EXPRESSIVE VISUAL TEXT TO SPEECH AND EXPRESSION ADAPTATION USING DEEP NEURAL NETWORKS

Jonathan Parker[1,2], *Ranniery Maia*[2], *Yannis Stylianou*[2,3], *Roberto Cipolla*[1,2]

[1]Department of Engineering, University of Cambridge, UK
[2]Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
[3]Department of Computer Science, University of Crete, Greece

ABSTRACT

In this paper, we present an expressive visual text to speech system (VTTS) based on a deep neural network (DNN). Given an input text sentence and a set of expression tags, the VTTS is able to produce not only the audio speech, but also the accompanying facial movements. The expressions can either be one of the expressions in the training corpus or a blend of expressions from the training corpus. Furthermore, we present a method of adapting a previously trained DNN to include a new expression using a small amount of training data. Experiments show that the proposed DNN-based VTTS is preferred by 57.9% over the baseline hidden Markov model based VTTS which uses cluster adaptive training.

Index Terms— Expressive Visual Text to Speech, Expression Adaptation, Deep Neural Network

1. INTRODUCTION

In this paper, we present a system that, given a text sentence and some expressive tag, produces a photorealistic talking head. Talking heads can be used to improve human-machine interactions systems. Such systems have been used in education, in reading news or eBooks and even in post processing of film production.

In this paper, we present a DNN-based text driven expressive talking head. The contributions of this work include an adaptation of audio DNN-based text to speech (TTS) to VTTS, a novel method of producing expressive visual speech using a DNN, yielding superior results compared to HMM-based VTTS, and a method of adapting a pretrained TTS system to incorporate a new expression using a small amount of adaptation data.

2. PRIOR WORK

Previously, various approaches have been presented for VTTS. In particular unit selection methods [11], interpolation methods [2], statistical methods based on hidden Markov models (HMM) [1] as well as hybrid systems that use HMMs to guide sample selection [16]. In this work, we pursue a DNN-based method for expressive VTTS.

Deep learning approaches have proven successful in many areas of signal processing (in computer vision [7], natural language processing [10] and automatic speech recognition [4] to name a few), including speech synthesis. Deep learning approaches to speech synthesis are statistical parametric speech synthesis (SPSS) systems. Recently, some deep learning approaches to TTS include [8], where a deep belief network is used to jointly model the acoustic and linguistic features and [19], where DNNs are used to map the linguistic features to acoustic features.

2.1. Talking Heads

Many talking heads, including the one presented in this work, use active appearance models (AAM) [3] to model the face region. Given a training set of images, each marked with a set of landmark points, an AAM builds an orthogonal basis of both shape and texture modes to describe most of the variation in the training set. By taking a weighted sum of these modes, different configurations of the images can be produced and these weights, or AAM parameters, are the parameters used to model the image set.

Talking heads can be considered in two categories: speech driven and text driven. In speech driven talking heads the audio speech in some way drives the visual model. In text driven talking heads both the audio and visual speech are synthesised from an input text sentence.

Several speech driven neural network based talking head models have been proposed recently. In [6], a context label sequence is generated from the audio speech and an AAM models the movement in the lower half of the facial region and a deep bidirectional Long Short Term Memory (LSTM) is used to map the label sequence to AAM parameters. In [9], a deep bidirectional LSTM with a bottleneck layer is trained to regress hand crafted low level descriptor features to contextual labels and the bottleneck features are regressed to facial animation parameters to drive the face model. In [13], a bidirectional LSTM is used to map MFCCs derived from the audio speech to Active Shape Model parameters and argue

this implicitly captures expression through the MFCCs and therefore can model expressive speech.

In text driven talking heads, the closest work to this one is by Anderson et al [1]. In [1], the audio features of a HMM-based TTS system are augmented with AAM parameters to form the full talking head. Different expressions are modelled using cluster adaptive training (CAT) [18]. The visual model used in [1] is the same as what is used here. In contrast to [6], the AAM covers the entire facial region. However, this work differs in that [1] uses a CAT HMM-based synthesis model to model different expressions, whereas this work models all expressions together in a single, multiple output DNN.

2.2. Adaptation

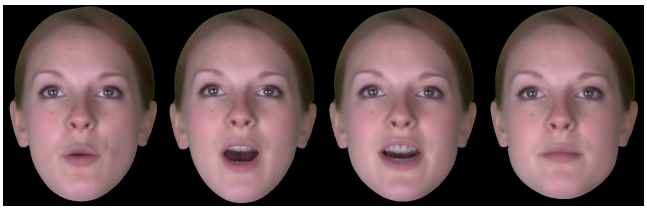
Speaker adaptation refers to the use of a small amount of training data from a novel speaker to modify a system that has already been trained on a large amount of training data. Speaker adaptation in DNN-based synthesis systems is performed in [17] and [5]. In [17], the system is modified in three ways. A speaker-specific code is appended to the input, a speaker-specific reweighting of the hidden unit contributions is learnt and a speaker-dependent mapping layer is learnt for each speaker. In [5], a single DNN models all speakers, with each speaker being modelled with a separate output.

3. MODEL

3.1. Visual Model

The visual model determines how the image of the face is constructed. In our system this is done parametrically, using an AAM. In this work, following [1], the modes are separated into semantically meaningful actions and regions: one mode is used to model blinking, two modes are used to model 3D head rotation, eight modes model the lower half of the face and six modes model the upper half of the face. Please refer to [1] for further details.

The interior of the mouth region is poorly modelled by the AAM and so is modelled separately with a static model of teeth and tongue. In addition, a static image of hair and ears is taken from the training set and added to the synthesised facial image. Figure 1 shows some samples of the visual model.



(a) (b) (c) (d)

Fig. 1: This figure shows the visual model performing several frames of the word "welcome".

3.2. Synthesis Model

3.2.1. Feature normalisation

The input text sentence is analysed and 503 numeric and binary features are extracted at every frame. These features include binary features regarding the current and neighbouring phonemes as well as numeric features regarding the quantities, relative positions and durations of phonemes, syllables and words. These features are then normalised to a range of $[0.01, 0.99]$ in order to avoid numerically large features from dominating numerically small features. These normalised features form the input to the DNN.

The output features are the audio and visual speech parameters. In this work we model the audio speech using 45 Mel-cepstral coefficients, logarithmic fundamental frequency, 25-band aperiodicities, 44 phase parameters, and their first and second time derivatives, in addition to a voiced/unvoiced decision. The visual speech is modelled using 17 AAM coefficients and their first derivatives. These are concatenated to form a 380 dimensional feature vector. The audio features were extracted according to the complex cepstrum analysis method of [12]. All features, except for the voiced/unvoiced flag are normalised to zero mean and unit variance. These normalised output features form the output from the DNN.

Because the video sample rate (25 frames per second) is lower than the audio frame rate (32 kHz), the video frame rate is increased to match the audio frame rate using cubic interpolation.

3.2.2. DNN

A DNN transforms the (normalised) linguistic features derived in 3.2.1 to (normalised) acoustic and visual features. The acoustic and visual features are modelled jointly because it is apparent that there is a large degree of correlation between audio and visual speech. By modelling them together in a single model, the mutual information between the audio and visual speech can be exploited.

In order to produce expressive speech, multiple emotions are modelled together in a single DNN with multiple outputs, one per emotion. Thus all layers of the DNN except the output layer are shared between the different expression classes and benefit from being trained by the entire expressive corpus, in contrast to any system which models each expression separately. This approach is inspired by [5]. This is shown in Figure 2.

At training time, for every frame in the training corpus, linguistic features are extracted from the aligned text and

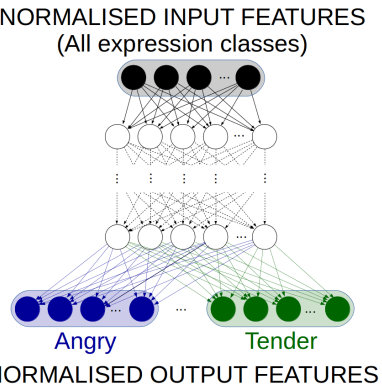


Fig. 2: Multiple output DNN, each output representing a different expression.

phonetic labels and acoustic and visual features are extracted from the audio and video respectively. The DNN is trained using backpropagation with care taken to propagate the error signal from the appropriate output only.

3.2.3. Postprocessing

The output of the DNN is normalised acoustic and visual features. Trajectories of the acoustic and visual parameters which best match the static and dynamic elements of the features are then found using the maximum likelihood parameter generation algorithm [15]. These trajectories are then sampled at the appropriate rate, the acoustic parameters are then used by a vocoder to generate audio speech, while the visual parameters drive an AAM to generate the appropriate accompanying facial movements.

The different expression outputs can be linearly interpolated to yield a continuous expression space. This can be done by simply taking a weighted sum of the different outputs. Because the output layer is linear, this is the same as forming a new output layer which has the same weighted sum of the different output layers.

3.3. Expression Adaptation

The penultimate layer of the DNN, which is the final shared layer, can be thought to represent a mean expression space, with the final expression dependent layers mapping to specific expressions. Given a small amount of training data for a new expression, a new output layer for the network can be added for that expression.

The input linguistic features are propagated to the penultimate layer of a previously trained DNN. No weight adaptation is performed at this stage. In this way, the small amount of new data can take advantage of the much larger training set that the DNN has already been trained on. Now, to estimate the final output layer, a linear regression is learnt between the penultimate layer and the output acoustic and visual features. This can be done using a linear least squares algorithm. Because there could potentially be only a small amount of new training data however, a regularised form of the linear least squares algorithm should be used to avoid overfitting on the small amount of adaptation data. In this work Tikhonov regularisation [14] was used.

Briefly, Tikhonov regularisation is a form of L_2 regularisation. In solving $A\mathbf{x} = \mathbf{b}$, ordinary least squares minimises $\|A\mathbf{x} - \mathbf{b}\|^2$, while least squares with Tikhonov regularisation minimises $\|A\mathbf{x} - \mathbf{b}\|^2 + \|\Gamma\mathbf{x}\|^2$, where $\Gamma = \alpha I$. The explicit solution is $\hat{\mathbf{x}} = \left(A^T A + \Gamma^T \Gamma\right)^{-1} A^T \mathbf{b}$. Larger values of α correspond to a higher level of regularisation.

Note that the rest of the DNN is not altered and that linear least squares algorithm is computationally much less expensive than retraining the DNN.

A particularly useful application would be to train an expressive talking head using a large corpus of neutral speech and then a small corpus of expressive speech. This can be achieved using an extension of the approach outlined above by training the DNN on the neutral speech alone and then adding the expression layers later using only a small amount of data for each expression.

4. EXPERIMENTS AND RESULTS

4.1. Corpus

The training corpus is populated with 6591 utterances (735 angry, 696 fearful, 697 happy, 3078 neutral, 691 sad and 694 tender utterances) of a female native British English speaker. There are a further 100 utterances (50 neutral and 10 for each expression) in the validation corpus.

4.2. Model

The DNN used has six hidden layers, each with 1024 neurons and six different output layers, one for each of the expression classes - angry, fearful, happy, neutral, sad and tender. The weights of the DNN were initialised randomly and optimised using stochastic gradient descent with gradients estimated using backpropagation. The error function used was the mean square error between the predicted and actual output features over a minibatch. Minibatches were formed of 256 samples. The network was regularised by penalising the L_2-norm of the weights with a penalty of 0.001.

4.3. Evaluation

4.3.1. Quantitative

The synthesised results have been evaluated quantitatively by examining the root-mean-square (RMS) difference in the landmark points of the AAM, as well as the shape-normalised

	All layers trained	One layer adapted				All expressive layers adapted
		$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$	$\alpha = 10000$	
Angry	315.8	330.0	325.4	**323.6**	331.9	329.8
Fear	330.9	346.7	**316.8**	338.4	343.4	345.2
Happy	320.1	332.6	**317.1**	326.3	334.0	332.1
Sad	310.3	319.9	316.4	**315.5**	319.6	319.5
Tender	304.4	315.4	316.4	**310.7**	316.9	314.4
Average	316.3	328.9	**318.4**	322.9	329.2	328.2

Table 1: Comparing the average DNN output error on a test set on the various expressive subsets for three different experiments. Firstly, an experiment where all output layers are trained. Secondly, an experiment where one expressive layer is estimated, for various values of α. The lowest error is given in bold. Thirdly, an experiment where the network is trained only on neutral data and all expressive layers are estimated.

texture (pixel values) between the synthesised and real data on samples not used in training. The RMS shape error across all landmark points of all test samples is 4.4 pixels. The RMS texture error across all pixel values is 2.9. Note that the texture error should be considered with an understanding that RGB values of a pixel are in the range $[0, 255]$. The Mel cepstral distance from the synthesised audio is 5.69 dB.

The quality of the expression adaptation as outlined in Section 3.3 is assessed quantitatively by comparing the error in output of the synthesis model from a fully trained network and from a network with an adapted output layer. 20 utterances were used as adaptation data.

In Table 1, the average DNN output errors for a fully trained network are compared to networks with output layers that were estimated using regularised least squares. The training samples from the expression class corresponding to the estimated output layer were omitted during training. Furthermore, the value of the α was varied; for $\alpha < 100$, the layer weights are too large and overfit the adaptation data and for $\alpha > 1000$, the layer weights are too small and do not model the adaptation data well. For further comparison, the results from a network only trained on neutral speech and adapted for five different expressions is also given. While the error of the adapted layers does increase slightly over the trained layers, they are still comparable.

4.3.2. Qualitative

The DNN-based system presented here is compared to another expressive talking head based on an HMM system [1] that models different expressions using CAT. The test subjects were presented with the 36 pairs of test sentences (not present in the training or validation sets), where the samples in each pair were produced by the DNN-based system and the HMM-based system, with six sentences from each of the six expression classes. Each pair was evaluated by 18 subjects. The subjects were asked, after watching the pair of synthesised samples, to indicate their preference or lack of preference. The results are given in Table 2. The DNN-based system is preferred to the HMM-based system overall and in every expression category. Some sample video clips can be viewed by visiting http://mi.eng.cam.ac.uk/~jwp37.

	Preferred HMM	Preferred DNN	No Preference
Overall	30.0 ± 3.6	$\mathbf{57.9 \pm 3.9}$	12.1 ± 2.6
Angry	38.9 ± 9.4	$\mathbf{55.6 \pm 9.6}$	5.6 ± 4.4
Fearful	38.9 ± 9.3	$\mathbf{41.7 \pm 9.5}$	19.4 ± 7.6
Happy	25.0 ± 8.3	$\mathbf{67.6 \pm 9.0}$	7.4 ± 5.0
Neutral	20.7 ± 7.9	$\mathbf{65.1 \pm 9.3}$	14.1 ± 6.8
Sad	32.4 ± 9.3	$\mathbf{46.1 \pm 9.9}$	21.6 ± 8.1
Tender	24.5 ± 8.5	$\mathbf{71.6 \pm 9.3}$	3.9 ± 3.8

Table 2: Preference scores (%) with 95% confidence intervals between HMM-based model and DNN-based model.

5. CONCLUSION

This paper presented a DNN-based expressive talking head. Unlike other systems, the entire face is modelled, which is important for expressive visual speech. Furthermore, new expressions can be added to the model with a small amount of adaptation data and without retraining the network. Our method outperforms an expressive HMM-based talking head.

Attention is being turned to investigate alternate methods of modelling different expressions, in particular using a gated final layer, where the final layer models a threeway relationship between the penultimate layer, the expression and the output. Furthermore, we wish to investigate the use of similar techniques to jointly model speaker and expression and so perform both speaker and expression adaptation.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Alexandros Papangelis for providing samples of the HMM-based talking head.

7. REFERENCES

[1] Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive visual text-to-speech using active appearance models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[2] Volker Blanz, Curzio Basso, Thomas Vetter, and Tomaso Poggio. Reanimating faces in images and video. In Pere Brunet and Dieter W. Fellner, editors, *EUROGRAPHICS 2003 (EUROGRAPHICS-03) : the European Association for Computer Graphics, 24th Annual Conference*, volume 22 of *Computer Graphics Forum*, pages 641–650, Granada, Spain, 2003. The Eurographics Association, Blackwell.

[3] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 484–498. Springer, 1998.

[4] George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *ICASSP*, pages 8609–8613, 2013.

[5] 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.

[6] Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K. Soong. A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools Appl.*, 75(9):5287–5309, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *arXiv prepring arXiv:1506.01497*, 2015.

[8] 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.

[9] Xinyu Lan, Xu Li, Yishuang Ning, Zhiyong Wu, Helen Meng, Jia Jia, and Lianhong Cai. Low level descriptors based DBLSTM bottleneck feature for speech driven talking avatar. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 5550–5554, 2016.

[10] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[11] Kang Liu and J. Ostermann. Realistic facial expression synthesis for an image-based talking head. In *IEEE Int. Conf. Multimedia and Expo*, pages 1–6, July 2011.

[12] Ranniery Maia, Masami Akamine, and Mark J.F. Gales. Complex cepstrum for statistical parametric speech synthesis. *Speech Communication*, 5(55):606–618, June 2013.

[13] Taiki Shimba, Ryuhei Sakurai, Hirotake Yamazoe, and Joo-Ho Lee. Talking heads synthesis from audio with deep neural networks. In *2015 IEEE/SICE International Symposium on System Integration, SII 2015, Nagoya, Japan, December 11-13, 2015*, pages 100–105, 2015.

[14] Andrey Nikolaevitch Tikhonov. *Numerical methods for the solution of ill-posed problems*. Mathematics and its applications. Kluwer Academic Publishers, Dordrecht, Boston, 1995.

[15] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1315–1318. IEEE, 2000.

[16] Lijuan Wang and Frank K. Soong. Hmm trajectory-guided sample selection for photo-realistic talking head. *Multimedia Tools and Applications*, 74(22):9849–9869, 2015.

[17] 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.

[18] Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark J. F. Gales, Kate Knill, Sacha Krstulovic, and Javier Latorre. Statistical parametric speech synthesis based on speaker and language factorization. *IEEE Trans. Audio Speech Lang. Process.*, 20(5), 2012.

[19] 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.