's voice. Experiments show that the proposed adaptation technique achieves higher MOS scores on both neutral and expressive speech, and higher speaker similarity and slightly lower expression similarity scores on the expressive speech when compared with another DNN speaker adaptation technique.

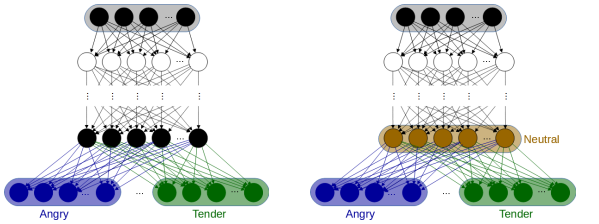
Index Terms— DNN, expressive speech, expressive speaker adaptation, expression transplantation

1. INTRODUCTION

Recently the success of DNNs in speech synthesis has been demonstrated [13, 6, 4]. Furthermore, techniques for DNN TTS speaker adaptation have been proposed. In [11], Swietojanski et al proposes a model-based adaptation technique, known as Learning Hidden Unit Contributions (LHUC). While originally proposed in the context of speaker adaptation for speech recognition, [12] successfully use LHUC for speaker adaptation for TTS. In [3, 10, 5], multiple speakers are used to train a single network, with all speakers sharing the hidden layers but having speaker specific outputs. [2, 7, 12] use speaker codes to identify different speakers. In [14], a linear adaptation layer (expressed as a product of two low rank matrices) is inserted into the network and adapted to a new speaker.

Expression transplantation is an extension to speaker adaptation, where the expressions from one speaker are transferred to a neutral speaker. In [8], this is done by constructing an eigenvoice space of neutral speech and the differences between neutral and expressive speech. In [1], two distinct cluster adaptive training (CAT) subspaces are created, one representing speakers and another representing expression. To transplant the expression from one speaker to another, the expressive utterance is projected into the expression subspace to obtain the appropriate expressive CAT weights while the neutral utterance is then projected into the speaker subspace to obtain the appropriate speaker CAT weights.

This work seeks to achieve the following: adapt an expressive single speaker DNN TTS model to a new speaker using only neutral speech, such that the adapted model will not only produce neutral speech that sounds like the new speaker but also produce expressive speech that sounds like the new speaker with appropriate expressiveness i.e. expression transplantation. Obtaining expressive speech can be challenging and expensive, so it is useful to adapt an expressive single speaker model. There are two challenges that distinguish this task from works mentioned above. First, the adaptation is performed on a DNN that had only been trained on a single speaker. In the above systems, models are trained on multiple speakers, and therefore learn more generic feature detectors that are useful for the speech of multiple speakers. Thus repurposing a single speaker DNN to a new speaker is more challenging. Secondly, when trained one expressive speech from a single speaker, the distinction between what is specific to the speaker and what is specific to the expression is unclear. This is in contrast to systems where there are multiple speakers with each expression.



(a) DNN with neutral and expression outputs in parallel.

(b) DNN with expression outputs following sequentially from neutral output

Fig. 1: DNN regressing normalised linguistic features to normalised acoustic features with different outputs for each expressive class

2. EXPRESSIVE SINGLE SPEAKER MODEL

2.1. Single speaker model

We begin with a single speaker model similar to that presented in [9], a single speaker, expressive DNN-based model, refer to [9] for details. In short, linguistic features are extracted from the text, acoustic features are extracted from the speech, these feature sets are normalised and a DNN learns to maps linguistic features to acoustic features. The DNN model has a separate output for each of the 6 expression classes, namely, angry, fearful, happy, neutral, sad and tender. This is shown in Figure 1a. Significantly, the expressive outputs are in parallel to the neutral output. The results of this model (albeit as part of a larger visual speech synthesis model) were evaluated in [9] and were shown to be superior to HMM models, with clear, expressive speech.

2.2. Expression specific normalisation

Normalisation is performed separately for each of the expression classes: μ_n and σ_n for the neutral speech and $\{\mu_{e_1}, \mu_{e_2}, \ldots, \mu_{e_n}\}$ and $\{\sigma_{e_1}, \sigma_{e_2}, \ldots, \sigma_{e_n}\}$ for each of the expressive speech classes.

This allows an estimation of the expression specific means and standard deviations for a new neutral speaker: $\mu_{e_i}^{(2)} = \mu_n^{(2)} + \left(\mu_{e_i}^{(1)} - \mu_n^{(1)}\right)$ and $\sigma_{e_i}^{(2)} = \sigma_n^{(2)} \cdot \frac{\sigma_e^{(1)}}{\sigma_n^{(1)}}$, where the superscript indices denote the speaker. Informal testing revealed that this greatly enhanced the expressiveness of the adapted speaker.

3. ADAPTATION USING LHUC

3.1. Preliminary experiments and analysis

The first attempt to perform speaker adaptation was using LHUC, refer to [11] for details. Because the adaptation data is labelled, it is possible to optimise the LHUC weights using backprop to minimise the squared error over the adaptation set between the target normalised acoustic features and the neutral output. The DNN was adapted twice, in separate experiments, once with a British female speaker and then with a British male speaker. 100 utterances of each was used for adaptation.

An informal examination of the test utterances showed that the algorithm was somewhat successful with neutral speech, however, the performance on expressive speech was very poor with significant artefacts and audible distortions.

3.2. Sequential Expression

With the understanding that the artefacts identified above were due to a lack of disentanglement between speech modelling and transforming neutral speech to expressive speech, the architecture is modified by placing the expressive speech outputs sequentially after the neutral speech output, as opposed to in parallel with it. A non-linearity (sigmoid) is placed after the neutral speech output, with the input to expressive speech layers coming from this non-linearity. This is demonstrated in Figure 1b.

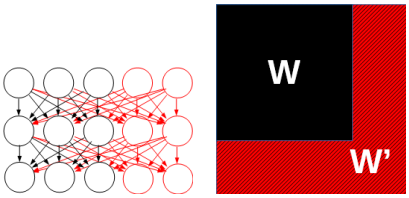
3.3. Preliminary experiments and analysis

Again, LHUC weights were applied to all hidden neurons and trained using backprop. An informal comparison of the expressive results between using parallel and sequential expressive outputs shows that there is a significant decrease in the speech artefacts and distortions. However, there is a significant difference in performance of the expressive speech between the male and female speakers, which stems from the fact that LHUC is being used to adapt a single speaker model, trained on a female speaker, not an average voice mode, trained on multiple speakers which is likely to have more generic features for modelling the speech of different speakers. However, LHUC "does not change the learned feature detectors" [11, p. 173]. Thus if there are feature detectors that are required by the new speaker that are not present in the pretrained network, the performance of LHUC will degrade significantly. This would seem to be the case where one speaker is male and the other female.

4. ADAPTATION USING HIDDEN LAYER AUGMENTATION

4.1. Hidden Layer Augmentation

We introduce a novel speaker adaptation technique designed to overcome the shortcomings of LHUC. Furthermore, we combine this speaker adaptation technique with two different techniques of modelling expression. Instead of adjusting the contributions of the existing feature detectors, additional neurons are added to each hidden layer of the network. This is called Hidden Layer Augmentation (HLA). The additional neurons are fully connected to the preceding and succeeding layers; they can therefore learn new features and their activations are part of the receptive field of the succeeding layer. This is illustrated in Figures 2a and 2b. During adaptation, only the new weights, that is, the weights connected to the additional neurons are trained.



(a) DNN schematic (b) Extension of the weight matrices

Fig. 2: Hidden layer augmentation. Black elements are part of the original network, while red elements are part of the augmentation.

Intuitively, the augmentation of the hidden layers allows them to learn whatever the pretrained network has not already learnt, that is, the new features represented by the new neurons are thus able to capture that which the pretrained network does not. The pretrained network will go some way to predicting the correct acoustic features, but the new neurons will make up the difference between the original and new speaker, thereby reducing the error rate over the adaptation training set. In particular, this addresses the shortcomings outlined above when attempting to adapt a DNN trained on a single speaker. To perform speaker adaptation, backpropagation is used to train the additional, speaker-specifc weights, while the rest of the weights of the network are frozen.

4.2. Expression modelling

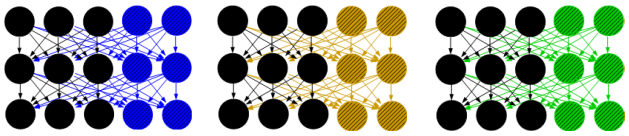
We combine this new method for speaker adaptation with two different techniques of modelling expression: using expression inputs (EI) and expression specific neurons (ESN). In contrast to the model of Sections 2.1 and 3.2, which models expression at the output, EI uses expression information at the input, while ESN uses expression information throughout the hidden layers of the model.

4.2.1. Expression Inputs

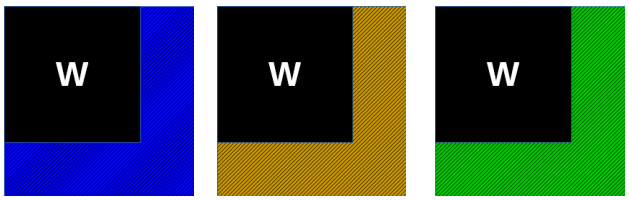
Using expression inputs simply means appending the linguistic features with an expression flag, a one-hot vector denoting the expression class. This approach is similar to [7]. Informal testing showed that the results when trained with the original expressive speaker are indistinguishable from those of Section 2.1. Applying HLA for speaker adaptation is as is described in Section 4.1. Typically the speech of the new speaker is neutral and therefore the neutral flag is set when performing speaker adaptation.

4.2.2. Expression Specific Neurons

In this technique the hidden neurons are divided into core and expression specific neurons. The core neurons behave as part of a standard DNN. However, each hidden layer has a set of neurons specific to a particular expression class. This is shown in Figure 3.



(a) Expression specific neurons and associated weights



(b) Section of weight matrix specific to each expression class.

Fig. 3: Expression Specific Neurons. Different coloured neurons and weights belong to different expression classes

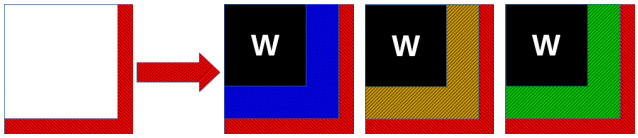


Fig. 4: Augmenting HLA weights (in red) to ESN weights modelling different expressions.

When data is propagated through the network, it will only pass through the neurons that are associated with the expression class of the data. Furthermore, the weights associated with the ESNs are only modified with data from the appropriate expression. The weights associated with the core neurons are modified with all the neutral and expressive speech data. By considering the case of two utterances with the same speech, (i.e. same linguistic information), one neutral and the other expressive, say "angry", the expressive utterance will pass through the same neurons as the neutral utterance, but will also pass through the "angry" set of ESNs, while not passing through the "neutral" set of ESNs. Therefore, replacing the neutral ESNs with an expressive set of ESNs can be thought of as a transformation of neutral speech to expressive speech. Informal testing showed that the results when trained with the original expressive speaker are indistinguishable from those of Section 2.1

Applying HLA to this model is straightforward. Because the speech of the new speaker is neutral, it uses the neutral ESNs. Once the HLA weights have been trained using backprop to achieve neutral speech, the HLA weights can be augmented to the ESN weights to achieve expressive speech. This is shown in Figure 4.

5. EXPERIMENTS

5.1. Experimental setup

Three models are evaluated:

- A: The model of Section 3.2, which uses multiple outputs to model different expressions and uses LHUC to adapt to a new speaker.
- B: The model of Section 4.2.1, which uses a one-hot vector representation of expressive class at the input to model different expressions and uses HLA to adapt to a new speaker.
- C: The model of Section 4.2.2, which uses ESNs to model different expressions and uses HLA to adapt to a new speaker.

The acoustic features used were 45 Mel-cepstral coefficients, logarithmic fundamental frequency, 25-band aperiodicities, and their first and second time derivatives, in addition to a voiced/unvoiced decision. The sampling rate of the speech recordings was 32kHz.

The base model is a DNN with six hidden layers, each with 1024 neurons, trained on 6591 utterances (735 angry, 696 fearful, 697 happy, 3078 neutral, 691 sad and 694 tender utterances) of a female native British English speaker. The neuron output nonlinearity is sigmoid.

Adaptation on all three models is performed with two separate experiments, one with adaptation to a female British English speaker and one with a male British English Speaker, using 100 neutral utterances for adaptation in each case.

The LHUC weights of model A are optimised using 25 iterations of backprop at a learning rate of 0.1. The HLA weights of models B and C are optimised using 25 iterations of backprop with a learning rate of 0.001 and a large L_2-norm weight penalty of 0.1.In model C, 960 core neurons and 64 expression specific neurons were used in each hidden layer. For models B and C, in performing speaker adaptation, each hidden layer was extended by 128 neurons. Some speech samples can be heard by visiting http://mi.eng.cam.ac.uk/~jwp37 .

	Baseline	A	B	C
Mel-cepstra (dB)	5.20	5.93	5.68	5.53
F_0	31.2	36.7	28.7	32.2
Band Aperiodicity	0.446	0.529	0.516	0.483
V/UV errors (%)	4.73	7.439	5.30	6.41

Table 1: Average measure of distortion for different acoustic features on male and female neutral test utterances.

5.2. Quantitative Evaluation

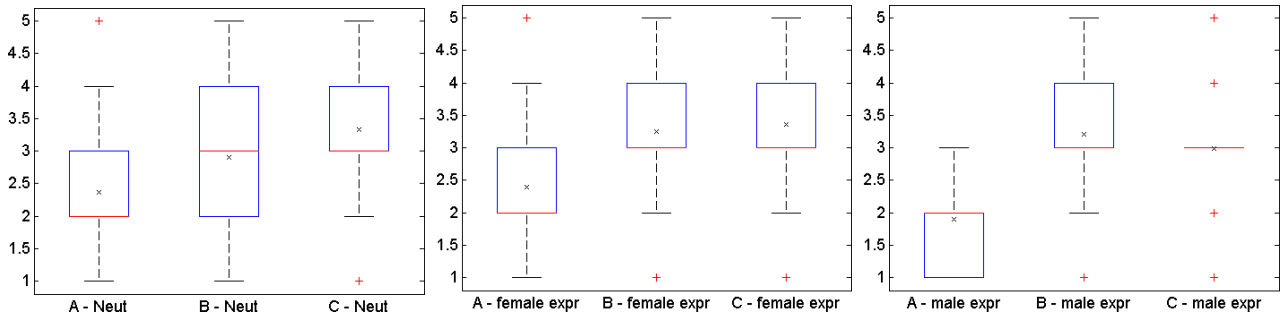
Table 1 shows the objective measures of distortion on some test data between the original acoustic features are those synthesised by the models in question. This can only be performed on neutral speech as there is no expressive speech data for these speakers. These are compared against a baseline results of a single-speaker DNN that is trained on 1024 utterances of each speaker in turn. These results suggest the superiority of HLA over LHUC in for neutral speaker adaptation in this case of adapting a single speaker DNN.

5.3. Qualitative Evaluation

The neutral and expressive adapted speech is assessed in three qualitative experiments: mean opinion scoring, speaker similarity and expression similarity.

5.3.1. Mean Opinion Scoring

Test subjects were presented with 10 neutral utterances and 20 expressive utterances (4 utterances from the 5 expression classes) from each adapted speaker from each of the three models. Test subjects were asked to assess the quality of the speech on a 1-5 scale. 8 test subjects were used. Results are shown in box plots in Figure 5 and the means are given in Table 2. As a reference, the baseline results of the single speaker DNNs are included.



(a) Neutral speech

(b) Female expressive speech

(c) Male expressive speech. The box plot for system C is flat because the 25-th and 75-th percentile are both 3.

Fig. 5: Box plots of the MOS

	A	B	C
Neutral	2.369	2.907	**3.325**
Expressive - female	2.389	3.252	**3.357**
Expressive - male	1.892	**3.200**	2.993

Table 2: Mean score of each system for neutral and expressive speech

	A	B	C	Neither
A vs B	30.4 ± 5.6	55.4 ± 6.0	-	14.2 ± 4.2
A vs C	33.3 ± 8.4	-	50.0 ± 9.0	16.7 ± 6.7

Table 3: Speaker Similarity (%). Which system did test subjects find closer to the adapted speaker (with 95% confidence intervals).

These results show that the quality of the synthesised speech from the models using HLA is consistently superior to those from models using LHUC.

5.3.2. Speaker similarity

Test subjects were asked to compare the expressive speech results from models A, B and C, in terms of their fidelity to the intended speaker. This was done in an ABX test with two expressive utterance from the different models being compared against a neutral reference. 20 expressive utterances (4 utterances from the 5 expression classes) from each adapted speaker were used. 8 test subjects were used. Models A and B are compared and models A and C are compared. The results are given in Table 3.

These results suggest that the expressive speech from HLA models is closer to the intended speaker than that from LHUC models.

5.3.3. Expression

Evaluating expression is a difficult and ill-defined problem. To assess the expressive speech results from models A, B and C, two expressive utterances from an adapted speaker from different models are compared with the original expressive utterance from the corpus used to train the original single speaker expressive model in an ABX test. Again, 20 expressive utterances (4 utterances from the 5 expression classes) from each adapted speaker were used. 10 test subjects were used. Models A and B are compared and models A and C are compared. The results are given in Table 4.

	A	B	C	Neither
A vs B	56.0 ± 4.6	37.9 ± 4.5	-	6.2 ± 2.2
A vs C	50.4 ± 4.5	-	41.0 ± 4.4	8.5 ± 2.2

Table 4: Expression Similarity (%). Which system did test subjects find closer to the desired expression (with 95% confidence intervals)

These results suggest that the expressive speech from LHUC model is closer to the intended expression than that from HLA models. This result is surprising and contradicts the findings of informal testing and is at odds with the result of Section 5.3 which indicated that the HLA models produced higher quality speech. However, this may be due to the difficulty that test subjects had in assessing the similarity of expression of two samples or indeed ranking which of two expressions is closer to a third expression. Therefore, different tests, such as expression classification (testing whether test subjects can correctly identify the expression in the synthetic speech), will be conducted to see if the expressions in the LHUC model is really closer to the desired expression than the HLA models. Furthermore, more test subjects will used.

6. CONCLUSION

We have presented a novel speaker adaptation technique for neural network based speech synthesis systems. We show that, in the case of adapting an expressive single speaker network, it produces higher quality neutral and expressive speech. Furthermore, the expressive speech from the adapted speaker has higher fidelity to the speaker but slightly lower fidelity to the expression. We think this is not a reflection on the results but rather a reflection on the difficulty with which people can judge the similarity of expressions. This merits further investigation into the quality of the expressions of the output speech. Furthermore, we intend on applying this technique to other neural network based speech synthesis models.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Norbert Braunschweiler for his advice and assistance in preparing the speech corpora.

8. REFERENCES

[1] Langzhou Chen, Norbert Braunschweiler, and Mark J. F. Gales. Speaker and expression factorization for audiobook data: Expressiveness and transplantation. *Trans. Audio, Speech and Lang. Proc.*, 23(4):605–618, April 2015.

[2] Rama Doddipatla, Norbert Braunschweiler, and Ranniery Maia. Speaker adaptation in dnn-based speech synthesis using d-vectors. 2017.

[3] Bo Fan, Lijuan Wang, Frank K. Soong, and Lei Xie. Photo-real talking head with deep bidirectional LSTM. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 4884–4888, 2015.

[4] Shiyin Kang, Xiaojun Qian, and Helen Meng. Multi-distribution deep belief network for speech synthesis. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 8012–8016, Vancouver, Canada, 2013. IEEE.

[5] Bo Li and Heiga Zen. Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis. 2016.

[6] Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Trans. on Audio, Speech & Language Processing*, 21(10):2129–2139, 2013.

[7] Hieu-Thi Luong, Shinji Takaki, Gustav Henter, and Junichi Yamagishi. *Adapting and Controlling DNN-Based Speech Synthesis Using Input Codes*, pages 1905–1909. IEEE, 6 2017.

[8] Yamato Ohtani, Yu Nasu, Masahiro Morita, and Masami Akamine. Emotional transplant in statistical speech synthesis based on emotion additive model. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 274–278, 2015.

[9] Jonathan Parker, Ranniery Maia, Yannis Stylianou, and Roberto Cipolla. Expressive visual text to speech and expression adaptation using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4920–4924, 2017.

[10] S. Pascual and A. Bonafonte. Multi-output rnn-lstm for multiple speaker speech synthesis and adaptation. In *European Signal Processing Conference*, pages 2325–2329, Sep 2016.

[11] P. Swietojanski and S. Renals. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *Proc. IEEE Workshop on Spoken Language Technology*, Lake Tahoe, USA, December 2014.

[12] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King. A study of speaker adaptation for dnn-based speech synthesis. In *INTERSPEECH*, pages 879–883. ISCA, 2015.

[13] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966, Vancouver, Canada, 2013.

[14] Yong Zhao, Jinyu Li, and Yifan Gong. Low-rank plus diagonal adaptation for deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 5005–5009, 2016.