Visual speech synthesis using dynamic visemes, contextual features and DNNs

Ausdang Thangthai, Ben Milner, Sarah Taylor

School of Computing Sciences, University of East Anglia, Norwich, UK

a.thangthai@uea.ac.uk, b.milner@uea.ac.uk, s.l.taylor@uea.ac.uk

Abstract

This paper examines methods to improve visual speech synthesis from a text input using a deep neural network (DNN). Two representations of the input text are considered, namely into phoneme sequences or dynamic viseme sequences. From these sequences, contextual features are extracted that include information at varying linguistic levels, from frame level down to the utterance level. These are extracted from a broad sliding window that captures context and produces features that are input into the DNN to estimate visual features. Experiments first compare the accuracy of these visual features against an HMM baseline method which establishes that both the phoneme and dynamic viseme systems perform better with best performance obtained by a combined phoneme-dynamic viseme system. An investigation into the features then reveals the importance of the frame level information which is able to avoid discontinuities in the visual feature sequence and produces a smooth and realistic output.

Index Terms: talking head, visual speech synthesis, deep neural network, dynamic visemes

1. Introduction

Visual speech synthesis has application in generating a speech animation or talking head and is used in, for example, the entertainment industry for animated characters in games and films. In practice, artists produce speech animations manually or use motion capture technology with a human actor. The quality of the animation is usually of primary concern which is why such methods are employed, even given the time and cost required. Generating animations automatically is much less expensive and can be performed in real-time but is not generally considered to be good enough for industry applications, although hybrid approaches can be used where automated animations are refined by artists. The aim of this work is to improve the quality of visual speech synthesis produced automatically from a text input of words.

Methods of visual synthesis can be broadly divided into model-based, sample-based and statistical methods. Model-based approaches construct a sequence of frames for each phoneme and then interpolate between them to generate animations. For example, [1] extracts visual parameters from each phoneme using a hand-tuned dominance function and produces animations using a blend function, although the method is time-consuming. More flexible data-driven approaches extract visual speech parameters from a speech corpus but the resulting animation can often be rather unrealistic [2, 3]. Sample-based approaches concatenate visual speech units contained in a database, where the units might be fixed-length (e.g. phonemes, visemes, or words [4, 5, 6, 7]) or of variable length [8, 9, 10]. A cost function, based on phonetic context and smoothness of concatenation, is then minimised to find the set of units which form the animation. Having a sufficiently large database is important to allow natural smoothing between concatenated units and avoids discontinuities. Statistical approaches aim to overcome these problems by learning and then predicting visual speech parameters from phonetic context, such as with Gaussian mixture models (GMMs), hidden Markov models (HMMs) or deep neural networks (DNNs) [11, 12, 13, 14, 15]. HMMs have been state of the art in visual speech synthesis for the past decade and typically employ decision tree clustered context-dependent models, although a drawback has been an oversmoothed output [16]. DNN approaches have more recently been proposed to address these limitations and can predict visual parameters from contextual features [17]. Related work has also converted acoustic speech features (e.g. filter bank, MFCC, LPC) into head motion parameters (nod, yaw, roll) using a feed-forward neural network model [18].

This paper continues with the DNN-based approach for predicting visual features from a text input but aims to improve the resulting naturalness of the animation. First, from the text input, two kinds of speech units are considered. The first decomposes the input text into phonetic units. Although phonemes have been used widely in speech processing they have been shown to be suboptimal as visual speech units [6]. Instead, we propose using dynamic visemes as speech units and compare their performance to phonetic units before combining both. Secondly, we consider using more low-level (frame-based) contextual information in the feature vector applied to the DNN which is derived from the speech unit annotations, with the aim of producing a more realistic and smooth visual feature trajectory. Finally, in several earlier works the audio-visual databases have been relatively small which has limited the size of the DNNs. We now use a large database (14 hours) that allows larger DNNs to be trained and an optimal configuration is then identified.

The paper is organised as follows. Section 2 explains the full pipeline of the system with Section 3 describing the two visual speech units and Section 4 the contextual and visual features. Section 5 presents experimental results that examine the different configurations proposed and finally present the results of comparative subjective tests using human viewers.

2. DNN visual speech synthesis framework

The proposed method of transforming text to visual features suitable for visual speech synthesis is based on a feed-forward neural network with a number of hidden layers. A given text input is first converted to a sequence of contextual features which comprises a combination of binary features for categorical contexts (e.g. phonetic labels) and numerical features to represent values (e.g. number of phonemes in a syllable). Specific details of the input features are given in Section 4. The output features from the DNN are visual features (specifically active appearance model (AAM) features) along with their time derivatives

[19] which are then used for visual synthesis.

For training, the input and output features are time-aligned frame-by-frame. For each hidden and output unit in the DNN, a nonlinear activation function including a sigmoid, a hyperbolic tangent (tanh) and rectifier (relu) function is used to map all inputs from the previous layer to the next layer. Commonly, the activation function is controlled by connection weights and biases which are initialised by a uniform function or a pre-training algorithm. The goal of training is to find an optimal set of weight parameters using the backpropagation algorithm. For neural network regression purposes, a nonlinear activation function is used for hidden layers while a linear activation function is adopted in the output layer.

In DNN synthesis, the input text is first converted into input a sequence of features and the output sequence of visual features computed using forward propagation from the set of trained weights and biases. The output features which comprise static and dynamic visual features are rather disjoint so to improve their trajectories they are input into a speech parameter generation algorithm that generates a sequence of smooth static visual parameters frame-by-frame [20]. Finally, a rendering module re-synthesises a lip animation using the smoothed static AAMs parameters [21].

3. Phonetic and dynamic viseme units

Existing methods of predicting visual speech features from text input involves creating a phonetic annotation that is input into either an HMM or DNN synthesis system to create the sequence of visual features [13, 14, 15, 18]. In this work we use a set of 41 ARPAbet phonemes including short pause and silence [22]. The phonetic annotations contain the sequence of phonemes and their durations and can be created either manually or automatically, using for example forced alignment [23]. We compare this method to a more novel approach that is based on segmenting the text into a sequence of dynamic viseme units that subsequently form the input into the DNN [10].

Dynamic visemes are novel units designed for visual speech processing and represent groupings of similar lip-shapes (gestures) as opposed to groupings of similar speech sounds (i.e. phonemes). A set of dynamic visemes is learnt by clustering visual speech parameters which in this work are AAM features (see Section 4.4 for details). The visual speech is segmented by identifying points where AAM acceleration coefficients change sign which identifies instances where visible articulators change direction. This produces a set of variable length and non-overlapping visual gestures which are then clustered to produce a set of N dynamic visemes classes [10]. To determine N a series of HMM-based visual synthesisers were trained using different numbers of dynamic visemes. Examining the reproduced video established that using N=160 produced good quality animation with a relatively small number of classes.

4. Contextual and visual features

Feature extraction begins with either a time-aligned phoneme or dynamic viseme sequence that can be generated automatically from, for example, HMM decoding or from human annotation. To transform this sequence of speech units into a time sequence of visual features as required for visual synthesis, contextual labels at the phonetic/dynamic viseme and linguistic levels are extracted and used to create a suitable feature vector. For HMM synthesis of visual vectors this level of contextual labelling is sufficient but for input into a DNN, to create smooth trajectories, it is necessary to include frame-level features. In practice many contextual factors affect the way people speak which includes the number of syllables in current word, the phoneme/dynamic viseme context and the part-of-speech. We consider a number of such factors in our features and extract information at the frame, segment, syllable, word, phrase and utterance level. The importance of these is examined in Section 5.1 in terms of their effect on the synthesised visual features. The full set of features considered is summarised in Table 1 which shows those for phonetic units (PH) and for dynamic viseme units (DV).

Table 1: *Contextual features for phonetic (PH) and dynamic visemes (DV) units at varying levels.*

Level	Feature	PH	DV
Frame	Centre phoneme	x	
	Phonetic window context	x	
	Position and number of frames in phoneme	x	
	Forward phoneme span	x	
	Acoustic class	x	
	Centre dynamic viseme		x
	Dynamic viseme window context		x
	Position and number of frames in dynamic viseme		x
	Forward dynamic viseme span		x
Segment	Phoneme context	x	
	Dynamic viseme context		x
	Number of phonemes in dynamic viseme		x
Syllable	Position and number of phonemes in syllable	x	
	Position and number of dynamic visemes in syllable		x
Word	Position and number of syllables in word	x	x
Phrase	Position and number of syllables in phrase	x	x
	Position and number of words in phrase	x	x
Utterance	Position of syllable, word and phrase in utterance	x	x

4.1. Frame level features

To include contextual information, and improve the resulting predicted visual contour, a sliding window is used so that frame level information preceding and ahead of the current frame is included. The width of the window needs to be wide enough to include articulation movements but short enough to avoid over-smoothing of features. Related studies have reported a window width of $K = 11$ frames applied to 30 frame-per-second (fps) data which equates to a width of 330ms [24]. In this work the visual frame rate is 100fps which gives an equivalent window width of $K = 33$ frames (16 preceding and 16 ahead) which preliminary tests have established is a satisfactory value.

Considering the frame level features in Table 1, the **centre phoneme** feature is a 41-D binary feature that indicates the phonetic class of the current frame (i.e. centre of the window). Frame level phonetic context is included for the 16 frames pre-

ceding and ahead of the current phoneme which form the 32 × 41 dimensional **Phonetic window context** binary feature. The **position** feature has three binary elements that correspond to whether the centre frame is at the *start*, *middle* or *end* of the current phoneme, while **number** indicates how many frames are in the phoneme [25]. The **forward phoneme span** indicates how many frames the current phoneme are present before changing to another phoneme [24]. **Acoustic class** is represented by a 57-D binary feature where each element is a response to questions such as '*Is the current phoneme voiced?*' or '*Is the current phoneme nasalised?*', which are taken from the contextual questions in HTS [26]. The final column of Table 1 shows a similar set of features defined for dynamic viseme units. These form longer binary features given that 160 dynamic visemes are used as opposed to 41 phonemes and no equivalent acoustic class feature exists as the units are visually-derived.

4.2. Segment level features

We define a segment as being five phonemes or five dynamic visemes in duration, centred about the middle unit, as preliminary tests found this to give best performance. The five phonemes in the segment are represented by the 41 × 5 dimensional **Phoneme context** binary feature that indicates the current, two preceding and two following phonemes. Similarly, **Dynamic viseme context** is a 160 × 5 dimensional feature that indicates the five DVs in the segment. **Phonemes in DV** is a numeric feature representing the number of phonemes in the dynamic viseme.

4.3. Syllable, word, phrase and utterance features

The syllable level features of **number** and **position** indicate how many phonemes or DVs are in the current syllable and the current position (*start*, *middle* or *end*) within the syllable. At the word, phrase and utterance levels the number and position features indicate similar information but are no longer unique to phonemes or DVs.

4.4. Visual features

An active appearance model (AAM) is used to track and parameterise the facial region in each frame of the video [27]. From a set of 34 2-D vertices that define a mesh demarcating the contours of the lips, jaw and nostrils a 30-D AAM vector, $\boldsymbol{y}$, is extracted. Preliminary tests examined visual frame rates of between 30fps and 200fps and established highest accuracy was with 100fps which is used for all subsequent testing.

5. Experiments results

Experiments are performed on the KB-2k audiovisual speech dataset which contains 2543 phonetically balanced sentences from TIMIT totalling around 14 hours [10]. Recordings were captured in a reading style with no emotion in both frontal and side views of professional male speaker, although only the frontal view is used in this work. Video was recorded at 30fps and subsequently upsampled to 100fps. Objective experiments are presented first and analyse the effectiveness of different frame level features, the use of phonetic or dynamic viseme speech units and finally optimisation of the DNN. Subjective test results are then presented that compare the proposed method with an HMM-based method of visual synthesis that serves as a baseline. Objective tests use correlation and root mean square error (RMSE) to evaluate the effectiveness of predicting visual features. Correlation and RMSE are measured between each predicted AAM coefficient and its reference value and averaged across the first five AAM coefficients which gives a more stable measure of similarity.

The DNN for prediction is a feed-forward network with three hidden layers each consisting of 3000 units. A hyperbolic tangent activation function was used for hidden layers and a linear activation function was employed at the output layer. Mini-batch stochastic gradient descent was used and the size of mini-batch set to 100; a learning rate and momentum were fixed to 0.1 and 0.9, respectively. The maximum number of epochs was set to 100. 50% dropout was also applied to avoid overfitting on hidden units. 10% of the training set was held out for validation purposes while 50 held out sentences are used for testing. Numerical input and output features were normalised to give zero-mean unit-variance.

5.1. Analysis of frame level features

These experiments analyse the effect that frame level features have on prediction accuracy. Only phonetic units are considered and features for segment, syllable, word, phrase and utterance (as defined in Table 1) are included. Five frame level feature combinations (A to E) are defined and summarised in Table 2. System A uses only the frame level centre phonetic indicator and acoustic class. System B extends this with the frame position and number features while System C extends System A with the phonetic window context and forward span features. System D combines the features in Systems B and C. System E then excludes the acoustic class feature.

Table 2: *Frame level feature combinations.*

	A	B	C	D	E
Centre phoneme	×	×	×	×	×
Phonetic window context			×	×	×
Position and number of frames		×		×	×
Forward phoneme span			×	×	×
Acoustic class	×	×	×	×	

Table 3 presents correlation and RMSE results of the five systems which shows that including all frame level features gives best performance (System D). Comparing Systems B and C shows that the phonetic window context feature makes a significant contribution to performance by including wider frame level information that is not present with just knowledge of the central phoneme. This is illustrated in Figure 1 which shows AAM coefficient 5 for the utterance '*If dark came they would lose her*' which is changed from being discontinuous at frame boundaries without phonetic window context (System B) to continuous (System C) and much closer to the original track.

Table 3: *Correlation and RMSE of phonetic frame level feature combinations (brackets show ±standard deviation).*

	Correlation	RMSE
System A	0.75(0.07)	10.46(2.17)
System B	0.75(0.07)	10.43(2.16)
System C	0.80(0.07)	9.43(2.24)
System D	**0.81(0.07)**	**9.39(2.23)**
System E	0.79(0.07)	9.74(2.26)

Table 4: *Mean correlation and RMSE (±standard deviation) with different numbers of hidden layers and units for DNN-based approach using a combination of dynamic viseme and phonetic units.*

Units/ Layers	512		1024		2048		3000	
	Corr	Rmse	Corr	Rmse	Corr	Rmse	Corr	Rmse
1	**0.85(±0.05)**	**7.74(±1.32)**	**0.86(±0.05)**	**7.62(±1.25)**	0.86(±0.05)	7.66(±1.33)	0.86(±0.05)	7.54(±1.25)
2	0.85(±0.05)	7.95(±1.29)	0.86(±0.05)	7.71(±1.32)	0.86(±0.05)	7.60(±1.32)	0.86(±0.05)	7.49(±1.28)
3	0.84(±0.05)	8.05(±1.31)	0.86(±0.05)	7.68(±1.29)	0.86(±0.05)	7.55(±1.29)	0.86(±0.05)	7.66(±1.28)
4	0.85(±0.05)	8.03(±1.28)	0.86(±0.04)	7.66(±1.23)	0.86(±0.04)	7.41(±1.19)	0.87(±0.04)	7.32(±1.24)
5	0.85(±0.05)	8.08(±1.30)	0.86(±0.04)	7.75(±1.21)	**0.87(±0.04)**	**7.37(±1.15)**	**0.88(±0.04)**	**7.18(±1.26)**

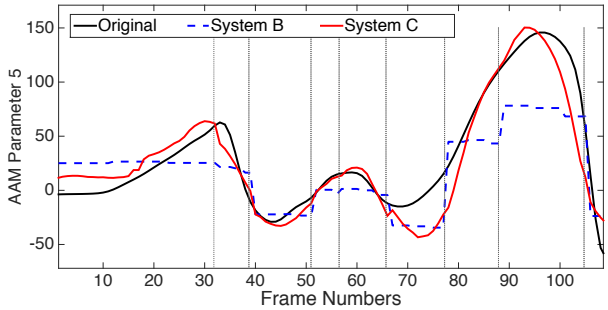

Figure 1: *Comparison of reference and predicted AAM coefficient 5 - with (System C) and without (System B) phonetic window context feature. Vertical lines show phoneme boundaries.*

5.2. Analysis of speech units

An investigation is now made into the effect of using either phonetic units or dynamic viseme units. For the phoneme based system all PH features shown in Table 1 are included while for the dynamic viseme system all DV features are included. A third configuration was also tested which combines the phonetic and dynamic viseme unit features and includes all features shown in Table 1. For comparison, a baseline HMM synthesis system was created which used five-state hidden semi-Markov models with each state modeled by a single Gaussian with diagonal covariance. Quinphone HMMs were created using decision tree clustering that considered phoneme, syllable, word, phrase and utterance level questions and resulted in 11,893 models [13].

Table 5 shows correlation and RMSE for the phoneme, DV and combined phomeme-DV systems using DNNs and the phoneme-based HMM system. Both the phoneme DNN and DV DNN systems outperform the HMM synthesis approach. Combining phoneme and DV features further improves performance which we attribute to their complementary information, one relating to acoustics and the other to visual information, which when combined improves the resulting visual features.

Table 5: *Correlation and RMSE performance of HMM and DNN approaches using phonemes and dynamic viseme units.*

	Correlation	RMSE
Phoneme HMM (Baseline)	0.75(±0.08)	10.82(±2.13)
Phoneme DNN	0.81(±0.07)	9.39(±2.23)
Dynamic-viseme DNN	0.80(±0.06)	8.79(±1.25)
Phoneme + DV DNN	0.86(±0.05)	7.66(±1.28)

5.3. Optimisation of DNN parameters

Optimisation of the DNN is now considered for the combined phoneme-DV system with the aim of further improving performance. The number of hidden layers was varied from 1 to 5 and the number of units from 512 to 3000. Table 4 shows correlation and RMSE for these combinations and identifies best performance with five hidden layers and 3000 units per layer. Increasing parameters further saw performance reduce.

5.4. Subjective evaluation

Subjective tests were performed to compare the naturalness of animations generated by AAM sequences from either the DNN/phoneme-DV system or the HMM system defined in Section 5.2. Viewers were asked to watch pairs of animations played side-by-side and to select the sequence that they found most natural. One animation was created from the DNN-Phoneme/DV system and the other from HMM synthesis, with the order of them randomised. Each test comprised 20 sentences that were selected randomly from the 50 test sentences. Viewers were asked to select the more natural animation and could watch the videos as many times as they wished. The original audio accompanied the video.

A total of 15 viewers took part in the tests and using a majority-wins voting system 80% found the DNN-based animations more natural. This confirms the objective test results in Table 5 that reported the DNN-based combination of phoneme and dynamic viseme units as being better than HMM synthesis. Several viewers reported that some of the animations (presumably the HMM synthesis) tended to lack audio-visual synchrony and to be under articulated in terms of mouth opening.

6. Conclusions

This paper has shown several improvements to visual speech synthesis from a text input using a DNN approach. Using phonetic or dynamic viseme speech units gives similar performance. However, combining the two speech units gives a substantial increase in performance which suggests that some complementary information exists between the two, and confirms other reported findings [10, 14]. An analysis of frame level features established that frame level context is very important for generating accurate AAM tracks, particularly for producing a smooth output that avoids discontinuities. Subjective tests confirm the findings made with objective tests with the proposed DNN-phonetic/DV systems producing much more natural animation than HMM synthesis.

7. References

[1] M. Cohen and D. Massaro, "Modeling coarticualtion in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. Thalmann and T. D, Eds. Springer-Verlag, 1994, pp. 141–155.

[2] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 388–398, Jul. 2002. [Online]. Available: http://doi.acm.org/10.1145/566654.566594

[3] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesisdetermination, adjustment, evaluation," *Speech Communication*, vol. 44, no. 14, pp. 141 – 154, 2004, special Issue on Audio Visual speech processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639304001128

[4] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of ACM SIGGRAPH*, 1997, pp. 353–360.

[5] F. Huang, E. Cosatto, and H. Graf, "Triphone based unit selection for concatenative visual speech synthesis," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2002, pp. 2037–2040.

[6] W. Mattheyses, L. Latacz, and W. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, no. 7–8, pp. 857–876, 2013.

[7] B. Theobald and I. Matthews, "Relating objective and subjective performance measures for AAM-based visual speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 8, p. 2378, 2012.

[8] E. Cosatto, G. Potamianos, and H. Graf, "Audio-visual unit selection for the synthesis of photo-realistictalking-heads," in *Proceedings of the International Conference on Multimedia and Expo*, 2000.

[9] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate visible speech synthesis based on concatenating variable length motion capture data," *IEEE Transactions on Visualization and Compuer Graphics*, vol. 12, no. 2, pp. 266–276, 2006.

[10] S. Taylor, B. Theobald, M. Mahler, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the Symposium on Computer Animation*, 2012, pp. 275–284.

[11] C. Luo, J. Yu, X. Li, and Z. Wang, "Realtime speech-driven facial animation using gaussian mixture models," in *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, July 2014, pp. 1–6.

[12] S. P. Deena, "Visual speech synthesis by learning joint probabilistic models of audio and video," Ph.D. dissertation, School of Computing Sciences, The University of Manchester, 2012.

[13] D. Schabus, M. Pucher, and G. Hofer, "Joint audiovisual hidden semi-markov model-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347, April 2014.

[14] A. Thangthai and B. Theobald, "Hmm-based visual speech synthesis using dynamic visemes," in *1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*, 2015.

[15] B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong, "A deep bidirectional lstm approach for video-realistic talking head," *Multimedia Tools and Applications*, pp. 1–23, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11042-015-2944-3

[16] S. Esmeir, S. Markovitch, and C. Sammut, "Anytime learning of decision trees," *Journal of Machine Learning Research*, vol. 8, p. 2007.

[17] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1561/2200000006

[18] C. Ding, L. Xie, and P. Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9871–9888, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11042-014-2156-2

[19] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, Feb 1986.

[20] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. 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*, vol. 3, 2000, pp. 1315–1318 vol.3.

[21] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135 – 164, November 2004.

[22] J. E. Shoup, "Phonological aspects of speech recognition:," 1980.

[23] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models," *Speech Commun.*, vol. 12, no. 4, pp. 357–370, Aug. 1993. [Online]. Available: http://dx.doi.org/10.1016/0167-6393(93)90083-W

[24] T. Kim, Y. Yue, S. Taylor, and I. Matthews, "A decision tree framework for spatiotemporal sequence prediction," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: ACM, 2015, pp. 577–586. [Online]. Available: http://doi.acm.org/10.1145/2783258.2783356

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

[26] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system version 2.0," in *Proc. of ISCA SSW6*, 2007, pp. 294–299.

[27] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, jun 2001. [Online]. Available: http://dx.doi.org/10.1109/34.927467