Transforming Neutral Visual Speech into Expressive Visual Speech

Felix Shaw and Barry-John Theobald

School of Computing Science, University of East Anglia, Norwich, Norfolk, UK

{f.shaw, b.theobald}@uea.ac.uk

Abstract

We present a method for transforming neutral visual speech sequences into realistic expressive visual speech sequences. By applying Independent Component Analysis (ICA) to visual features extracted from time aligned neutral and equivalent expressive sequences, a model that separates speech from expression can be learned. Analyzing the *behavior* of different speaking styles in terms of this model provides both a means for identifying the component(s) responsible for expression, and for learning the correspondence between different speaking styles. Exploiting this correspondence to transform neutral visual speech into expressive visual speech creates sequences that have the same time varying expressive dynamics as the equivalent ground-truth sequences, and an objective analysis shows that the neutral ICA parameters are shifted into the appropriate ranges for expressive visual speech.

Index Terms: expressive visual speech synthesis, independent component analysis, expressive style transformation

1. Introduction

Facial animation is employed extensively in computer games and in animated movies, but generating acceptable expressive speech animation remains a challenge. To ensure the highest quality animated sequences, animation studios usually resort either to using artists to handcraft animated sequences or use motion capture. Both of these approaches are expensive and time-consuming, whilst the latter makes it difficult to edit captured content subsequently. Automation of speech animation has been the focus of much research (see [1, 2] for an overview) and some approaches have reported highly realistic results [3, 4]. Despite this there has been limited, if any, adoption of these techniques by the animation industry. This is partly because the methods focus only on speech generation, and do not consider other communicative factors that are important, such as facial expressions and how these expressions interact with the accompanying facial movements due to speech.

In this paper, we describe a method for retargeting recorded neutral speech into the equivalent expressive speech using only a small training set of expressive sentences. Key to this is the idea of decomposing an expressive speech sequence into independent expressive and speech components [5, 6, 7]. Factorizing these components makes it possible to learn the relationship between neutral visual speech and expressive visual speech, and thus given a new sequence of neutral visual speech, expression can be added and manipulated (largely) independently of the speech content.

2. Related Work

Early approaches for expressive facial animation modeled the face using a geometric mesh, which was animated using either pseudo-muscle models embedded within the mesh [8] or a more complete physically-based approach [9]. Sifakis et al. [10] implemented a more complex and realistic muscle simulation based on tetrahedral meshes constructed using MRI and laser scanned data. Muscle activations corresponding to facial expression were then solved for using an inverse muscle activation algorithm. Muscle activations for expressions can be blended with activations for mouth shapes to produce expressive speech animation. However, these expressive activations are static, so the resultant animations lack dynamic subtlty. Furthermore, these more complicated models are increasingly computationally expensive and become more difficult to interact with.

More recently, statistical approaches have been applied to expressive video sequences [7, 11] or to motion capture data [12, 5, 13] in an attempt to parameterize speech and facial expression. Such approaches include bilinear [7, 11] and trilinear models [12], both of which factorize expressive speech into separate components so that parameterized neutral speech sequences can be modulated with expression parameters. Alternatively, independent component analysis (ICA) [14] provides a different approach based on the idea of an expressive speech signal being an additive mix of independent expression and speech signals [5, 13]. A mapping is learned between different emotional styles in ICA space by training a model on sequences containing two different styles (e.g. happy and sad). One ICA component (or mode) is then responsible for the facial movements associated with expression, whilst the rest are responsible for speech. By manipulating the expressive mode for

novel sequences, the corresponding facial expression can be switched between the two expressions on which the model was trained.

Inspired by [5, 13] we propose an approach for using ICA to separate speech and expression in expressive speech sequences. However, rather than requiring a model for each pair of facial expressions, we instead propose an approach that requires only a single model per expression.

3. Data Pre-processing

All work described in this paper was conducted using the B3D(AC)^2 corpus [15]. The data consists of 3D facial laser scans at 25Hz, of six males and eight females each speaking 80 sentences. Each sentence is said once in a neutral style and once as an expressive equivalent. A phonetic annotation of the sentences is provided with the corpus. The mesh for a single frame is represented as $\mathbf{x} = \{x_0, y_0, z_0, x_1, y_1, z_1, \ldots, x_{n-1}, y_{n-1}, z_{n-1}\}^T$, where $n = 23,370$. The scanned data are provided in vertex correspondence with a triangulation for surface rendering. A more compact model that allows for linear variation in the deformation of the mesh is given by:

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b}, \tag{1}$$

where the coefficients $\mathbf{b}$ are shape parameters used to encode the high-dimensional mesh. Such a model can be computed by applying principal component analysis (PCA) to a training set of meshes, so $\overline{\mathbf{x}}$ is the mean shape and the columns of $\mathbf{P}$ are the eigenvectors corresponding to the largest eigenvalues. The highly structured variation in the vertex locations means that each frame can be represented as a five-element vector, where five principal components account for 85% of the total variation in the original 70,110 dimensional space.

3.1. Independent Component Analysis

The time-varying principal component values, $\mathbf{b}$ in Equation 1, encode changes in position of the facial features due to both speech and expression. An underlying assumption is that this is a mixed signal composed of two independent signals, a speech and an expression signal, which are combined in an additive manner. This can be modelled as:

$$\mathbf{b} = \mathbf{A}\mathbf{s}, \tag{2}$$

where $\mathbf{s}$ represents the unknown source (speech and expression) signals and $\mathbf{A}$ is a matrix of unknown mixing coefficients. ICA provides a framework for estimating the mixing matrix such that the independent components can be computed using:

$$\mathbf{s} = \mathbf{W}\mathbf{b}. \tag{3}$$

where $\mathbf{W}$ is the pseudo-inverse of $\mathbf{A}$, and is calculated to maximize differential entropy and minimize mutual information of the random vector $\mathbf{b}$ [16].

In this work we use the publicly available FastICA algorithm [17] to calculate $\mathbf{A}$ and $\mathbf{W}$. To identify the independent components each utterance was projected onto the principal components using:

$$\mathbf{b} = \mathbf{P}^T \left(\overline{\mathbf{x}} - \mathbf{x}\right). \tag{4}$$

Next, the neutral utterances and the equivalent expressive sequences of a particular style (happy, sad, etc.) were time aligned using dynamic time warping (DTW). The number of ICA components retained by FastICA was made equal to the same number of PCA components to avoid any data loss.

4. Transforming Neutral Speech into Expressive Speech

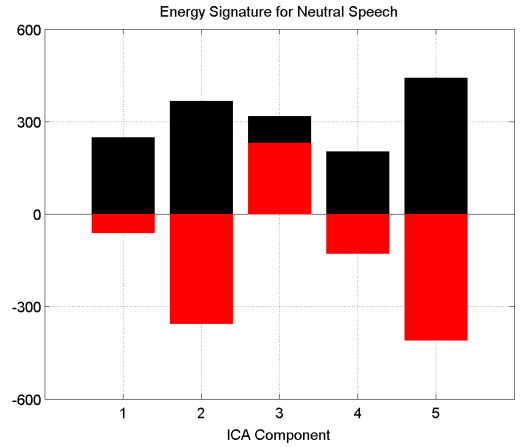
When transforming parameters that encode neutral visual speech into those which encode expressive visual speech, the dynamics of the expression must appear natural and the mouth movements corresponding to speech must remain valid. If all of the assumptions of ICA held, some of the independent components would correspond exactly to speech and some exactly to expression. However, we have found that a 'clean' separation of the signals does not occur, and each component tends to represent both speech and expression movements to varying degrees. This was verified by setting all values in every ICA mode except one to zero then reconstructing the mesh, thus showing the influence of only one ICA mode. An additional problem is that there is no ordering of the components returned by ICA, so there is no obvious objective way to discriminate between those components which predominantly represent speech and those which predominantly represent expression.

Although no component is fully responsible for expression, the distribution of the energy in the components is different for neutral speech and expressive speech. Figure 1 shows what we refer to as the *energy signatures* for neutral and expressive speech. The black bars in the Figure represent the energy in the components, computed using:

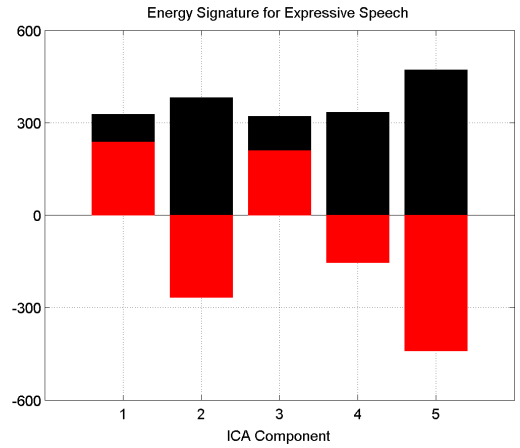
$$e^j = \sum_{t=1}^{k} |s^j(t)|, \tag{5}$$

where e^j represents the energy in the j^{th} component and $s^j(t)$ represents the value of the j^{th} component at time t. The red bars in Figure 1 show the amplitude of the components, computed using:

$$a^j = \sum_{t=1}^{k} s^j(t). \tag{6}$$



(A)



(B)

Figure 1: The energy (black) and amplitude (red) in the ICA components of (A) four neutral speech sequences and (B) the equivalent four expressive speech sequences.

The main differences between the neutral and expressive energy signatures is that one component tends to be negative for neutral speech, whereas it tends to be positive for expressive speech (in this case component 1). Another difference is that a component tends to have more overall energy in expressive speech than in neutral speech (in this case component 4). The task then is to compute the ICA components for a novel neutral speech sequence and then redistribute the energy in these components so that they better match those observed in expressive speech. This involves rescaling the values of the components and (possibly) changing the sign. The weights used to transform the neutral speech components are computed using:

$$w^j = \frac{e_e^j}{e_n^j} \tag{7}$$

where w^j is the scaling for the j^{th} component, e_e^j is the energy in the j^{th} component of expressive speech and e_n^j the energy in the j^{th} component of neutral speech. Thus, given a sequence of novel neutral speech projected into ICA space, the parameter values are adjusted according to:

$$u^j(t) = \begin{cases} w^j s^j(t) & \text{if } \mathrm{sgn}(a_e^j) = \mathrm{sgn}(a_n^j) \\ (-w^j(s^j(t) - \mu^j)) + \mu^j & \text{otherwise} \end{cases} \tag{8}$$

where μ^j is the mean value of the j^{th} component over the novel utterance, and sgn is $+1$ if the amplitude is positive and -1 if the amplitude is negative. The value $u^j(t)$ represents the new value of the j^{th} independent component at time t.

5. Results

Five sequences from the B3D(AC)^2 corpus were chosen for each expressive style and paired with their equivalent neutral speech. To maximize the limited data available leave-one-out training was used, where five ICA models for an expressive style were trained using four of the sequence pairs, with a fifth sequence held-out for testing. Figure 2 shows example time-varying trajectories in three of the five independent components for ground truth expressive, ground-truth neutral and the corresponding transformed neutral visual speech. Note that the transform is not attempting to recreate the expressive sequence exactly, rather the *style* of the expressive speech is being imposed onto the *content* of the neutral speech.

Sequences transformed from neutral to expressive styles using the process described in Section 4, not only show the correct change in facial expression, but also display the dynamics which are seen in the training set because real ICA data is being scaled rather than a style being statically imposed. Sample frames from video sequences containing real neutral speech, the same speech after transforming to an expressive style, and the corresponding real expressive sequences time-aligned to the neutral sequence are shown in Figure 3.

6. Evaluation

A small subjective evaluation involved a forced choice Turing test, where 14 viewers were each shown 8 sequence pairs (n=112 samples). Sequences of time aligned real and transformed expressive speech were shown as visual only to ensure that acoustic artefacts due to time aligning the sequences had no influence on the results. The left-right ordering of the pair was randomized and viewers were asked to identify the real sequence in the pair. Of the 112 samples, 43 of the responses were correct. Using a binomial significance test we find that viewers cannot reliably identify the real sequences from the transformed sequences ($p > 0.3$). In several cases, viewers stated that they found it difficult to choose between sequences in terms of realism, and so therefore chose their favorite. Responses tended to be biased in favor of transformed sequences being identified as real — so we ob-

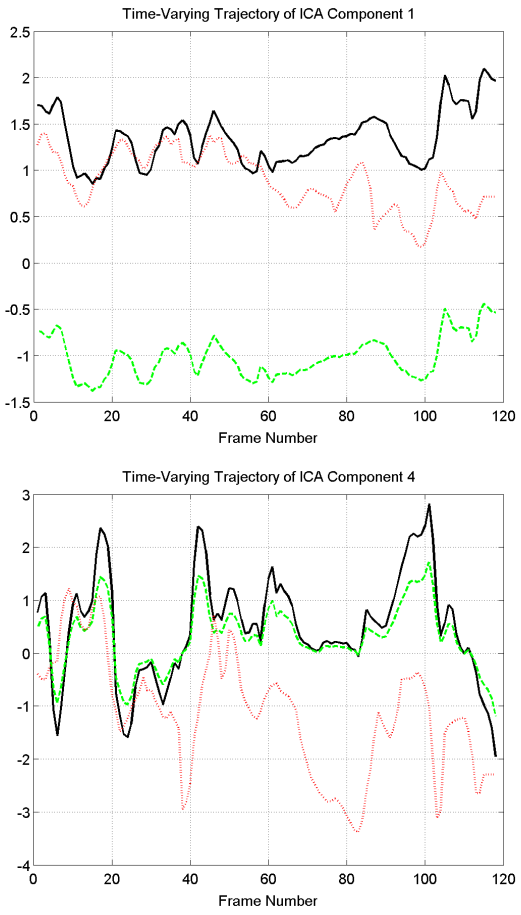


Figure 2: Time-varying independent components for a ground-truth neutral sequence (green dashed curve), the time aligned expressive equivalent sequence spoken in a happy style (red dotted curve), and the neutral sequence transformed into a happy style (black solid curve). In particular, note how component one has been shifted into the correct range for expressive speech, but still displays a similar shape to the ground truth neutral speech. Also, the bars in Figure 1 suggest that component four should have greater amplitude after transformation.

serve more false positives than false negatives. This is perhaps explained by the fact that transformed sequences tend to be slightly attenuated and thus smoother than the corresponding ground truth data.

7. Summary

We have described a method for transforming sequences of neutral visual speech into expressive visual speech. Independent component analysis is used to decompose time aligned neutral and expressive visual speech, and weights are learned to distribute the energy in the independent components of (novel) neutral speech to better match the energy observed in expressive visual speech. This transformation results in expressive utterances that appear to display the same kinds of expression as seen in the expressive training set, and importantly the integrity of mouth shapes remains intact. Our approach uses ICA to separate neutral from expressive speech, unlike previous attempts which are trained to separate expressive data of different types (e.g. happy from sad). The advantage of this is that the number of models grows linearly as we train for new expressions, whereas separating different expression types requires a model for each pair of expressions. This technique is flexible in that it allows any arbitrary neutral visual speech to be transformed into an expressive style using only a small training set of expressive and neutral speech.

Future work will focus on extending this work to represent multiple expressions in a single model to firstly reduce the number of models required for representing a broad set of expressions, and secondly to investigate how new expressions can be generated as combinations of existing expressions. We will also work on incorporating this technique into an expressive visual speech synthesizer. To date we have only applied the method to transform *real* neutral visual speech into the equivalent expressive visual speech. A neutral visual speech synthesiser can be trained from a large corpus of neutral visual speech, which would be able to produce realistic neutral synthesised visual speech. The technique presented in this paper could then be used to transform this synthesized neutral visual speech into expressive visual speech using only a few expressive training sequences. The advantage of this is that the neutral speech is relatively easy to capture (compared with expressive speech), and it avoids the need to capture a full corpus of speech for each expressive style.

8. References

[1] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," in *International Journal of Speech Technology*, vol. 6, 2003, pp. 331–346.

[2] B. Theobald, "Audiovisual speech synthesis," in *International Congress on Phonetic Sciences*, 2007, pp. 285–290.

[3] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic

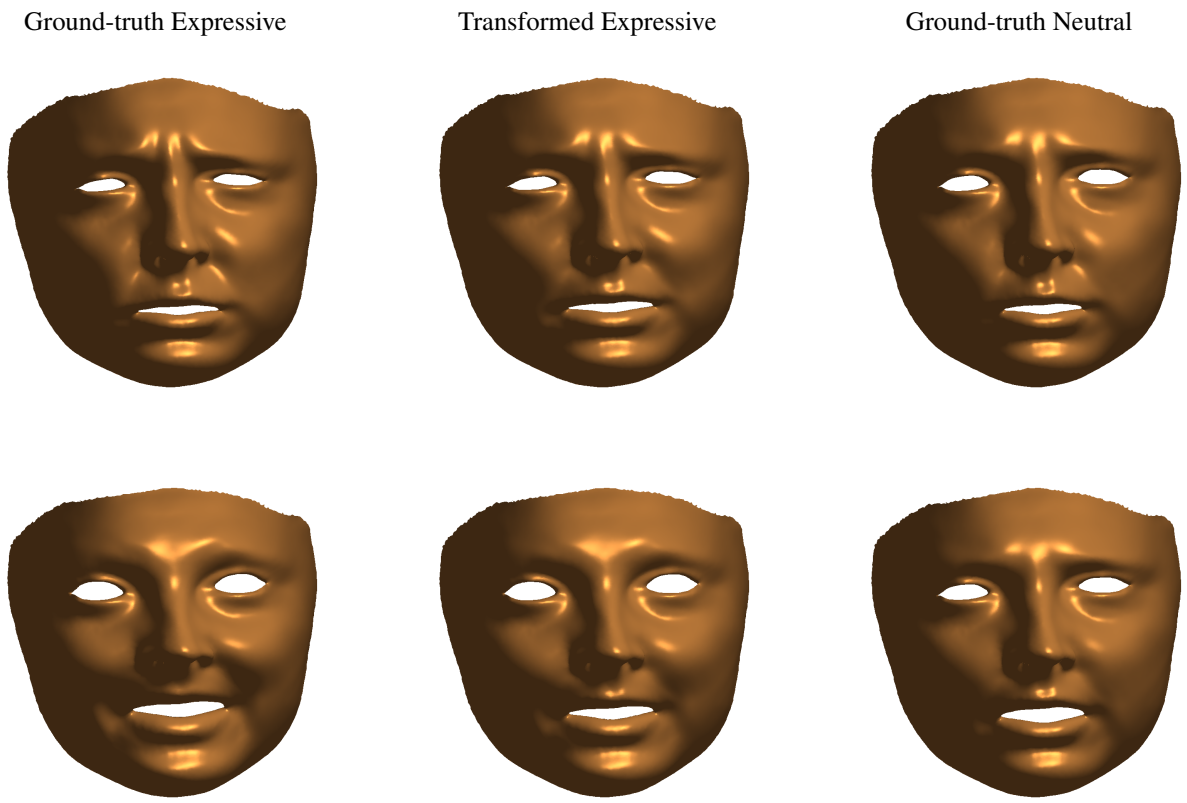


Figure 3: Each row corresponds to an equivalent video frame for (left) real expressive speech time-aligned to (right) real neutral speech. Row 1 shows sadness, and row 2 shows happiness. The neutral versions transformed to expressive (center) display the *style* of the expressive sequences, but with intact visual speech gestures from the neutral sequences.

speech animation," in *Proceedings of SIGGRAPH*, 2002, pp. 388–398.

[4] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," MIT, Cambrige, MA, Tech. Rep. CBCL Paper 224/AI Memo 2003-003, 2003.

[5] Y. Cao, P. Faloutsos, and F. Pighin, "Unsupervised learning for speech motion editing," in *Proceedings of the Symposium on Computer Animation*, 2003, pp. 225–231.

[6] E. Chuang, H. Deshpande, and C. Bregler, "Facial expression space learning," in *Pacific Graphics*, 2002.

[7] E. Chuang and C. Bregler, "Mood swings: Expressive speech animation," in *ACM Transactions on Graphics*, vol. 24, no. 2, 2005, pp. 331–347.

[8] K. Waters, "A muscle model for animating three-dimensional facial expressions," in *Proceedings of SIGGRAPH*, 1987, pp. 17–24.

[9] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proceedings of SIGGRAPH*, 1995, pp. 55–62.

[10] E. Sifakis, A. Selle, A. Robinson-Mosher, and R. Fedkiw, "Simulating speech with a physics-based facial muscle model," in *Proceedings of the Symposium on Computer Animation*, 2006, pp. 261–270.

[11] E. Chuang and C. Bregler, "Performance driven facial animation using blendshape interpolation," Stanford University, Tech. Rep. CS-TR-2002-02, April 2002.

[12] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models." *ACM Transactions on Graphcs*, vol. 24, no. 3, pp. 426–433, 2005.

[13] Y. Cao, E. Faloutsos, P. Kohler, and F. Pighin, "Real-time speech motion synthesis from recorded motions," in *Symposium on Computer Animation*, 2004, pp. 347–355.

[14] A. Hyvärinen, "A family of fixed-point algorithms for independent component analysis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 3917–3920.

[15] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-D Audio-Visual Corpus of Affective Communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[16] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[17] H. Gavert, J. Hurri, J. Sarela, and A. Hyvarinen, "FastICA," p. GNU GPL Version 2, Jan. 2005.