Towards 4D Coupled Models of Conversational Facial Expression Interactions

Jason Vandeventer[1]
VandeventerJM@Cardiff.ac.uk
Lukas Gräser[3]
Lukas.Graeser@Campus.tu-berlin.de
Magdalena Rychlowska[2]
Rychlowska@Cardiff.ac.uk
Paul L. Rosin[1]
RosinPL@Cardiff.ac.uk
David Marshall[1]
MarshallAD@Cardiff.ac.uk

[1] School of Computer Science and Informatics
Visual Computing Group
Cardiff University
Cardiff, Wales, UK

[2] School of Psychology
Cardiff University
Cardiff, Wales, UK

[3] School of Electrical Engineering and Computer Sciences
Berlin Institute of Technology
Berlin, Germany

Abstract

In this paper we introduce a novel approach for building 4D coupled statistical models of conversational facial expression interactions. To build these coupled models we use 3D AAMs for feature extraction, 4D polynomial fitting for sequence representation, and concatenated feature vectors of frontchannel-backchannel interactions (with offset values) for the coupled model.

Using a coupled model of conversation smile interactions, we predicted each sequence's backchannel signal. In a subsequent experiment, human observers rated predicted sequences as highly similar to the originals. Our results demonstrate the usefulness of coupled models as powerful tools to analyse and synthesise key aspects of conversational interactions, including conversation timings, backchannel responses to frontchannel signals, and the spatial and temporal dynamics of conversational facial expression interactions.

1 Introduction

Face-to-face conversations are a frequent and important part of social communication. These conversations, whether with well-known friends or complete strangers, consist of a variety of verbal and non-verbal signals (e.g. expressions, gestures), which determine the tone, content, and flow of a conversation [5, 9, 12, 47].

Dyadic, face-to-face conversations involve repeated exchanges between the listener and the speaker. Input from the listener often serves to control conversational flow [4, 5, 9, 11, 12, 28, 37, 44, 47]. In [47], Yngve coined the term *backchannel* to describe the signals being sent

© 2015. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

from the listener(s) to the speaker. This feedback can indicate comprehension (e.g. a look of confusion), provide an assessment (e.g. saying "correct"), control conversational flow, or even add new content (e.g. sentence completion). Conversely, the term *frontchannel* is used to describe the speaker's behaviour. Obviously, the frontchannel and backchannel roles may be swapped multiple times during a conversation; this dynamic relationship is what allows for the conversation's path to be altered based on expressed and received conversational expressions.

The importance and ubiquity of dyadic, face-to-face conversations in social interaction makes understanding the roles of frontchannel and backchannel signals important for a variety of fields, such as psychology, neuroscience, affective computing, etc. Statistical modelling of such interactions is a particularly promising research area because it allows for quantitative analysis of qualitative data. Modelling each side of a conversation may provide some information about so-called *conversational expressions* (e.g. thinking, confusion, agreement), however, to better understand the effect frontchannel and backchannel expressions have on each other (i.e. content, flow, etc.), coupled statistical models should be built. With coupled models we can analyse the effects one side of a conversation has on the other, understand important characteristics of conversational interactions, and better synthesise conversational expression interactions.

Existing research has used 2D data for modelling conversational aspects [1, 26]. While 2D data is useful for some cases, 3D data offers the advantage of providing intrinsic geometry which is invariant to pose and lighting. Moreover, 3D dynamic (4D) data is preferred over 3D static data because it includes temporal information, which is critical for modelling and synthesising realistic facial expression sequences. No such databases existed of 4D conversations, so our lab created the first 4D (3D video) database of natural, dyadic conversations. Details of this publicly available database can be found in [43].

By building 4D coupled statistical models of conversational expressions we can analyse and synthesise key aspects of frontchannel-backchannel interactions, including conversation timings, backchannel responses to frontchannel signals, and the spatial and temporal dynamics of conversational facial expressions. Such models will allow for advances in many areas, such as behaviour analysis, perceptual psychology, and digitally animated character facial expression modelling and synthesis.

2 Related Work

2.1 3D/4D and Conversational Databases

While many 3D/4D facial expression databases currently exist [10, 33, 40, 48], none contain natural conversations, and as a result, they lack conversational facial expressions; those expressions found more commonly in everyday conversation, such as laughing, thinking, confusion, agreement, etc. While these databases are potentially useful for modelling and synthesis of prototypical facial expressions, they can not be used for our purposes of creating coupled models of conversational expressions.

Some conversational databases exist (e.g. [20, 34, 35, 36, 45, 46]), however, they either do not focus on the face or they are not 4D datasets. Therefore, they are also unsuitable for our research. It is for these reasons our lab created the first 4D (3D video) database of natural, dyadic conversations. This database is used for building coupled statistical models of conversational facial expression interactions.

2.2 Coupled Models

Many statistical based approaches exist for learning and modelling behaviour. Such models that “couple” one behaviour model with another are termed “coupled statistical models”. Coupled statistical models are often used to enhance the information contained in data. Hogg *et al.* [26] was one of the first works to build coupled models of facial expressions. Using tracked data of head shakes and nod movements, a 2D coupled statistical model was built and used to synthesise a nodding head, which would be displayed when a real (non-synthesised) head was making a shaking motion. In [16], 2D coupled-view Active Appearance Models were used to determine the relationship of the frontal-view and profile of the face. Castelan *et al.* used 2D frontal photographs to approximate 3D face shape, by coupling intensity and height information [13]. In [38], Coupled Scaled Gaussian Process Regression (CSGPR) models are used for head pose normalisation, with the goal of head-pose invariant facial expression recognition.

All of these approaches couple actions which occur in the same time instance. The coupled models we build in this paper are sequential in time, as one action influences another. To our knowledge, no work on 4D coupled models of conversational expressions currently exists.

3 Building Coupled Models

Conversations are filled with dynamic facial expressions. These expressions can differ greatly in intensity and length, and their variations encode different, important aspects of conversational interactions. Since our goal is to build coupled statistical models of conversational expression interactions, the sequences will need to be comparable, while still maintaining their expression characteristics.

Finding an appropriate method of representing sequences of varying lengths and characteristics, as single, comparable entities, is a challenge. Approaches like Dynamic Time Warping (DTW) work well for signals which share similar characteristics, such as two smiles made by different people. However, it falls short when those sequences are vastly different. In conversations, the same type of expression, such as smile, may greatly differ in their trajectories due to factors such as the individual speaking before/during the expression, the individual holding what we have termed a “resting smile” (an individual maintaining a masking smile that remains unchanged for long periods of time), or because the individual is transitioning from a different expression to a smile. It is for these reasons that many machine learning approaches also fall short. One solution to this problem is polynomial fitting of the sequence data. It allows for multiple types and lengths of expression sequences to be represented as single, comparable entities and is described in Section 3.1.2.

3.1 Methodology

As stated in Section 1, we created a 4D (3D video) database of natural, dyadic conversations. The conversations were annotated for frontchannel and backchannel conversational expressions by four experienced annotators. A total of 764 frontchannel/backchannel expression periods were annotated. The scope of this paper does not include the details of this database, but full details can be found in [43]. These manual annotations provide the segmented data used for the experiments in Section 4.

To build an Active Appearance Model (AAM) [17] of conversation sequences, the data must be inter-subject registered. Sparse correspondence is achieved using a 4D tracking approach which uses 3D shape and texture. These tracked points are used as control points in a dense correspondence method. This method uses a Thin Plate Spline (TPS) based algorithm, with an additional "snapping" step, to modify the geometry of one mesh (reference mesh) so that it matches that of another mesh (target mesh). The tracking and inter-subject registration methods were developed in-lab and details for these approaches can be found in [23].

3.1.1 3D AAMs

Frontchannel and backchannel sequences of inter-subject registered data are used to build a 3D AAM. Each frame of a sequence is projected into the combined AAM model. This provides us with the *bVectors*, which are the principal component (PC) weights for the projected frame ("Original bVectors" in Figure 1). These bVectors specify the shape and texture features for each projected frame At this point, the sequence consists of bVector values for each individual frame. In order to represent the sequence as a continuous, length-invariant entity, polynomial fitting is used.

3.1.2 Polynomial Fitting

Before a polynomial fit is calculated on a sequence of bVector values (for each principal component), the sequence is first standard score normalised [30] and shifted to the mean. This is an important step for retaining each sequence's characteristics, while also ensuring all of the sequences reside in the same polynomial space. The importance of this becomes evident when we use the coupled models to predict sequence values.

An n^{th} degree polynomial is fitted to the sequence of bVectors, for a specific PC ("Original Fit" in Figure 1). In the resulting polynomial equation, the coefficients make up the feature vector which is used as input into the coupled model. This feature vector is concatenated with the normalisation and shifting values, as well as the number of frames in the sequence. This information helps for reversing the process when synthesising the sequence. The feature vectors for each PC are concatenated to produce a single, combined feature vector, which describes all parts of the expression sequence.

One of the major benefits to this approach is that it allows for the modification of sequences, such as changing the length or amplitude of an expression. For instance, to extend a sequence of i frames, you simply insert $i+1$ for each equation variable. Assuming the PC sequence represents expression variations, to amplify facial expressions you simply amplify the polynomial curve and calculate the new bVector values, as seen in Figure 1 as "70% Decreased" and "70% Increased". Projecting the bVectors out of the AAM models allows for the synthesis of newly modified sequences while, importantly, retaining the expression characteristics.

The main strength of this approach, however, is that it allows sequences of different lengths and characteristics to be represented by the same number of values, which is critical for building our coupled models of conversational interactions.

3.1.3 Coupled Models

For each frontchannel-backchannel interaction, the combined feature vectors for the frontchannel signal, an offset For each frontchannel-backchannel interaction, the combined feature

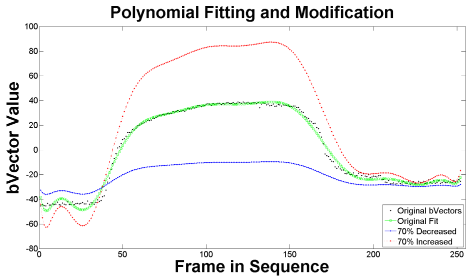


Figure 1: Original bVectors with a polynomial fit and two amplitude-modified curves

	Sequences			
FC	PC 1	PC 1	PC 1	PC 1
	PC 2	PC 2	PC 2	PC 2
	PC 3	PC 3	PC 3	PC 3
	Offset	Offset	Offset	Offset
BC	PC 1	PC 1	PC 1	PC 1
	PC 2	PC 2	PC 2	PC 2
	PC 3	PC 3	PC 3	PC 3

Table 1: Example of the coupled model's feature vectors.

vectors for the frontchannel signal, an offset value for the interaction, and the combined feature vectors of the backchannel signal are concatenated to produce a coupled feature vector. The combined feature vectors were calculated according to the process described in Section 3.1.2, and the *offset* value is the number of frames from the start of the frontchannel expression to the start of the reacting backchannel expression. The coupled model consists of these coupled feature vectors and now describes the interactions between a frontchannel signal and a backchannel signal, as a single feature vector. Table 1 provides a visualisation of the coupled model. This model can be used for a variety of applications. In this work we use it to predict backchannel responses to frontchannel signals, using k-Nearest Neighbour (kNN) imputation [3, 18].

3.1.4 Predicting Frontchannel and Backchannel Signals

kNN imputation is a popular approach, especially when using sparse data sets, for replacing missing data [3, 15, 24, 42]. This approach replaces missing data with a weighted-mean of the *k*-nearest columns, where the weight is inversely proportional to the distance from those neighbours. In our coupled model, each column is the feature vector for a single interaction. Given the feature vector values of a frontchannel sequence and using a coupled model of conversational interactions, kNN imputation can be used to predict the values of a backchannel feature vector. The same is true for predicting frontchannel feature vector values. This imputation process is performed on the data for use in Experiments 3 and 4 (Section 4.4).

4 Experiments

The steps described thus far allow us to predict the characteristics of a backchannel signal using the previously modelled frontchannel-backchannel interactions. To fully evaluate the effectiveness of our modelling and synthesis approach, we performed three experiments.

In Experiment 1 (Section 4.2) classified frontchannel and backchannel expressions. This was done to not only show that our in-lab tracking and inter-subject registration approach is sound, but also to demonstrate that both frontchannel and backchannel signals have intrinsic, separable properties.

In Experiment 2 (Section 4.3) we built 4D models of backchannel sequences, manipulated their amplitudes, and synthesised the manipulated sequences. These sequences were then evaluated by human observers to measure the effect of the modification on perceptions of realism, both for facial expression and image quality. Results of this experiment show that the models we built are perceived as realistic and believable, even after being modified.

In Experiment 3 (Section 4.4.1) and Experiment 4 (Section 4.4.2), we built a 4D coupled statistical model of smiles reciprocated during conversations. We use these models to predict, for each sequence, the characteristics of backchannel reactions to frontchannel signals. In Experiment 3, we classified the predicted frontchannel and backchannel sequences, using the original sequences as the training data. In Experiment 4, human observers viewed the original and predicted sequences. Their task was to rate the extent to which a predicted sequence was similar to the original.

4.1 Conversational Data

In addition to conveying complex and rich information, smiles appear in face-to-face interactions more frequently than other facial expressions [6]. They also constitute important backchannel signals in face-to-face conversations, similarly to other signals described in the literature [8, 47]. Consequently, the classification and perception experiments described below focus on smiles during naturalistic conversations.

Using the manually annotated dataset described in Section 3.1, interactions consisting of a frontchannel (FC) smile expression with a corresponding backchannel (BC) smile expression (occurring within 2 seconds of the FC smile), were selected. This resulted in 22 conversation interactions (44 sequences), which were tracked and registered using the process described in Section 3.1.

4.2 Experiment 1 - Classification

In this experiment we attempted to differentiate frontchannel from backchannel smile sequences, using 3D AAMs for feature extraction, polynomial fitting for 4D sequence representation, and Support Vector Machines (SVMs) for classifying the 4D sequences. For each subject, Sub$_{\text{target}}$, a 3D AAM was built using all sequence frames from every other subject, Sub$_{\text{others}}$. 95% of the eigenenergy was kept. For each sequence, bVectors (feature vectors) were calculated by projecting every frame into the AAM.

An n^{th} degree polynomial fit was performed on each sequence of bVectors. This approach allowed for a 4D representation of 3D discrete data. A grid search was performed to empirically find an appropriate polynomial degree and number of principal components to use for fitting, for each Sub$_{\text{target}}$ AAM model.

The polynomial coefficients were used as input into a Support Vector Machine (SVM) classifier, *libSVM* [14], where Sub$_{\text{others}}$ sequences comprised the training set, and Sub$_{\text{target}}$ sequences comprised the testing set. A ν-SVM with a Gaussian RBF kernel was used, and a grid search was performed for parameter optimisation, as suggested in [27, 41].

As stated above, these steps were performed for each subject, so as to provide a fully-generalised approach to classification. That is, this is a subject-independent, cross-validation approach.

4.2.1 Results and Analysis

For classification accuracy, Area Under the ROC Curve (AUC) was chosen as the performance metric because it has been shown to be more reliable and contain more preferable properties than raw classification accuracy, as described in [2, 7, 32]. The average accuracy for all four subjects was 97.54%.

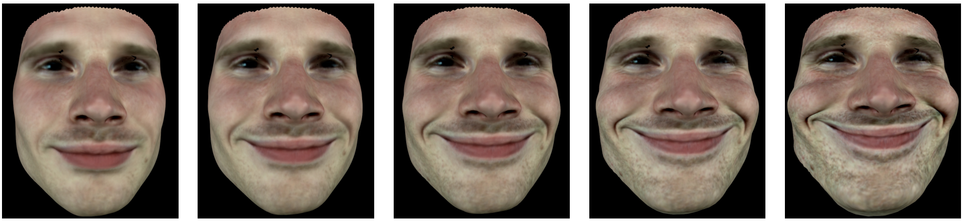
Experiment 1 was able to validate two main points. First, frontchannel and backchannel signals contain characteristics which allow them to be differentiated; this was most likely the vertical movement of the mouth of the speaker (frontchannel signal). Second, the positive results support the use of our in-lab tracking and inter-subject registration methods, as any tracking or registration issues would have resulted in warped shape and texture and led to poorer classification results. It was extremely important to confirm the quality of these approaches before undertaking the modelling and synthesis steps described in Section 4.3.

4.3 Experiment 2 - Model Modification

Experiment 2 focused on the realism of synthesised backchannel sequences, for both facial expression and rendered sequence image quality. 3D AAMs were built for every smile sequence (95% eigenenergy kept). By building an AAM for each expression sequence, we remove the identity variations that occur in multi-subject AAMs, and increase the probability that the top PC values will represent variations in expression. The fitting process described in Section 3.1.2 was performed using a 14$^{\text{th}}$ degree polynomial for three principal components. These three PCs represented, on average, 85% of the remaining model energy. After observing the output from various combinations of PCs, we chose to use the top three PCs because they provided the best balance between the number of PCs used and capturing the data and variability required for our experiments. That is, by reducing the number of PCs we are able to reduce processing time, while still retaining the quality and variation of data we require for our work. For this approach, since each sequence is completely independent, no normalisation or shifting step is performed. In addition to the independent nature, and since the goal is the creation of realistic stimuli, over-fitting is not a concern.

As well, it is true that oscillations tend to occur at the edges of a polynomial curve for high-order polynomials (*Runge's phenomenon* [39]), and a piecewise polynomial fitting approach (such as the use of *B-Splines* [19]) might be better for fitting these sequences. To produce the expected visual output there is more benefit to using a single polynomial function for fitting, rather than multiple sections of polynomial functions. For instance, amplifying each polynomial section of a piecewise polynomial-fitted smile sequence will not result in an overall amplified sequence, but rather disjointed amplified parts (which will not have the appearance of a normal, but more intense, smile). Each backchannel sequence's polynomial fit was modified using four amplitude values: 70% decreased, 30% decreased, 30% increased, and 70% increased. From these new polynomial curves, bVector values for each PC were calculated for each modification type, as described in Section 3.1.2 and shown in Figure 1. The original amplitude sequence was produced using the original polynomial coefficients and acted as a ground-truth of sorts. Figure 2 shows examples of the same peak frame for each modified sequence. The frontchannel sequence uses its original polynomial fitted values. It is worth noting that while the amplitude of the backchannel smile expressions were modified, the expression dynamics of the individual were preserved (Figure 2).

These videos of frontchannel-backchannel smile interactions were used in a subsequent experiment, in which 28 participants (17 male and 11 female) viewed the video sequences and evaluated the realism of the backchannel sequence(both for expression and for image quality), using a 4-point Likert-type scale ranging from (1) Not at all realistic to (4) Highly realistic. Before starting the task, participants were shown examples of realistic and unrealistic backchannel sequences. These example sequences were not included in the sequences



(a) 70% Decreased (b) 30% Decreased (c) Original (d) 30% Increased (e) 70% Increased

Figure 2: The peak expression frame for each modified sequence (same frame number).

evaluated by the participants during the task. Each sequence evaluated showed the original frontchannel sequence and one of the five amplitude-modified sequences, with the interaction offset preserved. Each participant thus rated 5 versions of 11 sequences – the original and the four modified versions – for a total of 1540 trials (55 trials per person). The interactions were shown to participants in a random order, using the efficient and unbiased *Durstenfeld-Knuth* shuffling algorithm [21, 29].

4.3.1 Results and Analysis

Four trials from three participants were discarded from the analysis due to web-based input errors which resulted in missing data. Both expression realism and image quality ratings were significantly affected by the manipulation of amplitude, $F(4,108) = 24.69$, $p < .001$ and $F(4,108) = 26.04$, $p < .001$, respectively. Perceived realism of facial expressions displayed in the original videos, in the videos using lower levels of amplitude, and in the videos using the 130% amplitude level was significantly higher than the scale midpoint (2.5, all $t's > 3.5$, $p's < .01$, Bonferroni corrected). These versions were therefore perceived as highly realistic. The high-amplitude (170%) level, however, was perceived as less realistic and not significantly higher than the scale midpoint ($t(27) = -1.19$, $p > .1$). Table 3(a) shows the average rating and standard deviation for expression realism, for each modification level. An identical pattern was observed for the ratings of image quality (Table 3(b)).

Modification	Avg. Rating	S.D.
30%	2.95	.62
70%	3.14	.49
100%	3.00	.43
130%	2.83	.41
170%	2.40	.43

(a) *Expression Realism* Ratings

Modification	Avg. Rating	S.D.
30%	2.98	.53
70%	3.04	.44
100%	2.99	.45
130%	2.84	.44
170%	2.43	.52

(b) *Image Quality* Ratings

Figure 3: Experiment 2 Results

Figure 4 shows the values plotted for each modification level, for expression realism ratings and image quality ratings. The red line in the figure represents the rating scale's mid-point. Thus, any values above this line represent stimuli that were perceived as realistic.

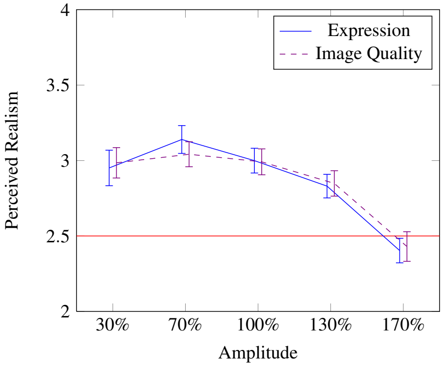


Figure 4: Experiment 2 Results

The observed decline in perceived realism of smiles as the amplitude/velocity increased is consistent with previous research linking short smile onset and offset with decreased perceptions of genuineness [22, 25, 31].

Having shown that we can model, modify, and synthesise realistic expression sequences separately, we move on to joining them in a coupled statistical model, used for predicting backchannel expressions given frontchannel stimuli.

4.4 Experiments 3 & 4 - Predicting Expressions

Experiment 3 and Experiment 4 focused on evaluating our coupled statistical modelling approach, specifically the similarity of original and predicted backchannel smile expressions. We built an AAM which contained all sequence frames for every subject (95% eigenenergy kept). The fitting process described in Section 3.1.2 was performed using a 7$^{\text{th}}$ degree polynomial for three principal components, which represented 77.63% of the remaining model energy. After observing the output from various combinations of PCs, we chose to use the top three PCs because they provided the best balance between the number of PCs used and capturing the data and variability required for our experiments. That is, by reducing the number of PCs we are able to reduce processing time, while still retaining the quality and variation of data we require for our work. The polynomial degree chosen was empirically determined, so as to avoid over-fitting. The coupled model was built according to the steps described in Section 3.1.3, and the imputation step was performed for the polynomial coefficients using an empirically determined k value of 15 for kNN. The shift and normalisation values (discussed in Section 3.1.2) were not removed as part of the imputation step. In the future we would like to be able to impute all parts (polynomial coefficients, shifting, and normalisation values) of the sequence feature vectors, however, more data is needed for predicting accurate enough shifting and normalisation values for re-synthesis purposes. These values are useful for better matching the appearance of the identity of the subjects.

For these experiments we used the data from our 4D conversational database. Clearly, for these interactions there is no missing data. Therefore, to predict a sequence's PCs, the data to be predicted is removed before imputation. That is, the implementation is such that the data is imputed on a PC-by-PC basis (i.e. PC 1, then PC 2, then PC 3, until all PC values have been imputed). The original data that has been removed acts as a ground truth and is used to help evaluate how well we predicted the values. The ground truth is used in the classification experiment (Section 4.4.1) as the training set, and is used to synthesise the original backchannel sequences for the human perceptual study (Section 4.4.2).

4.4.1 Experiment 3 - Classifying Predicted Sequences

To analytically evaluate our coupled model approach we classified frontchannel and backchannel predicted sequences. The methodology was identical to the one we used in Experiment 1 (Section 4.2). The training set was comprised of 22 frontchannel and 22 backchannel original sequences (i.e. polynomial coefficients used for building the coupled model). The testing set was comprised of 22 frontchannel and 22 backchannel predicted sequences. The optimal ν-SVM parameters were: $Cost = -10$, $\nu = 0.95455$. Classification accuracy (Raw and AUC) was 95.45%. The two incorrectly classified sequences were false-positives (i.e. predicted as frontchannel when actually backchannel sequences).

4.4.2 Experiment 4 - Perceptual Experiment

The imputed/predicted backchannel sequences were synthesised by projecting the new bVector values out of the AAM. The frontchannel sequences and the original backchannel sequences were synthesised using the original polynomial fit values. Participants (see Section 4.3) were instructed to rate the extent to which they perceived the imputed backchannel expressions as similar to the original expressions. They used a 4-point Likert-type scale ranging from (1) Very Dissimilar to (4) Very Similar. Each participant rated 11 sequences for a total of 308 trials. One trial was discarded from the analysis due to web-based input errors which resulted in missing data. We also discarded one trial for which the participant spent only 1.34 seconds (1.667 seconds being the length of the shortest backchannel video to rate), for a final sample of 306 trials. Similarity ratings, averaged within participants, were significantly higher than the scale midpoint (2.5), $M = 2.90$, $SD = 0.31$, $t(27) = 7.00$, $p < .001$, suggesting that participants consistently perceived the imputed videos as similar to the original versions.

5 Conclusion

In this paper we introduced a novel approach for building 4D coupled statistical models of conversational facial expressions. Using 3D AAMs for feature extraction, 4D polynomial fitting for sequence representation, and combined feature vectors of frontchannel-backchannel interactions, we built a 4D coupled model of smile expressions from conversational interactions. This model served to predict and synthesise backchannel expression sequences, which were used in a perceptual experiment. The results of this experiment support the use of the proposed methods. While this approach to building coupled models of conversational expression interactions is clearly promising, more evaluations are needed. This includes using different frontchannel-backchannel expression types, adding more subjects to the models, and conducting experiments which evaluate different attributes of the interactions. The approach described in this paper allows for the creation of realistic, modifiable stimuli that would not otherwise be possible. Therefore, future work will focus on utilising these coupled models to analyse and synthesise a larger variety of conversational expression interactions.

References

[1] Andrew J. Aubrey, Douglas W. Cunningham, David Marshall, Paul L. Rosin, AhYoung Shin, and Christian Wallraven. The Face Speaks: Contextual and temporal sensitivity to backchannel responses. In *The Asian Conference on Computer Vision (ACCV) Workshop on Face analysis: The intersection of computer vision and human perception*, pages 248–259. Springer, 2012.

[2] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Automatic Face and Gesture Recognition (FG 2006), IEEE International Conference on*, pages 223–230. IEEE, 2006.

[3] Gustavo E.A.P.A. Batista and Maria Carolina Monard. A study of k-nearest neighbour as an imputation method. In *Hybrid Intelligent Systems (HIS), 2002 International Conference on*, volume 87, pages 251–260. IOS Press, 2002.

[4] Janet B. Bavelas, Alex Black, Charles R. Lemery, and Jennifer Mullett. "I show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, 50(2):322–329, 1986.

[5] Janet B. Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[6] Janet Beavin Bavelas and Nicole Chovil. Faces in dialogue. In James A Russell and Jose Miguel Fernández-Dols, editors, *The psychology of facial expression*, chapter 15, pages 334–346. Cambridge University Press, 1997.

[7] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[8] Lawrence J. Brunner. Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5):728–734, 1979.

[9] Peter Bull. State of the art: Nonverbal communication. *The Psychologist*, 14(12): 644–647, 2001.

[10] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *Visualization and Computer Graphics (TVCG), IEEE Transactions on*, 20(3):413–425, 2014.

[11] Justine Cassell and Kristinn R Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999.

[12] Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1):55–64, 2001.

[13] Mario Castelán. *Face Shape Recovery from a Single Image View*. PhD thesis, University of York, 2006. URL `http://www.bmva.org/thesis-archive/2006/2006-castelan.pdf`.

[14] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *Intelligent Systems and Technology (TIST), ACM Transactions on*, 2(3):27:1–27:27, 2011.

[15] Hugh Chipman, Trevor J Hastie, and Robert Tibshirani. Clustering microarray data. In Terry Speed, editor, *Statistical Analysis of Gene Expression Microarray Data*, chapter 4, pages 159–200. Chapman & Hall/CRC, 2003.

[16] Timothy F. Cootes. Model-based methods in analysis of biomedical images. In R. Baldock and J. Graham, editors, *Image Processing and Analysis: A Practical approach*, chapter 7, pages 223–248. Oxford University Press, 1999.

[17] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *Pattern Analysis & Machine Intelligence (PAMI), IEEE Transactions on*, 23 (6):681–685, 2001.

[18] Brenda G Cox. The weighted sequential hot deck imputation procedure. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pages 721–726. American Statistical Association, 1980.

[19] Carl De Boor. A practical guide to splines. *Mathematics of Computation*, 1978.

[20] Iwan de Kok and Dirk Heylen. The MultiLis corpus – Dealing with individual differences in nonverbal listening behavior. In Anna Esposito, Antonietta M. Esposito, Raffaele Martone, VincentC. Müller, and Gaetano Scarpetta, editors, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, volume 6456, pages 362–375. Springer, 2011.

[21] Richard Durstenfeld. Algorithm 235: Random Permutation. *Communications of the ACM*, 7(7):420, 1964.

[22] Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.

[23] Lukas Gräser, Jason Vandeventer, Job van der Schalk, Paul L. Rosin, and David Marshall. 4D tracking and inter-subject registration for the synthesis of realistic facial expression sequences. *Under Review*, 2015.

[24] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays. Technical report, Division of Biostatistics, Stanford University, 1999.

[25] Holger Hoffmann, Harald C Traue, Franziska Bachmayr, and Henrik Kessler. Perception of dynamic facial expressions of emotion. In Elisabeth André, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, and Michael Weber, editors, *Perception and interactive technologies*, volume 4021, pages 175–178. Springer, 2006.

[26] David Hogg, Neil Johnson, Richard Morris, Dietrich Buesching, and Aphrodite Galata. Visual models of interaction. In *Proceedings of the 2nd International Workshop on Cooperative Distributed Vision*, 1998.

[27] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[28] Ellen A Isaacs and John C Tang. What video can and cannot do for collaboration: A case study. *Multimedia Systems*, 2(2):63–73, 1994.

[29] Donald E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, third edition, 1997.

[30] Erwin Kreyszig and Edward J Norminton. *Advanced engineering mathematics*. Wiley, fourth edition, 1979.

[31] Eva Krumhuber and Arvid Kappas. Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior*, 29(1): 3–24, 2005.

[32] Charles X Ling, Jin Huang, and Harry Zhang. AUC: A statistically consistent and more discriminating measure than accuracy. In *IJCAI*, volume 3, pages 519–524, 2003.

[33] Bogdan J Matuszewski, Wei Quan, Lik-Kwan Shark, Alison S Mcloughlin, Catherine E Lightbody, Hedley CA Emsley, and Caroline L Watkins. Hi4D-ADSIP 3-D dynamic facial articulation database. *Image and Vision Computing*, 30(10):713–727, 2012.

[34] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The SEMAINE Database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

[35] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. Predicting Listener Backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, volume 5208, pages 176–190. Springer, 2008.

[36] Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28, 2013.

[37] Isabella Poggi and Catherine Pelachaud. Performative facial expressions in animated faces. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth F. Churchill, editors, *Embodied Conversational Agents*, chapter 6, pages 155–189. MIT Press, 2000.

[38] Ognjen Rudovic, Maja Pantic, and Ioannis Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on*, 35(6):1357–1369, 2013.

[39] Carl Runge. Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten (About empirical functions and the interpolation between equidistant ordinates. *Zeitschrift für Mathematik und Physik (Journal for Maths and Physics)*, 46 (224-243):20, 1901.

[40] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[41] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[42] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[43] Jason Vandeventer, Andrew J. Aubrey, Paul L. Rosin, and David Marshall. 4D Cardiff Conversation Database (4D CCDb): A 4D database of natural, dyadic conversations. In *Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, 2015.

[44] Roel Vertegaal. Conversational awareness in multiparty VMC. In *Extended Abstracts on Human Factors in Computing Systems (CHI 1997)*, pages 496–503. ACM, 1997.

[45] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops (ACII 2009), 3rd International Conference on*, pages 1–4. IEEE, 2009.

[46] Ekaterina P Volkova, Betty J Mohler, Trevor J Dodds, Joachim Tesch, and Heinrich H Bülthoff. Emotion categorization of body expressions in narrative scenarios. *Frontiers in Psychology*, 5:623, 2014.

[47] Victor H Yngve. On getting a word in edgewise. In *Papers from the 6th Regional Meeting of the Chicago Linguistic Society*, pages 567–578, 1970.

[48] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3D dynamic facial expression database. In *Automatic Face and Gesture Recognition (FG 2013), IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.