Comparing text-driven and speech-driven visual speech synthesisers

Barry-John Theobald[1], *Gavin Cawley*[1], *Andrew Bangham*[1], *Iain Matthews*[2] *and Nicholas Wilkinson*[1]

[1]School of Computing Sciences University of East Anglia, Norwich, UK.
[2]Weta Digital Limited, Wellington, New Zealand

{bjt,gcc,ab}@cmp.uea.ac.uk, iainm@wetafx.co.nz, nw@cmp.uea.ac.uk

Abstract

We present a comparison of a text-driven and a speech driven visual speech synthesiser. Both are trained using the same data and both use the same Active Appearance Model (AAM) to encode and re-synthesise visual speech. Objective quality, measured using correlation, suggests the performance of both approaches is close, but subjective opinion ranks the text-driven approach significantly higher.
Index Terms: visual speech synthesis

1. Introduction

Visual speech synthesisers can be broadly categorised as speech-driven or text-driven — see [1, 2] for an overview. We compare both approaches using the same underlying model for synthesis. In particular, the text-driven system from [3] is compared with a speech-driven approach that maps Mel-frequency cepstral coefficients (MFCCs) to AAM parameters using an Artificial Neural Network (ANN). AAMs are adopted in our synthesisers as they encode the changes in both the *shape* and *appearance* of the face in a few tens of parameters, and can later re-synthesise near-photorealistic images of the face from those parameters — see [4] for a description of AAMs.

1.1. Text-Driven Synthesis

To synthesise visual speech from text, the similarity between phoneme pairs in terms of AAM parameters is computed using:

$$S_{ij} = e^{-\gamma\left(\sum_{m=1}^{k+l}\sum_{n=1}^{5}\left[\left(v_i P^i_{mn} - v_j P^j_{mn}\right)w_m\right]^2\right)}. \qquad (1)$$

P^i and P^j are representations of phonemes i and j computed from examples in the corpus, the first summation is over the dimensions of the AAM and the second over samples equally spaced over the phoneme sub-trajectories. The parameters v_i are inversely proportional to the variance of the i^{th} phoneme, and w_m reflects the significance of the m^{th} AAM parameter. The similarities obtained with this measure match intuitive expectation. For example, {/b/, /p/, /m/}, {/f/, /v/}, {/tʃ/, /dʒ/, /ʃ/, /ʒ/}, etc., are most similar to one another.

Synthesised sequences are generated by measuring the distance between a desired context and the contexts in which a phoneme appears in the training corpus using:

$$\delta_j = \sum_{i=1}^{C} \frac{\mathbf{S}_{l_i j}}{i+1} + \sum_{i=1}^{C} \frac{\mathbf{S}_{r_i j}}{i+1}, \qquad (2)$$

where C is the context width and $\mathbf{S}_{l_{ij}}$ and $\mathbf{S}_{r_{ij}}$ are the similarity between the left and right contexts respectively. The selected sub-trajectories for the *best* examples are temporally normalised to the desired duration, concatenated, smoothed and applied to the model.

1.2. Speech-Driven Synthesis

The acoustic speech in the training corpus is encoded as MFCCs at 10ms intervals and the AAM parameters are up-sampled from 25Hz to 100Hz to match the audio. At each time-step, five frames either side of each AAM feature vector are concatenated to provide temporal context. A three-layer ANN with a 50-node hidden layer is used to learn the mapping from MFCCs to AAM parameters and a network is trained for each sentence in the corpus. This leave-one-out methodology matches that used in the text-driven synthesis.

2. Results

One hundred sentences not included in training were synthesised using both systems and the correlation between ground-truth and synthesised parameters for the first three parameters of the AAM are shown in Table 1. Viewers ratings (on a five-point Likert scale) for sequences presented in a random order show the text-driven output is significantly preferred ($p < 0.02$).

Table 1: *Mean correlation ($\pm\sigma$) between original and synthesised parameters for a test set of 100 held-out sentences.*

Parameter	Text	Speech
Shape 1	0.81 ±0.04	0.79±0.08
Shape 2	0.80 ±0.08	0.77±0.08
Shape 3	0.64 ±0.15	0.68±0.15
Appearance 1	0.62 ±0.16	0.75±0.11
Appearance 2	0.83 ±0.08	0.79±0.09
Appearance 3	0.76 ±0.10	0.77±0.10

3. Acknowledgements

The authors gratefully acknowledge the support of EPSRC (EP/D049075/1) for funding.

4. References

[1] Bailly, G., Bérar, M., Elisei, F., and Odisio, M., "Audio-visual speech synthesis", International Journal of Speech Technology, 6:331–346, 2003.

[2] Theobald, B., "Audiovisual Speech Synthesis", International Congress on Phonetic Sciences, 285–290, 2007.

[3] B. Theobald, A. Bangham, I. Matthews, and G. Cawley. "Near-videorealistic synthetic talking faces: Implementation and evaluation," *Speech Communication*, **44**, pp. 127–140, 2004.

[4] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, pp. 681–685, June 2001.