Speaking faces for face-voice speaker identity verification

Girija Chetty and Michael Wagner

School of Information Sciences and Engineering
University of Canberra, Australia
girija.chetty@canberra.edu.au

Abstract

In this paper, we describe an approach for an animated speaking face synthesis and its application in modeling impostor/replay attack scenarios for face-voice based speaker verification systems. The speaking face reported here learns the spatio-temporal relationship between speech acoustics and MPEG4 compliant facial animation points. The influence of articulatory, perceptual, and prosodic acoustic features along with auditory context on prediction accuracy was examined. The results are indicative of vulnerability of audiovisual identity verification systems to impostor/replay attacks using synthetic faces. The level of vulnerability depends on several factors, such as the type of audiovisual features, the fusion techniques used for the audio and video features and their relative robustness. Also, the success of the synthetic impostor depends on the type of co-articulation models and acoustic features used for the audiovisual mapping of speaking face synthesis.

Index Terms: speaking face synthesis, identity verification, audio-visual mapping

1. Introduction

Speaker verification systems based on multimodal fusion of face-voice biometric information rate high in terms of user acceptance and deployment costs, due to less intrusiveness and to the ease of availability of low-cost off the shelf system components. However, the very use of user friendly face and voice biometric traits exposes the system to a different type of impostor attacks called "synthesis attacks", as compared to other biometrics such as retina or iris [1] because of the ease of accessing face and voice information for a speaker. In addition, with advances in computer graphics and animation technology, it has become relatively easy to create an artificial speaking face of the client with a high degree of photorealism [2] from prerecorded photographs and speech, and to surreptitiously replay a client's synthesized speaking face in order to access a secure facility.

Most of the speaking face synthesis techniques proposed so far are in the area of interactive entertainment and human-computer interaction, e.g. [2, 3] and although they produce photo-realistic perceptually pleasing animations, are not
specifically tailored for the investigation of impostor attack scenarios. To combat different types of potential impostor attacks, it is necessary to study the modeling of imposture techniques and scenarios, and present these scenarios for testing the identity verification systems. We argue that, if a well structured approach is used for testing and evaluation using tailored, realistic impostor/replay attack scenarios, it is possible to minimize the chances for fraudulent attacks on identity verification systems. Though there has been much reported work in the speaker verification literature on identity verification techniques, such as feature extraction, selection, enhancement, transformation, and classification techniques [4,5], there has been limited work on modeling realistic impostor/replay attack scenarios and techniques for testing the ability of identity verification systems to thwart fraudulent attacks, and evaluate any new security measures proposed [6.7] to combat such attacks.

A novel biometric information synthesis framework allowing the systematic testing and evaluation of face-voice speaker identity verification techniques against impostor and replay attacks is currently in progress in the Human Computer Communication Laboratory at the University of Canberra. The speaking face synthesis framework is based on data-driven facial deformation analysis and subsequent speech-to-facial-deformation mapping by using face-voice information from different types of data corpora featuring 2D and 3D, broadcast news speaking faces belonging to several race, gender and age groups. The artificial speaking face synthesized evolves progressively through every functional stage of the framework with basic lip-synch abilities at the lowest level for modeling simple replay attack scenarios. In subsequent stages, the speaking face has more realistic humanlike abilities such as nonverbal gestures including head motion and facial emotions/ expressions associated with speech production similar to real human speaking faces, in an attempt to capture more of the individual nuances of the speaker.

In this paper we report the investigations related to one of the stages of the speaking face synthesis framework, which models simple animated replay attack scenarios for face-voice based identity verification. In this approach, a Hidden Markov Model (HMM) learns to predict speech driven lip-motions from real speaking face video of the person. By examining previous methods used for speaking face synthesis, [1,2,3], the approach proposed here is tailored to address impostor/replay attack scenarios, and is based on the following considerations:

1. *Context Coding*: Context information directly encoded into the acoustic representation can increase the prediction performance of the A/V mapping. Although HMMs have the modeling power to learn context information, explicitly incorporating acoustic context delta vectors results in a consistent improvement of predictive accuracy.

2. *Acoustic fusion features*: Different acoustic representations carry complementary information regarding orofacial motion. The fusion of all features modeling articulatory (LPC, LSF), perceptual (MFCC, PCBF), and prosodic information (E, F0) can provide improved synthesis performance as compared to individual feature vectors.

3. *LDA on acoustic fusion vector*: With the fusion of all the acoustic feature vectors, the feature space would be too large for satisfactory HMM performance, hence we use linear discriminant analysis (LDA) in order to reduce the number of feature space dimensions.
4. *HMM model parameters*: We chose 35 model states with two Gaussian mixtures and diagonal covariance matrices, as there are around 35 visemes for spoken English. The speaking face sequences from the 24 male and 19 female speakers of the VidTimit corpus [9] were used for learning the correspondence between acoustic and visual features.

2. Feature extraction

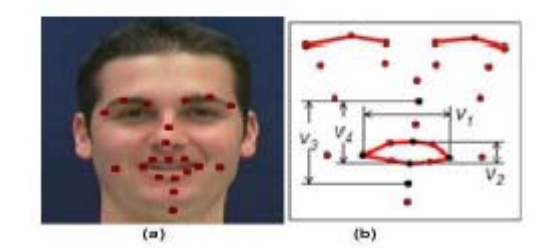

Figure 1: (a) Neutral face of the subject with visual markers, (b) reconstructed markers and orofacial features MW(=V_1), MH(=V_2), CH(=V_3) and LL(=V_4)

2.1. Audio features

Speech signals were down-sampled to 16 kHz, processed in blocks of 16 milliseconds, pre-emphasized, and Hamming windowed. The purpose of feature extraction was to determine robust acoustic features that are predictive of orofacial motion. Unfortunately, not much is known about which acoustic features are relevant for speaking face synthesis. Therefore, three general types of features, each providing a different characterization of the speech signal, are selected. They are the prosodic group comprising fundamental frequency (F0) and signal energy (E); the segmental group comprising linear prediction coefficients (LPC) and line spectral frequencies (LSF); and the perceptual group comprising mel-frequency cepstral coefficients (MFCC) and perceptual critical band features (PCBF). E and F0 were transformed to a logarithmic scale; each of the other features formed a 12-dim vector. The left and right context was incorporated into feature extraction by determining a 12-dim delta vector for each 12-dim feature vector by linear regression of an 11frame neighborhood.

2.2. Video Processing

To facilitate accurate tracking of the facial dynamics, 27 markers
are placed on the subject's face at various positions defined by the MPEG4 standard [3]. The 3D coordinates of these markers are then recovered through stereo vision from front and profile views [6]. This process yields a video vector with 81 (27×3) measurements, which are highly correlated since movements of the points on the face are clearly interdependent.
For this reason, principal component analysis (PCA) is used to project the redundant data onto a lower dimensional space that preserves most of the relevant motion. Based on previous results [1], we use four principal components as the reduced video vector for the HMM training. The coordinates of the 27 markers are then recovered from the four principal components through least squared back-projection. Figure 1(a) shows a neutral face with some of the MPEG4 markers. Figure 1(b) shows the derivation of four representative orofacial articulatory parameters from the marker positions, namely mouth width (MW), mouth height (MH), nose-to-chin height (CH) and nose-to-lower-lip height (LL). The prediction performance of our system for the facial movements will be determined in terms of these 4 orofacial parameters.

3. HMM based facial parameter generation

A hidden Markov model (HMM) was trained on the joint A/V space, by combining the audio and visual features into one joint observation vector. Once the joint HMM is trained, the extraction of an audio HMM is trivial since the audio parameters are part of the joint A/V distribution. To synthesize a video vector from a new audio input, the LMSHMM method operates in two stages. First, the most likely state sequence is found based on the learned audio HMM. Then, the audio input at and the Gaussian mixture model corresponding to each state *qt* is used to analytically derive the visual estimate $\hat{v}_t$

$$\hat{v}_t = \arg\max P(a_t, v_t, q_t \mid \lambda)$$

by way of an "inverted" Baum Welch re-estimation algorithm [14].

3.1. FAP animation

To verify the accuracy of the tracking method and the resulting predictions, we reconstruct the faces from facial animation parameters (FAP) using an MPEG4 face animation engine from the Xface toolkit [12]. Xface is a set of open-source tools for the creation of MPEG4 talking heads from FAPs. Xface FAE is a high level interface capable of animating MPEG4 compliant faces at high frame rates in synchrony with an audio track .Figure 2 shows the *Alice* character from Xface, which is used for mapping the predicted lip motion from VidTIMIT sentences.

Figure 2: Open source facial animation engine (FAE) Xface [12] and virtual character Alice

4. Experimental Results

The speaking face data for the audio-visual mapping experiments comprised utterances from the VidTIMIT corpus. Session I has 6 sentences per person, and Sessions II and III have two sentences each. Session I was used for training, Session II for validation and Session III for testing. Each sentence was repeated 10 times. Thus there were 60 sentences in the training set, 20 sentences in the validation set and 20 sentences in the test set. The use of validation and test sets allows us to identify an upper and lower bound of predictive accuracy. Two performance metrics were used to measure the predicted accuracy: normalized mean-squared error ε and correlation coefficient ρ between the predictions and the true trajectories, defined by:

$$\varepsilon = \frac{1}{\sigma_v^2}\frac{1}{N}\sum_{k=1}^{N}(\hat{v}(k) - v(k))^2$$

$$\rho = \frac{1}{N}\sum_{k=1}^{N}\frac{(\hat{v}(k) - \mu_{\hat{v}})(v(k) - \mu_v)}{\sigma_{\hat{v}}\sigma_v}$$

4.1. Influence of the acoustic features

The various acoustic features e.g. prosodic, articulatory and perceptual – all capture useful information for predicting orofacial motion [1]. Therefore, it is worthwhile to investigate the performance of different acoustic features first. Energy and fundamental frequency together form a 6-dim vector EF0 (2 features/frame × 3 frames), whereas each of the 4 feature groups LPC, LSF, MFCC and PCBF forms a 36-dim vector (12 coefficients/frame × 3 frames). The full acoustic fusion vector AFN therefore has 4×36+6=150 dimensions. After LDA projection, the fusion vector dimension reduces to 10.

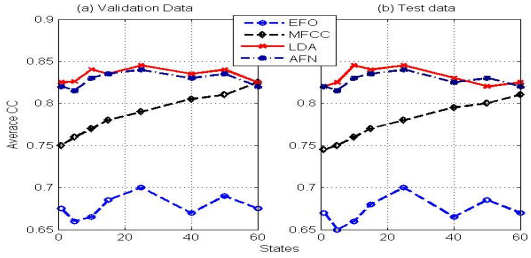

Figure 3: between different audio feature groups and the orofacial features for validation and test data sets.

The prediction performance in terms of the average correlation coefficients on the three orofacial parameters of some feature groups (EFO, MFCC), their combination AFN(LPC, LSF, MFCC and PCBF), and its LDA projection is shown in Figure 3. It can be noted that AFN and LDA, present similar performance on the validation data, followed by MFCC and EF0. On the test data, AFN and LDA produce the highest performance, followed by LSF, MFCC, PCBF and LPC and EF0(only MFCC and EF0 shown in figure 3 for clarity). EF0 consistently provides the lowest predictive accuracy, which confirms that prosody is less related to orofacial articulators than phonetic content. Considering these results, one can draw the conclusion that both AFN and LDA are capable of extracting information that generalizes well for phonetic sequences not included in the training set, even though the LDA transformation effects a 15:1 reduction in dimensionality. For this reason, further investigation will focus on HMMs trained with LDA features.

4.2. Predicted and actual lip trajectories

An example of the predicted and original trajectories of the 4 orofacial articulators, obtained from the LDA acoustic vector, is shown in Fig. 4 for the 2 VidTIMIT sentences:

(1) "*Please dig my potatoes up before frost*"
(2) "*I'd ride the subway, but I haven't enough change*"

The first sentence is from the validation set and the second sentence is from the test set. The predicted trajectories were passed through a 3 frame mean filter to remove jitter.

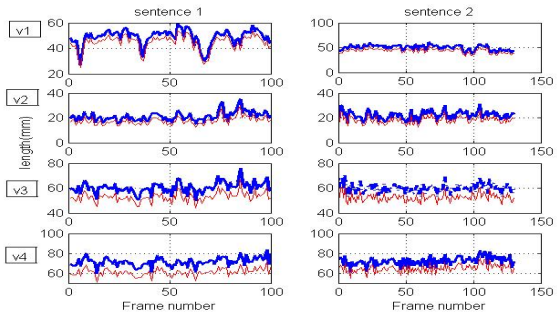

Figure 4: Predicted (thin red) versus actual (thick blue trajectories from the LDA acoustic vector for the four articulatory parameters.

4.3. Lip parameter prediction

Table 1 shows the performance of the HMM mapping technique for predicting lip articulatory parameters for different audio features on validation and test data. The best performer for each lip articulatory parameter is highlighted. Regardless of the type of acoustic feature vector, the correlations indicate that the mouth width is the most difficult articulator to predict. The AFN and LDA features produce higher correlations than the single acoustic feature groups, supporting our hypothesis that different audio processing techniques encode for different visual information and that, therefore, a combined feature vector gives better results than individual feature groups.

4.4. Performance for replay attack scenarios

For each subject in the VidTIMIT corpus, the synthetic speaking
face was reconstructed from four articulatory parameters using the Xface facial animation engine again. The synthesized face sequences and real face sequences of each subject were presented to a GMM based audiovisual speaker authentication system designed to detect impostor replay attacks. The system uses several different types of feature extraction and fusion techniques to distinguish fake clients (synthesized speaking faces) from genuine clients with details described elsewhere [13].

For the male only text dependent subset of the VidTIMIT corpus (DB1TDMO), the detection error tradeoff (DET) curves for the three feature sets are shown in Figure 5.

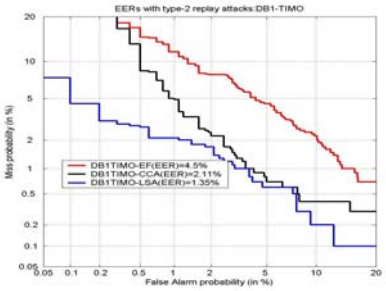

Figure 5: DET curves for face-voice verification of VidTIMIT male subjects against synthesized impostors using EF, AFN and LDA feature sets.

These curves show that the synthetic speaking faces were most confused with the original speaking faces when only the EF features were used (EER=3.02%), they were less confused with the original speaking faces when the LDA reduced full feature set was used (EER=1.92%) and they were least confused with the original speaking faces when the full feature set was used (EER=1.01%). EER results for this subset and for the text-dependent female, text-independent male and text-independent female subsets of VidTIMIT are listed in Table 2. These results all show that the synthetic speaking faces were most confused with the original speaking faces when only the EF features were used, less confused with the original when the LDA reduced full feature set was used and least confused with the original when the full feature set was used.

Table 1: Correlation coefficient for lip articulatory parameters for different acoustic features (VS-validation set; TS-Test set)

		LPC	LSF	MFCC	PCBF	EF0	AFN	LDA
VS	V_1	0.84	**0.88**	0.86	0.85	0.81	0.87	0.87
VS	V_2	0.70	0.74	0.72	0.69	0.48	0.74	**0.75**
VS	V_3	0.81	**0.85**	0.84	0.82	0.74	0.84	0.81
VS	V_4	0.79	0.83	**0.84**	0.81	0.77	0.83	0.80
TS	V_1	0.80	0.83	0.82	0.82	0.83	**0.89**	0.85
TS	V_2	0.68	0.70	0.64	0.68	0.43	0.75	**0.77**
TS	V_3	0.77	0.81	0.73	0.69	0.74	**0.75**	**0.75**
TS	V_4	0.78	0.80	0.80	0.83	0.73	**0.85**	0.76

5. Conclusions

A speaking face synthesis approach using HMM based acoustic-to-visual articulator mapping is presented. Correlations of different types of acoustic features and visual articulators were analyzed. We evaluated the reconstructed synthetic speaking faces by means of a face-voice verification system and the results suggest that the discrimination ability of the synthetic speaking faces from the originals depend on a number of factors, such as the features used for identity verification. Further work will include further details in synthesizing speaking faces such as nonverbal gestures like acoustic head motion, facial emotional expressions and the investigation of the robustness of acoustic and visual features against complex replay attacks.

Table 2: Equal error rates for verification of synthetic speaking faces: text-dependent/ text-independent male/female subsets of VidTIMIT for different acoustic feature sets

Data Set	EF	AFN	LDA
TD-M	3.02	1.01	1.92
TD-M	3.46	1.78	2.76
TI-M	4.50	1.38	2.11
TI-F	4.60	2.47	3.65

6. References

[1] P. Kakumanu, R. GutierrezOsuna, A. Esposito, R. Bryll, A. Goshtasby and O. N. Garcia, "Speech Driven Facial Animation", Proc. Workshop on Perceptual/Perceptive User Interfaces (PUI), Orlando, FL, PUI 2001.

[2] E. Cosatto and H. P. Graf, "Sample-based synthesis of Photorealistic talking heads," in Computer Animation, pp. 103-110, Philadelphia, Pennsylvania, June 810,1998.

[3] J. Ostermann and E. Haratsch, "An animation definition interface – rapid design of MPEG4 compliant animated faces and bodies," in Proc Int Workshop on SNHC and 3D Imaging, Rhodes, Greece, September 59 1997.

[4] S. Bengio, "Multimodal Authentication Using Asynchronous HMMs," Proc. Intl. Conf. AVBPA, 2003.

[5] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", Speech Communication, 17, pp. 179-192, 1995.

[6] G. Chetty and M. Wagner, "Face-Voice Authentication Based on 3D Face Models", Lecture Notes in Computer Science, Volume 3851, pp. 559-568, 2006.

[7] G. Chetty and M. Wagner, "'Liveness' Verification in Audio-Video Authentication", Proc. Int Conf on Spoken Language Processing ICSLP04, pp 2509-2512, 2004.

[8] D. Cosker, D. Marshall, P.L. Rosin, Y. Hicks, "Speech Driven Facial Animation using a Hidden Markov Co-articulation Model", Int. Conf. Pattern Recognition, vol. 1, pp. 128-131, 2004.

[9] C. Sanderson and K.K. Paliwal, "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters 24, 24092419, 2003.

[10] A.Wichmann, "The attitudinal effects of prosody and how they relate to emotion", ISCA Workshop on Speech and Emotion, Belfast 2000.

[11] G. Blais, M.D. Levine, Registering Multi-view Range Data to Create 3D Computer Objects, PAMI(17), No. 8, pp. 820-824, 1995.

[12] Xface, open source facial animation engine, http://xface.itc.it/.

[13] G. Chetty and M. Wagner, "Liveness detection using cross-modal correlations in face-voice person authentication", Proc. Interspeech 2005, 2181-2184, 2005.

[14] K. Choi, Y. Luo, J.N. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG4 facial animation system", J VLSI Proc, 29, 51-61, 2001.