Audio-Visual Correlation Modeling for Speaker Identification and Synthesis

by

Mehmet Emre Sargın

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical & Computer Engineering

Koç University

August, 2006

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Mehmet Emre Sargın

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

____
Prof. A. Murat Tekalp

____
Assist. Prof. Engin Erzin

____
Assist. Prof. Yücel Yemez

____
Assist. Prof. Alper T. Erdoğan

____
Dr. A. Tanju Erdem

Date: ________________________

To my family

ABSTRACT

This thesis addresses two major problems of multimodal signal processing using audio-visual correlation modeling: speaker recognition and speaker synthesis. We address the first problem, i.e., the audiovisual speaker recognition problem within an open-set identification framework, where audio (speech) and lip texture (intensity) modalities are fused employing a combination of early and late integration techniques. We first perform a canonical correlation analysis (CCA) on the audio and lip modalities so as to extract the correlated part of the information, and then employ an optimal combination of early and late integration techniques to fuse the extracted features. The results of the experiments indicate that the proposed multimodal fusion scheme improves the identification performance over the early and late integration of original modalities. We also demonstrate the importance of modality synchronization for the performance of early integration techniques and propose a CCA-based method to synchronize audio and lip modalities. We address the second problem, i.e., the speaker synthesis problem within the context of a speech-driven speaker animation application. More specifically, we present a Hidden Markov Model (HMM) based two-stage method for joint analysis of head gesture and speech prosody patterns of a speaker towards automatic realistic synthesis of head gestures from speech prosody. The analysis method is used to learn correlations between head gestures and prosody for a particular speaker from a training video sequence. The resulting audio-visual mapping model is then employed to synthesize natural head gestures on a given 3D head model for the speaker from arbitrary input test speech. Objective and subjective evaluations indicate that the proposed synthesis by analysis scheme provides natural looking head gestures for the speaker with any input test speech.

ÖZETÇE

Bu tezde görsel işitsel ilinti modellenmesi ile çok kipli sinyal işlemedeki iki önemli probleme cevap aranmaktadır: Konuşmacı tanıma ve konuşmacı sentezleme. Görsel-işitsel konuşmacı tanıma problemine, ses (konuşma) ve dudak doku (yeğinlik) özelliklerinin, erken ve geç tümleştirme tekniklerinin katışımı ile kaynaştırılarak, açık küme tanıma çatısı altında cevap aranmıştır. Ses ve dudak kiplerine doğal ilinti analizi (DİA) uygulayarak bilginin ilintili kısmını özütledikten sonra, erken ve geç tümleştirme tekniklerinin en iyi katışımı kullanılarak özütlenmiş öznitelikler kaynaştırılmıştır. Deneylerin sonuçları, önerilen çok kipli katıştırma tasarısının, özgün erken ve geç tümleştirme yöntemlerine göre tanıma performansını arttırdığını göstermiştir. Bununla beraber, kiplerin eşzamanlaşmasının erken tümleştirme tekniği için önemini gösterip, ses ve dudak kiplerinin eşzamanlaşması için DİA tabanlı metot önerilmiştir. Konuşmacı sentezleme problemine ise, konuşma ile sürülen konuşmacı animasyonu uygulaması bağlamında cevap aranmıştır. Konuşma tonlamasından otomatik gerçekçi kafa jesti sentezi doğrultusunda, Saklı Markov Model (SMM) tabanlı kafa jesti ve konuşma tonlaması örüntülerinin iki basamaklı metot ile bütünleşik analizi önerilmektedir. Analiz metodu, belirli bir konuşmacının eğitim video dizisi üzerinden, kafa jestleri ile konuşma tonlaması arasındaki ilintinin öğrenilmesi için kullanılmıştır. Ortaya çıkarılan ses-görüntü eşleşme modeli daha sonra gelişigüzel giriş test konuşmasından doğal kafa jestlerinin, konuşmacının 3B kafa modeli üzerinde sentezlenmesi için kullanılmıştır. Öznel ve nesnel değerlendirmeler, önerilen analiz ile sentezleme tasarısının herhangi giriş test konuşması ile gerçekçi görünen kafa jesti sentezlediğini belirlemiştir.

ACKNOWLEDGMENTS

First I would like to thank my supervisor Prof. A. Murat Tekalp and my co-advisors Assist. Prof. Engin Erzin and Assist. Prof. Yücel Yemez who have been a great source of inspiration and provided the right balance of suggestions, criticism, and freedom.

I am grateful to members of my thesis committee for critical reading of this thesis and for their valuable comments.

I would like to thank to Assist. Prof. Alper T. Erdoğan for his valuable help and comments in implementation of the head motion capture module. I also would like to thank Dr. Mehmet Özkan who has shared time, given acquisitions for our MVGL MASAL audio-visual database. I am again grateful to Dr. A. Tanju Erdem and Momentum-DMT Inc. who helped me to build the animation platform and integration of speech synchronized lip animation with the animation platform.

Finally I thank my family and my friends for providing me a morale support that helps me in hard days of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

LP	Linear Prediction
CCA	Canonical Correlation Analysis
DCT	Discrete Cosine Transform
EER	Equal Error Rate
FAR	False Acceptance Rate
FRR	False Reject Rate
GMM	Gaussian Mixture Models
IOHMM	Input-Output Hidden Markov Model
HMM	Hidden Markov Model
LS	Least-Squares
MFCC	Mel-Frequency Cepstral Coefficient

Chapter 1

INTRODUCTION

Multi-modal signal processing has recently been an active research area where combination of information from different sources are considered for various applications of human computer interface and biometrics. Audio and video signals are two major sources from which various modalities such as speech, prosody, voice, gestures, facial expressions and lip movements can be extracted [1],[2],[3]. Each of these modalities plays an important role in human-to-human communication and often carries some mutual information with other modalities. Hence the need for correlation analysis/modeling in multimodal signal processing applications.

This thesis addresses two major problems of multimodal signal processing using audio-visual correlation modeling: speaker recognition and speaker synthesis. Multi-modal fusion for the recognition task is often used to achieve more robust performance under adverse environmental conditions [4, 5, 6]. A typical task is audio-visual speaker identification, where the visual modality is introduced in addition to audio in order to make the system more robust in a noisy environment. With audiovisual fusion, even in case the audio modality is noisy and thus unreliable, the system performance is less effected due to the information coming from visual modality. Many methods to combine the multiple modalities for recognition task have been proposed in the literature but most of them do not take into account the correlation between modalities. For instance, in the audio-visual speaker identification scenario, when lip movement and speech modalities are employed, fusion strategies which assume that the modalities are independent of each other may not yield the optimum performance since these modalities are in fact physiologically coupled and highly correlated. Another related issue is that, even though the modalities are coupled, the features may contain some independent information, depending on the noise and selected features. In this case the features corresponding to each modality can be decomposed into correlated and uncorrelated components, which enables the design of an optimum fusion strategy [7]. Optimal fusion also requires synchronized features to better model the correlation between modalities. The asynchrony between modalities may be caused due to several reasons such as imperfections in the experimental acquisition setup or differences in co-articulation behavior.

Correlation between two modalities can also be exploited to predict one modality from the other. An interesting application is speech-driven speaker animation [8]. State of the art visual speaker animation methods are capable of generating synchronized lip movements automatically from speech content; however, they lack automatic synthesis of speaker gestures from speech. Since lip movements are strongly coupled with speech, it is relatively an easy task to find a mapping between the phonemes of speech and the visemes of lip movement. On the other hand, head and face gestures are usually added manually by artists, which is costly and often look unrealistic. There are several difficulties involved in modeling the correlation between gestures and speech. First there does not exist a well-established set of elementary gestural patterns for gesture synthesis. Second gestural patterns are speaker dependent, and may exhibit variations in time even for a given speaker and finally synchronicity of gesture and speech patterns exhibit variations. Hence, learning the correlation between gesture and speech patterns of a speaker towards automatic realistic synthesis of speaker gestures from speech remains as a challenging open research problem.

The organization of this thesis is as follows. The audiovisual speaker identification problem is addressed in Chapter 2. In Chapter 3, we present our preliminary work for speech driven speaker animation problem, which is based on co-occurrence analysis of gesture-speech patterns for synthesis of head and arm gestures driven by speech. In Chapter 4, we elaborate the idea presented in Chapter 3 and propose a more rigorous correlation modeling framework for speech-driven speaker animation, that specifically addresses the prosody-driven head gesture synthesis problem. A summary of the relevant past research related with each problem is provided in the introduction section of the associated chapter. Experimental results are presented at the end of each chapter. Finally the conclusions are given in Chapter 5.

1.1 Overview and Contributions

The major contribution of this thesis work is on audiovisual correlation modeling for speaker identification and synthesis, as described in more detail in the following.

In Chapter 2, we propose using canonical correlation analysis (CCA) to improve the performance of multimodal recognition systems that involve modalities having a mixture of correlated and uncorrelated components. More specifically, we address the audio-visual speaker recognition problem within an open-set identification framework. Audio and lip modalities are represented by mel-frequency cepstral and intensity-based DCT coefficients, respectively. There are two important contributions: First, we propose a simple CCA-based technique for synchronization of audio and lip modalities to optimize the performance of the data fusion process. Second, we propose a multimodal fusion strategy based on canonical correlation analysis that first extracts the correlated components of audio and lip features, and then employs an optimal combination of early and late integration schemes to fuse the extracted features.

In Chapter 3, we present our preliminary work for speech-driven synthesis of head and arm gestures. The audiovisual correlation analysis is based on a pre-designated gesture and speech dictionary. The gesture dictionary is composed of head nods and tilts and directional gestures whereas the dictionary of speech elements is composed of directional keywords and speech prosody. The gesture-speech correlation analysis finds an audio-visual mapping based on co-occurrences of the gesture-speech elements. The obtained mapping is then used to synthesize head and arm gestures driven by speech.

Chapter 4 presents a more rigorous and advanced framework for speech-driven gesture synthesis motivated by the prior work presented in Chapter 3. The framework is based on a two-stage method for joint analysis of head gesture and speech prosody patterns of a speaker. The analysis method is used to "learn" correlations between head gestures and prosody for a particular speaker from a training video sequence. The resulting audio-visual mapping model is then employed to synthesize natural head gestures on a given head model for the speaker from arbitrary input test speech. We represent head gestures by Euler angles associated with head rotations, and speech prosody by temporal variations in the pitch frequency and speech intensity. In the first stage analysis, we perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of head gesture and speech prosody features independently to determine elementary head gesture and speech prosody units, respectively. In the second stage, joint analysis of correlations between these elementary head gesture and prosody units is performed using Multi-Stream HMMs to determine an audio-visual mapping model. In the synthesis stage, the audio-visual mapping model is used to predict a sequence of gesture elements from the prosody element sequence computed for the input test speech. The Euler angles associated with each gesture element are then applied to animate the speaker head model.

Chapter 2

AUDIO-VISUAL SYNCHRONIZATION AND FUSION USING CANONICAL CORRELATION ANALYSIS

Speech and lip texture/movement are physiologically coupled modalities; hence, they are highly correlated. However, depending on the features employed for representation, they may also contain some uncorrelated components. Two fusion strategies are commonly employed in the literature: The late integration strategy [9], which is also referred to as decision or opinion fusion, is optimal in case the contributing modalities are uncorrelated, and thus the resulting partial decisions are statistically independent. On the other hand, early integration, which is also referred to as data fusion, combines modalities at the data or feature level and may be effective if the modalities are highly correlated. However, dimensionality is an important problem and in practice, decision fusion can outperform data fusion even if the modalities are tightly coupled. Neither of these two alternatives actually offers an optimal solution alone, especially when the modalities contain a mixture of correlated and uncorrelated components.

Lip information has extensively been employed in the state-of-the-art audio-visual speech recognition applications [1], since it is natural to expect that speech content can be revealed through lip reading. Lip movement patterns also contain information about the identity of the speaker. Yet, audio and lip information have been used for speaker identification in only few works such as [4, 5, 6]. These works are mainly based on decision fusion, where audio is generally modeled by mel frequency cepstral coefficients (MFCC). Several feature sets can be used for the lip modality such as shape, motion and texture. In texture-based approaches, pure or DCT-domain lip image intensities are commonly used as features [5, 2]. Dimension reduction techniques, such as principle component analysis (PCA), linear discriminant analysis (LDA) or Discrete Cosine Transform (DCT), are independently applied to the lip and speech modalities regardless of the mutual information between them.

There is relatively little work available on explicit analysis of audio-visual correlations.

In [10], the speaker association problem is addressed via an information theoretic method, which aims to maximize the mutual information between the projections of audiovisual measurements so as to detect the parts of video, that are highly correlated with the speech signal. In [11], the information fusion problem is addressed in the context of handwritten character recognition. The correlated projections of multiple features, which are assumed to be maximally informative, are first extracted by using CCA, and then concatenated. However their fusion scheme is not optimal, because the uncorrelated components, which may also be informative, are not taken into account; moreover, the combined features are all derived from a single modality. In [12], CCA is used for speaker adaptation to improve speech recognition performance.

Audiovisual correlation analysis has also been used in the literature to address the problem of temporal asynchrony between audio-visual features, such as in [13] that uses product HMMs (Hidden Markov Models) and as in [14] that uses CCA on audio and face video. In the case of lip movement and speech, asynchrony may occur not only due to imperfections of the acquisition setup, but also due to a natural delay between the acoustic and facial components of the speaking act.

In this chapter, we propose using CCA to improve the performance of multimodal recognition systems that involve modalities having a mixture of correlated and uncorrelated components. More specifically, we address the audio-visual speaker recognition problem within an open-set identification framework. Audio and lip modalities are represented by mel-frequency cepstral and intensity-based DCT coefficients, respectively. There are two important contributions: First, we propose a simple CCA-based technique for synchronization of audio and lip modalities to optimize the performance of the data fusion process.[1] Second, we propose a multimodal fusion strategy based on canonical correlation analysis that first extracts the correlated components of audio and lip features, and then employs an optimal combination of early and late integration schemes to fuse the extracted features. The chapter is organized as follows: In Section 2.1, we review basics of the open-set audio-visual speaker identification problem. We address the audiovisual synchronization problem in Section 2.2, and propose a CCA-based synchronization method. The proposed multimodal fusion scheme with canonical correlation analysis is presented in Section 2.3.

[1] A preliminary version of this method was presented in [15].

Experimental results are discussed in Section 2.4 and concluding remarks are given in Section 2.5. Finally in Appendix, we provide a brief review of the canonical correlation analysis problem, where we also clarify the terminology and notation used throughout the chapter.

2.1 The Audio-Visual Speaker Identification Problem

This section provides an overview of the open-set audio-visual speaker identification problem, since we present the proposed fusion strategy in the context of this application.

In open-set speaker identification, the objective is to find whether the given input audio and video features belong to one of the R subjects registered in the database or not; the system identifies the speaker if there is a match, rejects otherwise. Hence, the problem can be formulated as an $R + 1$ class identification problem, where there are R subjects and a reject class. For the open-set identification problem, we employ a maximum likelihood solution through the likelihood ratio test as described in [5]. The likelihood ratio is defined as

$$\rho(\lambda_r) = \log \frac{P(\boldsymbol{f}|\lambda_r)}{P(\boldsymbol{f}|\lambda_{R+1})} \tag{2.1}$$

where $\boldsymbol{f}$ is the observation from an unknown speaker, λ_r is the r-th registered speaker class, and λ_{R+1} is the impostor (reject) class. The conditional probability for the reject class, $P(\boldsymbol{f}|\lambda_{R+1})$, is approximated by using all available training data across all subjects. Then, the decision strategy can be implemented in two steps. First, determine

$$\lambda_* = \underset{\lambda_1,\ldots,\lambda_R}{\arg\max}\ \rho(\lambda_r), \tag{2.2}$$

and then

$$\text{if} \quad \rho(\lambda_*) \underset{\text{reject}}{\overset{\text{accept}}{\gtrless}} \tau \tag{2.3}$$

where λ_* denotes the speaker class with the maximum likelihood ratio and τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate.

2.1.1 Computation of Class Conditional Probabilities

Computation of class-conditional probabilities needs a prior modeling step. Hidden Markov Models (HMM) are known to be effective structures to model the temporal behavior of the speech signal and hence they are widely used in audio-based speaker identification and speech recognition applications. In this study, we address a text-dependent open-set speaker identification application and our database consists of audio and video signals belonging to individuals of a certain population. We use word-level continuous-density HMM structures for temporal characterization of both lip-texture and audio modalities. Each speaker in the database is modeled using a separate HMM, which is trained over some repetitions of the feature streams observed from the corresponding speaker and modality. An HMM model for the impostor class, λ_{R+1}, is also trained over the whole training data of the population. In the recognition process, given a test feature stream, each HMM structure associated with a speaker produces a likelihood ratio. The likelihood ratio test as defined in Eq. (2.3) identifies the person if there is a match and rejects otherwise.

The performance of speaker verification/identification systems is often measured using the equal error rate (EER). The EER is calculated as the operating point, where the false accept rate (FAR) equals the false reject rate (FRR). In the open-set identification problem, the false accept and false reject rates can be defined as,

$$\mathrm{FAR} = 100 \times \frac{F_a}{N_a + N_r} \quad \text{and} \quad \mathrm{FRR} = 100 \times \frac{F_r}{N_a}, \tag{2.4}$$

where F_a and F_r are the number of false accepts and rejects, and N_a and N_r are the total number of trials for the true and impostor clients in the testing, respectively.

2.1.2 Audio-Visual Feature Extraction

We use the mel-frequency cepstral coefficients (MFCC) as features for the audio modality, which are known to be robust and effective features and thus commonly employed in speaker recognition systems. The audio feature vector is formed as a collection of 13 MFCC coefficients together with the first and second derivatives, for a total of 39 coefficients. We denote the audio feature vector by $\boldsymbol{f}_A$ and its dimension by N_A.

We assume that each video sequence contains frontal view of a talking face. For the visual features, a preprocessing step is employed to locate the lip region in each frame, and to eliminate global motion of the head between the frames so that the extracted motion features within the lip region provides us with only the motion of the speaking act. To this effect, each face frame is aligned with the first frame of the sequence using a 2D parametric motion estimator. For every two consecutive face images, global head motion parameters are calculated using hierarchical Gaussian pyramids and 12-parameter quadratic motion model [16]. The face images are successively warped according to these calculated parameters [17] to align the lip regions. Thus, by hand-labeling the mid-point of the lip region only in the first frame, we automatically extract a sequence of lip frames of size 128 × 80 in all frames. The lip texture features, denoted by $\boldsymbol{f}_L$ of dimension N_L, are the first 50 zig-zag scanned 2D DCT coefficients of the luminance component within this rectangular lip region. These features implicitly represent lip movements with texture. The texture information itself might sometimes carry additional useful information for discrimination; but in some other cases it may also degrade the recognition performance since it is sensitive to acquisition conditions.

2.2 Audio-Visual Feature Synchronization using CCA

Early integration techniques require the features extracted from different modalities to be exactly at the same rate and in synchrony. In our case, the audio features are extracted at a rate of 100 audio frames per second (fps), whereas the lip features have only a frame rate of 15 video fps. Thus prior to early integration, the lip features are interpolated using cubic splines to match the audio frame rate. Let us denote the audio and lip features of the k-th 10ms frame by $\boldsymbol{f}_A^k$ and $\boldsymbol{f}_L^k$, respectively. The audio and visual features need to be precisely synchronized in the interpolated frame scale before the data fusion, so that the correlations between them can better be exploited. We propose using the canonical correlation analysis (CCA) to achieve synchronization (see Appendix A for a brief review of the canonical correlation analysis). The problem then becomes, given a set of realizations of $\boldsymbol{f}_A^{k+s}$ and $\boldsymbol{f}_L^k$, finding the delay s^* between audio and lip features, that maximizes the mutual information.

The CCA requires the covariance matrix of the concatenated audio and lip feature vector to be estimated using the whole set of realizations. The canonical correlations γ_i, $i = 1, 2, ..., N$, where N is the minimum of the audio and lip feature dimensions, which turns out to be 39 in our case, can then be computed from the estimated joint covariance matrix as described in Appendix A. Based upon these canonical correlations, we define an overall audio-visual correlation measure $\gamma_{AL}(s)$ between audio-visual features $\boldsymbol{f}_A^{k+s}$ and $\boldsymbol{f}_L^k$ as,

$$\gamma_{AL}(s) = \sum_{i=1}^{N} \gamma_i^2 \tag{2.5}$$

which is a function of the delay variable s. The CCA is applied to the audio-visual features with varying values of s, and for each s the value of the correlation measure $\gamma_{AL}(s)$ is computed.

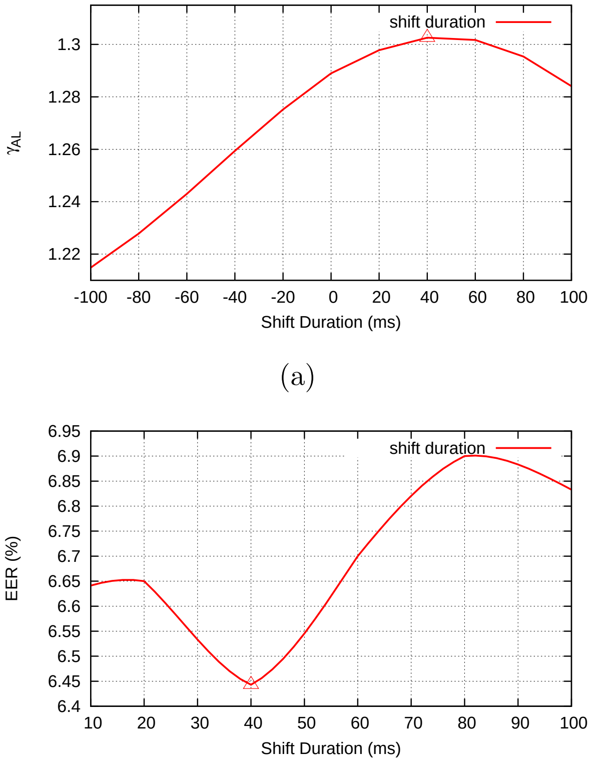
Figure 2.1(a) displays the behavior of $\gamma_{AL}(s)$ with varying s. As observed from the figure, the correlation measure, $\gamma_{AL}(s)$, is maximized for $s = 4$. This indicates that there is a 40 ms asynchrony between the features $\boldsymbol{f}_A$ and $\boldsymbol{f}_L$. Hence, for the rest of the chapter, the lip features are shifted by 4 frames prior to their fusion with the audio features. This inference is also supported with the speaker identification results that we obtained using early integration of audiovisual features. The equal error rates obtained for varying shift durations are plotted in Figure 2.1(b), where we observe that the optimal shift s^* found by our CCA-based synchronization method yields the best EER performance.

2.3 Multimodal Fusion Using CCA

In this section, we propose a combination of early and late integration of the synchronized audio and lip texture features. For the early integration, the audio and lip features are first transformed using the CCA. New strategies for early integration of the correlated CCA components are proposed in Section 2.3.1, whereas the best combination of early and late integration schemes for the overall multimodal fusion strategy is presented in Section 2.3.2.

2.3.1 Integration of Correlated CCA Components

Let the N-dimensional CCA-transformed audio and lip features be represented with $\boldsymbol{f}'_A$ and $\boldsymbol{f}'_L$, respectively, where N is chosen as the minimum of the audio feature dimension N_A and the lip feature dimension N_L. The between-set covariance matrix of $\boldsymbol{f}'_A$ and $\boldsymbol{f}'_L$ is a diagonal matrix with N diagonal terms, each of which corresponds to a squared canonical correlation (see Appendix A). However, each of these diagonal terms does not necessarily exhibit a strong correlation. Hence, one can pick the highly correlated components from the transformed vectors, discarding those with small canonical correlations. Fig. 2.2 plots the canonical correlations of the audio-visual features, obtained by applying CCA to our



(a)

(b)

Figure 2.1: CCA-based synchronization results: (a) Correlation measure γ_{AL} and (b) speaker identification equal error rates, for varying values of shift duration s.

database. As observed from Fig. 2.2, the maximum correlation coefficient is around 0.65, and 18 correlation coefficients out of 39 are higher than 0.05 threshold.

We define the highly correlated components as the projections of the original features onto the CCA basis vectors along which the canonical correlations are above a certain threshold T_h. Let us denote the two transformations corresponding to these canonical basis vectors by $\tilde{\mathbf{H}}_A$ and $\tilde{\mathbf{H}}_L$, respectively for the audio and lip modalities. Then, the correlated projections, $\tilde{\boldsymbol{f}}_A$ and $\tilde{\boldsymbol{f}}_L$, each with dimension M, are given by

$$\begin{aligned} \tilde{\boldsymbol{f}}_A &= \tilde{\mathbf{H}}_A^T \boldsymbol{f}_A \\ \tilde{\boldsymbol{f}}_L &= \tilde{\mathbf{H}}_L^T \boldsymbol{f}_L \end{aligned} \tag{2.6}$$

Here, $\tilde{\boldsymbol{f}}_A$ and $\tilde{\boldsymbol{f}}_L$ can be regarded as the correlated components embedded in $\boldsymbol{f}_A$ and $\boldsymbol{f}_L$.

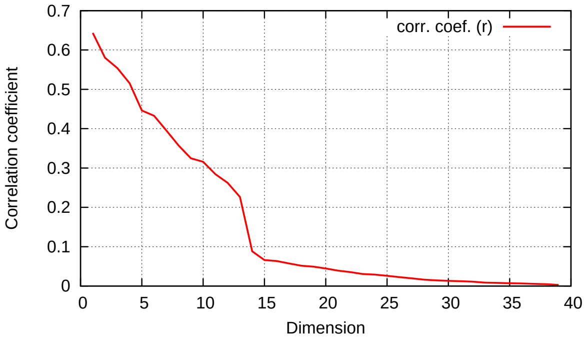


Figure 2.2: Canonical correlations resulting from audio-lip CCA analysis (sorted in decreasing order).

Early Integration by Concatenation

The early integration can simply be performed by concatenation of these correlated M dimensional projection vectors. The resulting combined audio-visual feature vector is thus given by

$$\tilde{\boldsymbol{f}}_{AL} = [\ \tilde{\boldsymbol{f}}_A^T\ \ \tilde{\boldsymbol{f}}_L^T\ ]^T_{2M\times 1} \tag{2.7}$$

Integration by Combining Weak Classifiers

An alternative integration strategy can be developed by decomposing the correlated CCA components, $\tilde{\boldsymbol{f}}_A$ and $\tilde{\boldsymbol{f}}_L$, into pairs of components, which are statistically independent from each other, but pairwise highly correlated. Recall from Appendix that the M pairs of canonical components, $(\tilde{f}_{Ai}, \tilde{f}_{Li})$, that are statistically independent from each other, can be computed via the projections

$$\begin{aligned} \tilde{f}_{Ai} &= \tilde{\boldsymbol{h}}_{Ai}^T \boldsymbol{f}_A \\ \tilde{f}_{Li} &= \tilde{\boldsymbol{h}}_{Li}^T \boldsymbol{f}_L \end{aligned} \tag{2.8}$$

where $\tilde{\boldsymbol{h}}_{Ai}$ and $\tilde{\boldsymbol{h}}_{Li}$ are the corresponding CCA basis vectors on which the projections are highly correlated so that $\gamma_i > T_h$.

In the new integration scheme, we employ M different *weak* classifiers, one for each pair of correlated speech-lip canonical components. Each canonical pair, that is, a two-dimensional concatenated vector, becomes input to the associated weak classifier. The decisions of these M weak classifiers are then combined using a late integration technique, as depicted in Fig. 2.3. The use of a weak classifier combination avoids the dimensionality problem of feature concatenation, and thus eases the task of feature modeling. Moreover, the late integration technique that combines the canonical pairs is optimal since these pairs of feature components are statistically independent.

2.3.2 The Proposed Multimodal Fusion Scheme

The two options presented in Section 2.3.1 for integration of the correlated CCA components do not take into account the mutually independent information embedded in the features that might also convey discriminative information.

The solution that we propose to exploit the mutually independent information is to employ a final step of late integration that incorporates the original audio and lip feature vectors, $\boldsymbol{f}_A$ and $\boldsymbol{f}_L$, as depicted in Figure 2.3. The experiments that we have conducted show that the uncorrelated components of the intensity-based lip feature vector can be noisy and do not carry useful additional discriminative information about a speaker's identity. Hence, our optimal configuration discards the original lip feature vector and incorporates only the audio features into the fusion scheme.

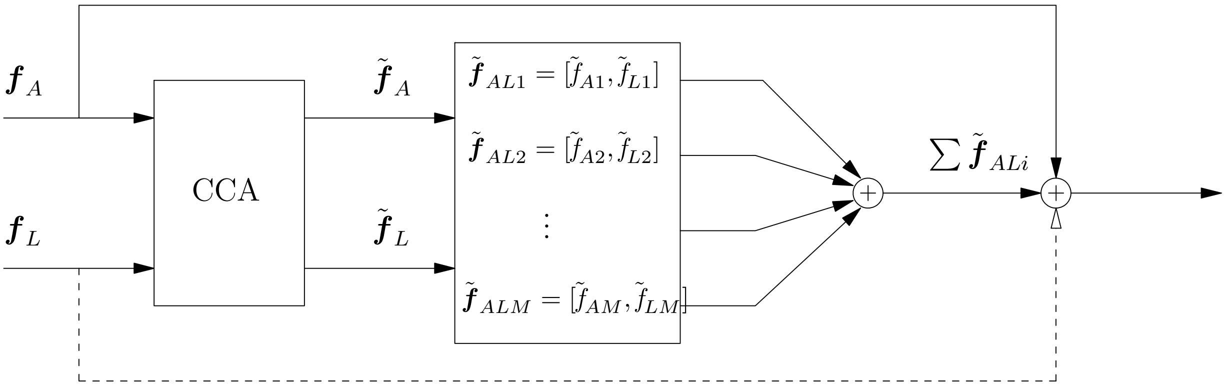

$\boldsymbol{f}_A$
$\boldsymbol{f}_L$
CCA
$\tilde{\boldsymbol{f}}_A$
$\tilde{\boldsymbol{f}}_L$
$\tilde{\boldsymbol{f}}_{AL1} = [\tilde{f}_{A1}, \tilde{f}_{L1}]$
$\tilde{\boldsymbol{f}}_{AL2} = [\tilde{f}_{A2}, \tilde{f}_{L2}]$
$\vdots$
$\tilde{\boldsymbol{f}}_{ALM} = [\tilde{f}_{AM}, \tilde{f}_{LM}]$
$\sum \tilde{\boldsymbol{f}}_{ALi}$
+
+

Figure 2.3: The proposed fusion scheme. The optimal configuration discards the late integration of the lip feature vector and incorporates only the audio features.

2.4 Experimental Results

The MVGL-AVD audio-visual database, denoted by $\mathcal{D}$, includes 50 subjects, where each subject utters ten repetitions of her/his name as the secret phrase [5]. A set of impostor data is also available with each subject in the population, uttering five different names from the population. The database is partitioned into two equal sets in two different ways, so that four different and independent training and testing sessions are deployed. Let $\mathcal{D}_T$ represent the whole database for the non-imposter data. In the experimental simulations, $\mathcal{D}_T$ is partitioned in two ways as $\{\mathcal{D}_{T_A}, \mathcal{D}_{\bar{T}_A}\}$ and $\{\mathcal{D}_{T_B}, \mathcal{D}_{\bar{T}_B}\}$, where $\mathcal{D}_{T_A}$ and $\mathcal{D}_{T_B}$ are disjoint sets, each having five repetitions from each subject in the database. Training and testing are performed over four independent sessions, where $\{\mathcal{D}_{T_A}, \mathcal{D}_{\bar{T}_A}\}$, $\{\mathcal{D}_{\bar{T}_A}, \mathcal{D}_{T_A}\}$, $\{\mathcal{D}_{T_B}, \mathcal{D}_{\bar{T}_B}\}$ and $\{\mathcal{D}_{\bar{T}_B}, \mathcal{D}_{T_B}\}$ pairs are respectively used for training and testing. Since there are 50 subjects and five repetitions for each true and impostor client tests, the resulting total number of trials for both the true accepts and true rejects is 1000. Note that the training sessions include training of speakers' HMM structures and CCA analysis of the audio-visual features.

The EER results for various fusion strategies using CCA are presented in Table 2.1 for several values of the correlation threshold T_h, where M denotes the number of correlated components above the threshold. In Table 2.1, $\tilde{\boldsymbol{f}}_{AL}$ and $\sum \tilde{\boldsymbol{f}}_{ALi}$ respectively denote integration by concatenation and integration by combining weak classifiers as described in Section 2.3.1, whereas + stands for Bayesian decision fusion (also called product rule) [5]. The minimum equal error rates in each row are indicated in bold. We observe that for early integration by concatenation, as the threshold T_h decreases, that is, as the transformed vector dimension M increases, the EER for the concatenated audio-lip feature, $\tilde{\boldsymbol{f}}_{AL}$, first decreases and then increases, achieving an optimal 3.8% EER value at the threshold $T_h = 0.25$. On the other hand, the EER obtained using a combination of weak classifiers, $\sum \tilde{\boldsymbol{f}}_{ALi}$, first decreases with decreasing threshold and then saturates at 3.8% EER. Hence, the EER performance in this case is more robust to selection of the threshold value.

In the next two rows of Table 2.1, the decision fusion results of the audio-only and the correlated audio-lip based classifiers are presented. When the audio-only classifier is combined with the concatenated audio-lip classifier, $(\boldsymbol{f}_A + \tilde{\boldsymbol{f}}_{AL})$, the best EER performance is observed as 0.6%. Furthermore, the EER drops to 0.3% for the proposed fusion structure in Fig. 2.3, that is, for Bayesian fusion of the audio-only classifier and the combined weak classifiers, $(\boldsymbol{f}_A + \sum \tilde{\boldsymbol{f}}_{ALi})$. Note that the performance saturates at this optimal EER value. Hence, the proposed fusion scheme is also robust to selection of the threshold T_h, or equivalently, to selection of the optimal correlated audio-visual feature dimension M.

The last two rows of Table 2.1 present the EER performances when the lip-only classifier is further included in the final decision fusion. The optimal EER performance degrades to 1.0% and 0.8% with $\boldsymbol{f}_A + \boldsymbol{f}_L + \tilde{\boldsymbol{f}}_{AL}$ and $\boldsymbol{f}_A + \boldsymbol{f}_L + \sum \tilde{\boldsymbol{f}}_{ALi}$ decision fusion schemes, respectively. The performance degradation is due to the inclusion of the uncorrelated lip information which is noisy, mainly because the lip texture alone is very sensitive to lighting conditions during acquisition.

Table 2.1: Speaker identification results for multimodal fusion using CCA: EER for varying values of the correlation threshold (T_h) and the corresponding projection dimension (M).

	EER (%) at (T_h, M)								
T_h	0.0	0.01	0.02	0.05	0.25	0.30	0.35	0.45	0.50
M	39	30	24	15	13	11	8	6	3
$\tilde{\boldsymbol{f}}_{AL}$	6.2	5.3	5.1	4.1	**3.8**	3.9	4.2	5.8	10.0
$\sum \tilde{\boldsymbol{f}}_{ALi}$	**3.8**	**3.8**	3.9	3.9	**3.8**	4.6	5.8	7.5	13.5
$\boldsymbol{f}_A + \tilde{\boldsymbol{f}}_{AL}$	0.9	0.8	**0.6**	**0.6**	**0.6**	**0.6**	0.8	1.0	2.7
$\boldsymbol{f}_A + \sum \tilde{\boldsymbol{f}}_{ALi}$	0.4	0.4	**0.3**	0.5	0.7	0.9	0.8	1.3	4.3
$\boldsymbol{f}_A + \boldsymbol{f}_L + \tilde{\boldsymbol{f}}_{AL}$	1.3	1.1	1.2	1.1	**1.0**	**1.0**	1.1	1.1	2.1
$\boldsymbol{f}_A + \boldsymbol{f}_L + \sum \tilde{\boldsymbol{f}}_{ALi}$	0.9	**0.8**	**0.8**	0.9	1.1	1.2	1.2	1.5	2.4

For benchmarking, we also present the EER results for unimodal and multimodal audio-visual speaker identification schemes in comparison with the best EER from Table 2.1. We observe that the conventional early fusion by means of concatenation of audio-visual features, ($\boldsymbol{f}_{AL}$), does not bring any performance gain, and performs worse than the audio-only identification. The late integration of audio-visual classifiers, including $\boldsymbol{f}_A + \boldsymbol{f}_L$ and $\boldsymbol{f}_A + \boldsymbol{f}_{AL}$, brings performance gain over audio-only identification. The last two rows present the best EER obtained with fusion schemes using CCA. We observe that the best EER is achieved by using the proposed fusion structure in Fig. 2.3.

Table 2.2: Comparison of EER for various audio-visual speaker identification strategies.

Strategy		EER (%)
Unimodal audio:	$\boldsymbol{f}_A$	1.1
Unimodal lip:	$\boldsymbol{f}_L$	7.0
Data fusion (concatenation):	$\boldsymbol{f}_{AL}$	6.4
Decision fusion:	$\boldsymbol{f}_A + \boldsymbol{f}_L$	0.7
Combined fusion (no CCA):	$\boldsymbol{f}_A + \boldsymbol{f}_{AL}$	0.8
Combined fusion using CCA:	$\boldsymbol{f}_A + \tilde{\boldsymbol{f}}_{AL}$	0.6
Combined optimal fusion:	$\boldsymbol{f}_A + \sum \tilde{\boldsymbol{f}}_{ALi}$	0.3

2.5 Discussions

We have presented new methods for multimodal synchronization and fusion using canonical correlation analysis. Experimental results show that the precise synchronization of modalities prior to fusion improves the speaker identification performance, and the proposed fusion strategy, following the proposed synchronization method, yields the best EER performance for the open-set audio-visual speaker identification. More specifically, we observe that i) in the late integration of weak classifiers, as the number of CCA transformed correlated audio-visual feature pairs increases, the equal error rate robustly drops to a minimum level and stays there, and ii) the best multimodal fusion strategy is constructed when the combination of weak classifiers is further integrated with the audio-only classifier.

Although the proposed fusion strategy using the CCA has only been demonstrated for the fusion of speech and lip modalities in the context of open-set speaker identification, it can indeed be applied to the fusion of any pair of modalities, which can be modeled as a mixture of correlated and uncorrelated components.

Chapter 3

COMBINED GESTURE-SPEECH ANALYSIS AND SPEECH DRIVEN GESTURE SYNTHESIS

The role of vision in human speech perception and processing is multi-faceted. The complementary nature of the information provided by the combinations of visual speech gestures used in phoneme production (such as lip and tongue movements) has been well researched and shown to be instinctively combined by listeners with acoustic and phonological information to correctly identify what is being said. In fact, speech perception is highly dependent on the visual gestures like lip movements. McGurk showed in [18] that perception of a speech sound is affected by a non matching lip/mouth movement. His experiments showed that when subject utters /b/ but lip movements corresponding to /g/ is seen, /d/ is perceived.

Non facial modalities like arm and head gestures are not tightly coupled with speech as the case of facial modalities but they are linked to present the same semantic idea units. The correlation between these two modalities and the analysis methodology is of interest in a number of fields including psychology and linguistics [19, 20].

The origin of the correlation between gestural and acoustical modalities are based on two hypothesis named as excitatory and inhibitory. The excitatory hypothesis states that vocal and gestural events are co-activated by a parallel processing system. In this case, human thoughts are processed by the cerebellum, then the motor neurons associated with vocal and muscular activation are stimulated simultaneously. The latter one called inhibitory hypothesis, in which vocal and gestural events are using the resources of single processing system. In this case, events of each modality co-occur with the counter modalities pauses. Detailed information about these hypothesis can be found in [21].

When audio and visual modalities are highly correlated, one modality's events can be used to predict the complementary modality's events. The more the modalities are correlated, the more reliable will be the predictions. This boils down the problem of prediction to selection of the most correlated features or events which was also studied in our previous work [15]. Estimation of correlated gestural events given speech, can be used to provide natural gesture patterns for the task of artificial gesture synthesis. Artificial gesture synthesis given speech is used in edutainment applications, where humans expect interactive conversations that animated person's speech is aided and complemented by other sensory modalities, including expression, gaze, gesture, grasp, signing, emotion, and factors beyond the textual equivalent of speech [22].

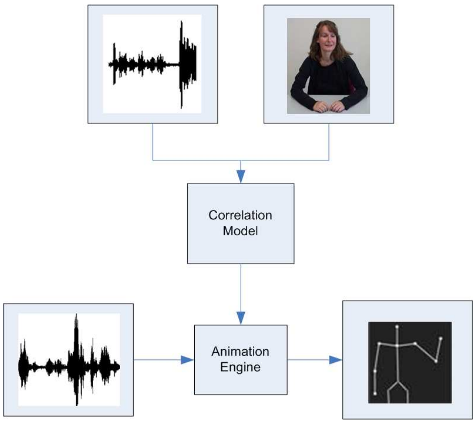


Figure 3.1: Proposed System Overview

3.1 Motivations and Initial Observations

A primary motivation of the work presented here was to identify natural classes of gestures that conveyed real linguistic meaning, that is, to identify gestures or groups of gestural patterns that could be clearly correlated with information conveyed in the speech signal. Once identified, these classes would be used to synthesize ”natural” gesture patterns using an animated stick figure, given an input speech signal. The work detailed below is intended to be a preliminary investigation and so is restricted to analyzing gestures in a limited but gesture-rich database. An audio-visual database was prepared, comprising 25 minutes of video data. A single native speaker of Canadian English was recorded, providing directions to a number of known destinations in response to assisted questions.

An initial informal analysis was carried out, in order to ascertain potential lexical candidates that had recurring patterns of significant gestures. This involved close viewing of the video data by two investigators with experience of gesture identification and speech annotation. Initial observation highlighted three candidates, ”left”, ”right”, and ”straight”, for further study. The three lexical items were chosen as they showed a high co-occurrence with periods of significant manual gestural activity. Furthermore, they had a high distribution throughout the database indicating a potentially rich source of data for analysis. It was informally noted that 28 instances of the candidate ”left”, appeared to be accompanied by some sort of gesture. Similarly 31 occurrences of ”right” had accompanying gestures, while ”straight” had associated gestures 32 times throughout the database. Other candidate words included ”across”, ”no”, and ”down”, but these were dismissed as having too few gesture-marked occurrences (8, 8, and 6 respectively).

3.2 Speech Event Detection

In this section we investigate automatic spotting of semantic and prosodic events. MFCC coefficients are used in the extraction of semantic events. Pitch, formant frequencies and intensity values are considered as features for prosodic event spotting.

3.2.1 Feature Extraction

Semantic features are represented with 13 MFCC, 13 delta coefficients and 13 acceleration coefficients. MFCC coefficients are calculated over 25 ms windows for each 10 ms frame, where the resulting speech feature rate is 100 fps.

The nature of prosodic speech events are well described with the temporal variations of intensity, pitch and formant frequencies. Therefore in this study, these three features, pitch (p), intensity (I) and the first three formant frequencies (f), are considered as the potential prosodic features.

The pitch contour is extracted from the speech signal using the autocorrelation method as described in [23]. The squared sound intensities are weighted with 32 ms Kaiser-20 window, and the speech signal intensity is calculated as the sum of these weighted samples. The 32 ms window is shifted by 10 ms for each frame such that, the intensity values have a 100 Hz frame rate. A linear prediction (LP) filter is calculated over 50 ms Hamming window for each 10 ms frame. The first three formant frequencies are extracted by tracking the peaks of the LP magnitude spectra.

3.2.2 Recognition of Semantic Events

Semantic events are considered as keywords uttered in speech. Frequently used words in the speech database are picked as the keywords, which are *left, right* and *straight.* In this section we present an HMM based automatic keyword spotter.

Keyword spotting task is performed using the methodology described in [22]. Manually labelled portion (80%) of the entire database is used for training and the remaining part is used for testing. Each keyword in the training database has at least 30 repetitions. Five HMM models are used for three keywords (*left, right* and *straight*) and two non-keywords (*silence* and *garbage*). The *silence* model is defined as segments corresponding to background noise. The *garbage* model corresponds to any non-keyword utterances. Continuous observation densities are modelled using varying number of Gaussian mixtures and the optimum number of Gaussian mixtures are selected considering the keyword spotting accuracy and false alarm rate.

In the experiments, we obtained 94.3% (33 out of 35) true detection and 1.6% (10 out of 600) false alarm rate for keyword spotting where the optimum number of Gaussian Mixtures is 15.

3.2.3 Recognition of Prosodic Events

Prosodic events that are correlated with speech signal are defined as pitch accents. Three different sets of features are used in proposed accent detector scheme: $[p, I]$, $[f, p, I]$ and $[\Delta f, p, I]$. Here, the "," operator represents concatenation of features.

In order to establish an initial working hypothesis, an experienced ToBI labeller marked training portion of speech for pitch accents and phrase boundaries. Within the training set, 122 pitch accents are identified. Manually labelled speech sequence is partitioned as *accent*, *non-pitch* and *non-accent.* The *accent* and *non-accent* labels correspond to syllables that are accented and non-accented, respectively. The *non-pitch* label is used for the syllables that pitch can not be extracted. Three left-to-right HMM structures with 6 states and 5 mixtures are used to model these three events.

Feature Set	$RRate$	$1 - FAR$
$[\Delta f, p, I]$	0.7810	0.6668
$[f, p, I]$	0.7140	0.6724
$[p, I]$	0.7479	0.6966

Table 3.1: Accent detection performance

The system is trained using the features corresponding to 80% portion of manually labelled pitch accents. Remaining 20% portion is used for testing. The position of testing portion is shifted 4 times with 4 new trainings to cover all labelled data in the testing. Table 3.1 presents the accent recognition rate $RRate$ and false alarm rate FAR. The use of $[\Delta f, p, I]$ feature set yields optimum performance. The $1 - FAR$ is maximized with the $[p, I]$ features, however, considering the trade off between false alarms and the recognized accents, the $[\Delta f, p, I]$ feature set still yields better performance than the other two.

3.3 Gestural Event Detection

In this section, HMM-based hand and head gesture recognition system is presented. The usage of HMMs for gesture recognition is motivated by the similarities between gesture and speech. Yang *et. al.,* summarizes these similarities in [24]. HMMs have been applied to the speech recognition problem to partition every word into a finite number of speech elements called phonemes. Similarly, the usage of HMMs for gesture recognition allows us to take the advantage of partitioning each gesture into gesture units where hidden states are associated with them. Therefore, the number of states for each HMM associated with a specific gesture should be selected according to the number of tactemes corresponding to that gesture.

3.3.1 Feature Extraction

In this study, head gesture features are chosen as the 8 global quadratic head motion parameters calculated over the face region. The extraction of head gesture features are described in detail in [22].

A hand gesture is represented with a single numeric feature which is the center of mass position of each hand. The center of mass is tracked over video using a Kalman filter where the states correspond to position and velocity [22].

3.3.2 Hand Gesture Recognition

Based on the initial observation of directional words and gestures that were salient in the video, three hand gestures were selected. Right and Left Gestures: The right or left hand turns to make a 90° with the arm, pointing to the right for right gesture, or to the left for left gesture. Straight Gesture: The subject starts with her hands in parallel, palms facing each other, fingers directed up, and moves the hands away from the body by extending her elbows. The finishing position is with hands parallel, palms facing each other, fingers pointing away from the subject's body.

An isolated hand gesture recognition scheme using continuous density HMMs with 5 states is employed. The performances obtained on the test video for left, right and straight gestures are 83%, 71% and 70% respectively.

3.3.3 Head Gesture Recognition

Head gestures, when examined, seemed to be correlated with prominences in speech. Since the evidence for correlation between sharp head movements and prosodic events in speech has previously been presented in gesture literature [25], we have decided to narrow down our investigation of head gestures to nods and head tilts. During nod gesture, the head comes down with chin closer to the body and sharply comes back up. During tilt gesture, the head rotates right or left 45° from its natural vertical position.

Given a set of training examples, three left to right continuous density HMMs are trained to model head gestures related with *nod*, *tilt* and *non-gesture*. These HMMs are then used to spot these gestures in testing sequence. The Viterbi algorithm is applied to determine the most probable gesture labels.

By changing the number of states used in HMMs, different performance metrics are obtained. The optimal number of states for head gesture recognition is achieved when *RRate* and 1-FAR metrics are equal to each other. The optimal number of states for

HMMs is 4 where *RRate* metric is 80%.

3.4 Correlation Analysis

After labels have been provided for speech and gesture events, correlation analysis has been conducted in order to provide justification for the two hypotheses. Directional hand gestures are closely correlated with the identified lexical candidate tokens, such as "left", "right" and "straight". Sharp head movements, such as nods and tilts, are closely correlated with speech prominences marked as pitch accents. In this section we will describe the correlation analysis procedure and results for both of these two hypotheses.

3.4.1 Directional Hand Gestures

Within the training portion labelled for directional hand gestures and speech keywords, 23 gestures were manually identified. Of the 23 gestures, 18 were matches with the candidate words "left", "right", and "straight", meaning that there was some degree of temporal overlap between the gestures and corresponding keywords. Of the remaining 5 gestures, 3 were wrongly identified as being related and 2 were designated as "confused", meaning that the speaker has correctly used the gesture to indicate going left, right or straight, but the phase of the gesture has overlapped with another word, usually being used in a different context. For example, the phrase: "Take a left and go *straight* down that street" had two accompanying left hand gestures. The first overlapped with the keyword "left" and was deemed a match, the second with the keyword "straight" and was marked as "confused".

3.4.2 Head Gestures

The training portion labelled for prosody and sharp head movements was found to contain 122 pitch accents and 81 head gestures 66 nods and 15 tilts. Of the 122 pitch accents, 79 or 64.75% overlapped with a head gesture, either a nod or a tilt. It is worth noting that from the 43 pitch accents that did not overlap with a head gesture, 23 or 53.5% were phrase initial accents, which are known to be problematic in prosody labelling. Often phrase-initial stressed syllables are misidentified as pitch accents due to the fact that both pitch accents and phrase-initial syllables are accompanied by "tense" voice quality [26].

If we disregard the 23 phrase initial syllables that were labelled as accents, only 20 of the 100 pitch accents identified did not overlap with a sharp head movement, that is 80% of remaining accents co-occurred with a head gesture.

The 79 accents that overlapped with a nod or a tilt were also examined for temporal correlation with the relevant head gesture. Time-stamp labels of the accented syllable were compared to the start and end time-stamps of the overlapping gesture using the statistical test of Pearson's correlation and the correlation test produced a correlation coefficient, $r = 0.994$, which implies an almost perfect correlation.

3.5 Animation

Given a speech sequence, the keyword spotter and the accent detector are used to extract time-stamps of auditory events. These time-stamps and speech sequence are provided to animation engine to animate the virtual body. In this work, we realized two animation schemes:

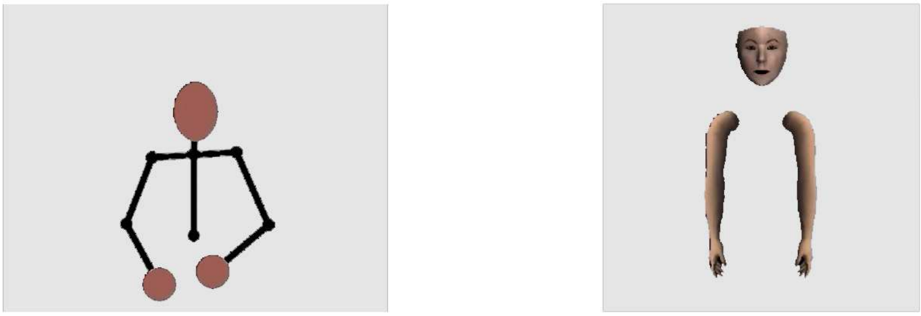
The Stick Model consists of line segments that corresponds to forearm and upper arm where starting and ending points of these line segments are determined as hand, shoulder and elbow positions. Together with these line segments, head is included with a line segment between head position and the center of the line segment between left and right shoulder. Animation engine for Stick Model uses 2D coordinates of the corresponding points.

The 3D Body Model consists of 2 arms and head without the body. Animation engine for this model uses a dictionary of gestural events and frames are constructed manually for each event in the dictionary. Animation engine uses each event independently for the animation of head, left arm and right arm. Sample stick and 3D body models are illustrated in Figure 3.2

In order to animate the body model, the center of mass positions of head and both hands is required by the animation engine. For each acoustical event, related gesture synthesized by considering the duration of acoustical event and the previously recognized gestures.

3.5.1 Hand Motion Model

During the left gesture, the motion of the right hand is limited when compared to the motion of the left hand. Similarly during the the right gesture, the motion of the left hand



(a) Stick Model

(b) 3D Body Model

Figure 3.2: Body Models

is limited when compared to the motion of the right hand. However for the straight gesture, both hands have large trajectories. The hand models for each hand gesture are constructed by HMMs. For the left gesture, we train an HMM by using only the left hand trajectory; for the right gesture, we train an HMM by using only the right hand trajectory and for the straight gesture we train two HMMs: one for the left hand and one for the right hand.

To construct an observation sequence from the HMM models, we use the model parameters: state transaction probabilities, parameters of Gaussian distribution for each state and prior probabilities of the states. Using this information, we construct an observation sequence by just providing a sequence length. The methodology used for constructing the observation sequence, given a sequence length and model parameters can be found on [22].

By using this methodology, we produce hand trajectories for each gesture where, for the *left* gesture, only left hand moves; for the *right* gesture, only right hand moves; and for the *straight* gesture both hands move.

Using the 20% portion of the database, we first run the keyword spotting algorithm for finding the time-stamps for words *left*, *right* and *straight.* We then produce the related hand gestures which are animated during the same period with the keyword.

3.5.2 Head Motion Model

Head motion model is generated according to the duration of accents. Let the duration of the accent be t_a seconds. For $t_a/2$ seconds head center of mass is shifted in $+y$ direction with 25 pixels/second. For the remaining $t_a/2$ seconds head center of mass is shifted back to its resting positions. The practical aspect of this methodology is that, the accents with short period are visually eliminated and the accents with long period are visually amplified.

3.6 Discussions

In this chapter, a gesture synthesizer based an audio-visual correlation is presented. Audio-visual correlation analysis is conducted using acoustic and visual events. Acoustic events are divided into semantic and prosodic categories. Visual events are selected as hand and head gestures. The types of events are defined by investigating a portion of the database. The repetitive patterns for acoustic events are mainly keywords (*left*, *right* and *straight*) and accents. The repetitive patterns for head gestures are *nod* and *tilt*. *Left* movement of left hand, *right* movement of right hand and *down* movement of both hands are defined as hand gestures.

Investigating the co-occurring patterns, we concluded that keywords and corresponding hand movements are strongly correlated. Moreover, *nod* movement of head is found out to be highly correlated with accents. Motivated from this fact, using the test portion of the database, first, keywords and accents are detected. Then the virtual body is animated using corresponding visual event at those detected acoustic events. Animation of the virtual body using both stick and 3D model can be found on [27].

Chapter 4

JOINT ANALYSIS OF HEAD GESTURE AND PROSODY PATTERNS FOR PROSODY-DRIVEN HEAD-GESTURE ANIMATION

State of the art visual speaker animation methods are capable of generating synchronized lip movements automatically from speech content; however, they lack automatic synthesis of speaker gestures from speech. Head and face gestures are usually added manually by artists, which is costly and often look unrealistic. Hence, learning the correlation between gesture and speech patterns of a speaker towards automatic realistic synthesis of speaker gestures from speech remains as a challenging research problem.

There exists significant literature on speaker lip animation, that is, rendering lip movements synchronized with the speech signal [8]. Since lip movement is physiologically tightly coupled with acoustic speech, it is relatively an easy task to find a mapping between the phonemes of speech and the visemes of lip movement. Many schemes exist to find such audio-to-visual mappings among which the HMM (Hidden Markov Model)-based techniques are the most common as they yield smooth animations exploiting temporal dynamics of speech [28, 29, 30, 31, 32, 33, 34, 35]. Some of these works also incorporate synthesis of facial expressions along with the lip movements to make animated faces look more natural [29, 32, 34, 35]. The common strategy in these techniques is to train a joint HMM structure with extracted visual and audio feature vectors and then to use the trained HMM structure to generate speech-driven facial expressions and lip movements.

Despite exhibiting variations from person to person and in time, head and body gestures are also correlated with speech. For example, it has been observed that manual gestures are correlated with prosody [21, 36] and verbal content of the speech [37], whereas head gestures are mostly correlated with the prosody [38, 39, 36]. Although correlations between speech and head/body gestures have been investigated in several works, there are only a limited number of publications that present preliminary results for speech-driven head and body gesture synthesis. In [40], a deterministic speech-to gesture mapping is first found by

K-Nearest Neighbor based dynamic programming and then used to synthesize head motion from speech. In [3], we presented a preliminary demonstration of natural looking head and arm gesture synthesis from speech using a manually determined audio-visual mapping from speech to head and arm motions. The aim of this chapter is to present a framework for joint analysis of head gesture and speech prosody patterns towards automatic generation of the audio-visual mapping from speech prosody to head gestures. Although the same framework can also be applied to analysis of co-occurring arm gesture and speech patterns, this is beyond the scope of the current chapter. There are some open challenges involved in the joint analysis of head gestures and prosody towards prosody-driven head gesture synthesis: First, there does not exist a well-established set of elementary prosody and gesture patterns for gesture synthesis, unlike phonemes and visemes in speech articulation. Second, prosody and gesture patterns are speaker dependent, and may exhibit variations in time even for the same speaker. Third, synchronicity of gesture and prosody patterns may exhibit variations. For instance, a speaker can move her/his head before the corresponding prosodic utterance with a variable time lag. Moreover, gestural patterns may span time intervals of different length with respect to its prosodic counterpart. We address these challenges by first processing the head gesture and prosody features separately by a parallel HMM structure to learn and model the gestural and prosodic elements (elementary patterns), respectively, over training data for a particular speaker. We then employ a multi-stream parallel HMM structure to find the jointly recurring gesture-prosody patterns and the corresponding audio-to-visual mapping.

HMM-based segmentation techniques are commonly employed in modeling multi-stream correlations; for example, for speech-driven lip animation in [33, 34, 35] and for audio-visual event detection in [41]. We can classify HMM based modeling techniques as supervised and unsupervised. Speech and lip motion correlation modeling can be thought of as a supervised analysis/segmentation problem, since phonemes and visemes constitute well-established elementary units for these modalities. Hence, speech-driven lip animation task is often equivalent to find a mapping between the phonemes of speech and the visemes of lip movement. On the other hand, we shall consider the audio-visual gesture modeling/mapping as an unsupervised segmentation problem, where the recurrent joint events are not well defined and to be extracted from the joint feature streams.

The organization of this chapter is as follows: In Section 4.1, we first provide an overview of the proposed HMM-based analysis-synthesis framework, and then describe the computation of head gesture and speech prosody features. Robust and accurate tracking of the speaker head motions is an integral part of the overall system; hence, it is described in detail. Section 4.2 presents the proposed two-stage unsupervised analysis procedure to identify and model jointly recurring head gesture and prosody patterns. Section 4.3 explains HMM-based synthesis of head gesture parameters from input test speech. In Section 4.4, we describe the experiments conducted, and present objective and subjective evaluation of the prosody-driven head gesture synthesis results. Finally, Section 4.5 provides discussions.

4.1 Overview of the Proposed System and Feature Extraction

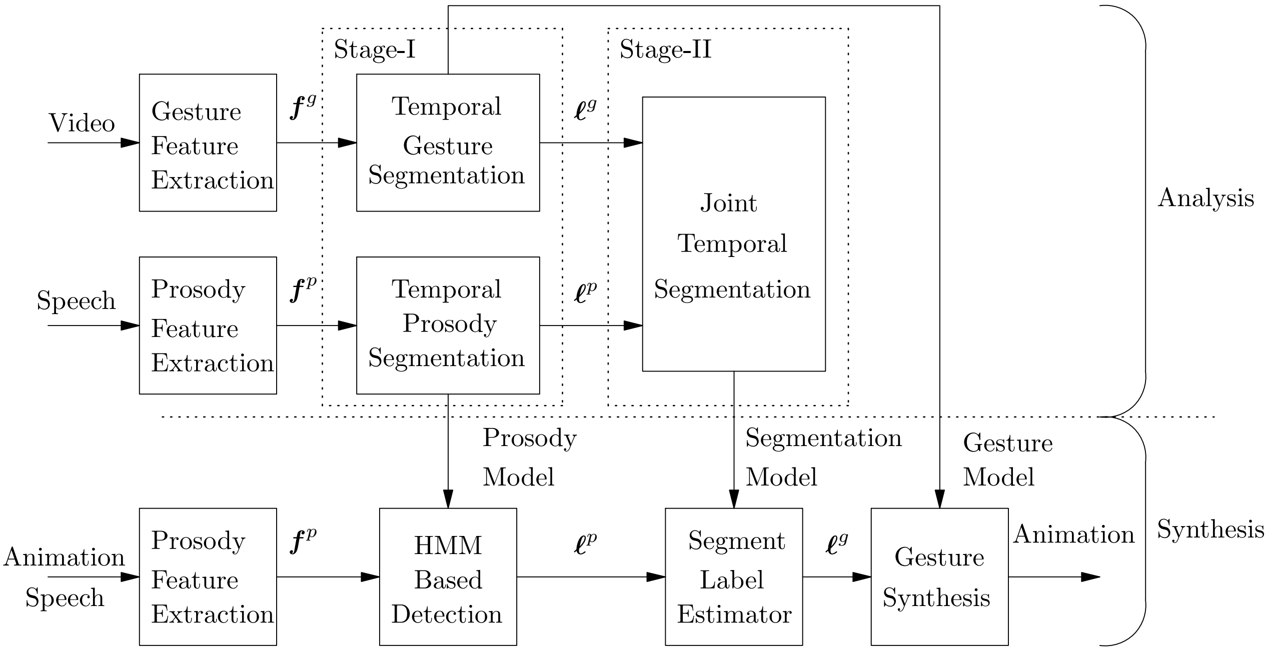


Figure 4.1: Overview of the proposed synthesis-by-analysis system.

A block diagram of the proposed system for prosody-driven head gesture animation, which consists of analysis and synthesis parts, is depicted in Fig. 4.1. The analysis part includes two feature extraction modules and two-stages of analysis. Feature extraction modules compute the head gesture features $\boldsymbol{f}^g$ and speech prosody features $\boldsymbol{f}^p$, respectively, from training stereo video sequences of a speaker. At the first stage analysis, individual feature streams are used to train separate parallel HMM structures, which provide probabilistic models for temporal recurrent patterns in the corresponding modalities, respectively. The segments corresponding to these patterns are detected and labeled over the training video streams, where pattern labels for prosody and gesture are denoted by $\boldsymbol{l}^p$ and $\boldsymbol{l}^g$, respectively. At the second stage, the labels of temporally segmented gesture and prosody streams are used together to train a discrete multi-stream parallel HMM to identify jointly recurring patterns. The resulting joint HMM structure models the correlation between speech prosody and head gestures. The synthesis part makes use of the joint HMM to predict the gesture labels from the prosody labels computed for a test input speech using the prosody HMM obtained by the first stage analysis. The corresponding gesture features, i.e., head motion parameters, are synthesized using the gesture HMM obtained at the first stage analysis and finally animated on a 3D head model. The details of two stages of analysis, shown by Stage-I and Stage-II blocks in Fig. 4.1 are presented in Section 4.2, whereas the gesture synthesis part is described in detail in Section 4.3. In the remainder of this section, we discuss extraction of head gesture and speech prosody features.

4.1.1 Extraction of Head Gesture Features

We define the head gesture feature vector, $\boldsymbol{f}^g_k$, for frame k to include the Euler angles associated with the 3D head rotation and their first differences,

$$\boldsymbol{f}^g_k = [\theta_k, \phi_k, \psi_k, \Delta\theta_k, \Delta\phi_k, \Delta\psi_k]^T \tag{4.1}$$

where θ_k, ϕ_k and ψ_k are the Euler angles of rotation, with respect to a reference frame k_r, around the x, y and z axes, respectively, and $\Delta\theta_k$, $\Delta\phi_k$, $\Delta\psi_k$ denote their respective first differences. The reference frame k_r can be selected as the first frame in which the subject's head is assumed to be at neutral position.

The set-up and algorithm for extraction of the feature vectors at each frame can be summarized as follows: We use a rectified stereo camera system with two identical cameras, and assume that the intrinsic camera parameters are known *a priori*. We first locate an ellipse for the head region in the reference frame. For each frame k, the 2D optical flow vectors are computed within the head region with respect to the reference frame k_r. Then, the 3D world coordinates of the 2D image points within the head region are calculated using disparity estimation and triangulation. Next, the rigid rotation and translation matrices are computed based on the resulting 3D point correspondences. Finally the Euler angles are extracted from the rotation matrix. We also employ a Kalman filter for post-smoothing of the estimated Euler angles. The steps of the head motion capture algorithm are detailed below.

2D Head Localization

The head region is initially detected in the reference frame k_r in one of the stereo views (e.g., the right or the left but not both) using a boosted Haar based cascade classifier structure which was initially proposed by Viola [42] and later improved by Lienhart [43]. The detected rectangular head region is then used to initialize the search window within which skin colored pixels corresponding to the facial region are found using color information. To this effect, we assume that the distribution of $\boldsymbol{c} = [c_1, c_2]$, which denote the color vector of a pixel (e.g., Cr and Cb components) belonging to skin region is Gaussian. The mean vector $\boldsymbol{\mu}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$ of this distribution are computed using a training set of sampled skin colors. The skin pixels are then detected based on the resulting probability density function. An ellipse $\mathcal{E}_{k_r}$ is then fitted to the skin region.

3D Point Tracking

Let P_{k_r} denote the set of image points within the ellipsoid $\mathcal{E}_{k_r}$ of the reference frame k_r so that $P_{k_r} = \{\boldsymbol{p}_{k_r,1}, \boldsymbol{p}_{k_r,2}, \dots, \boldsymbol{p}_{k_r,N}\}$ and $\boldsymbol{p}_{k_r,n} = [x_n, y_n]^T$. For each frame k, we employ the hierarchical Lukas-Kanade technique [44] to find the optical flow vectors, $\{\boldsymbol{v}_{k,1}, \boldsymbol{v}_{k,2}, \dots, \boldsymbol{v}_{k,N}\}$, from frame k_r to frame k. The set P_k of the corresponding image points in frame k is then obtained by $\boldsymbol{p}_{k,n} = \boldsymbol{p}_{k_r,n} + \boldsymbol{v}_{k,n}$, $n = 0, 1, ..., N$.

The 2D head localization procedure described above is also repeated for every frame of the sequence. We exclude those points in P_k that fall outside the ellipse $\mathcal{E}_k$ due to possible erroneous optical flow vectors. The excluded points are outliers which may corrupt the 3D motion capture process.

In order to find the 3D world coordinates of the remaining points, we need to estimate the disparity vector for each point at each frame. The disparity vectors are found using band-passed images and a cross correlation measure based on the sum of absolute differences [45]; and they are validated using several criteria [46]. Given the disparity vectors and the intrinsic parameters of the rectified stereo camera system, the 3D world coordinates of the 2D points from both sets P_{k_r} and P_k, excluding the outliers, are calculated by the well-known triangulation technique. Let $\boldsymbol{W}_k$ denote the $3 \times M$ matrix formed by the 3D world coordinates of the points associated with P_k, so that $\boldsymbol{W}_k = [\boldsymbol{w}_{k,1}, \boldsymbol{w}_{k,2}, \ldots, \boldsymbol{w}_{k,M}]$ and $\boldsymbol{w}_{k,m} = [X_m, Y_m, Z_m]^T$. We note that the dimension M of the matrices $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ are re-determined at each frame k according to the number of points that fall within the ellipse $\mathcal{E}_k$.

Computation of the Euler Angles

Let $\boldsymbol{R}_k$ and $\boldsymbol{t}_k$ denote the rotation matrix and the translation vector, respectively, of the rigid head motion from frame k_r to k. Then, $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ are related by

$$\boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{R}_k & \boldsymbol{t}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{k_r} \\ \\ \mathbf{1}^T \end{bmatrix}. \tag{4.2}$$

The rotation matrix $\boldsymbol{R}_k$ and translation vector $\boldsymbol{t}_k$ are estimated by a unitary constraint optimization technique as explained in the Appendix B. Once estimated, the rotation matrix $\boldsymbol{R}_k$ can be decomposed into three matrices:

$$\boldsymbol{R}_k = [r_{ij}^k] = \boldsymbol{R}_x(\theta_k)\boldsymbol{R}_y(\phi_k)\boldsymbol{R}_z(\psi_k) \tag{4.3}$$

where $\boldsymbol{R}_x(\theta_k)$, $\boldsymbol{R}_y(\phi_k)$ and $\boldsymbol{R}_z(\psi_k)$ are the matrices that specify rotations around x, y and z axes, respectively [47], [48]. The Euler angle vector $\boldsymbol{e}_k = [\theta_k, \phi_k, \psi_k]^T$ which denotes the mapping from $\boldsymbol{W}_{k_r}$ to $\boldsymbol{W}_k$, is finally extracted from this decomposition by

$$\boldsymbol{e}_k = \left[\arctan(-r_{23}^k/r_{33}^k),\ \arcsin(r_{13}^k),\ \arctan(-r_{12}^k/r_{11}^k)\right]^T. \tag{4.4}$$

In cases where the head rotation between the current frame k and reference frame k_r is larger than a threshold angle (e.g., if $|\theta_k| > 25°$ or $|\phi_k| > 25°$ or $|\psi_k| > 25°$), the optical flow vectors, hence the 3D point correspondences between two frames, may become unreliable. In such cases, we switch to incremental motion estimation, where the reference frame for frame k is set to frame $k-1$. Thus, we recompute optical flow vectors with respect to frame $k-1$; hence, the new 3D point correspondences and the resulting incremental Euler angle vector $\boldsymbol{\delta}_{k-1}$, which defines the rotation between frames k and $k-1$ are computed. Then, the Euler angle vector with respect to the reference frame k_r is given by

$$\boldsymbol{e}_k = \boldsymbol{e}_{k-1} + \boldsymbol{\delta}_{k-1} \tag{4.5}$$

Smoothing of the Feature Vector by Kalman Filtering

We finally employ a Kalman filter for post smoothing of the computed (estimated) Euler angles, which are input as observations $\boldsymbol{z}_k$ to the Kalman filter. The measurement noise $\boldsymbol{r}_k$ models the estimation errors in the Euler angles. The head gesture feature vector, $\boldsymbol{f}_k$ (the superscript g is omitted for ease of notation) consisting of the Euler angles and their first differences, is selected as the state vector. The state-space representation of the Kalman filter is given by

$$\begin{aligned} \boldsymbol{f}_{k+1} &= \boldsymbol{F}\boldsymbol{f}_k + \boldsymbol{G}\boldsymbol{u}_k \\ \boldsymbol{z}_k &= \boldsymbol{H}\boldsymbol{f}_k + \boldsymbol{r}_k \end{aligned} \tag{4.6}$$

where

$$\boldsymbol{f}_k = \begin{bmatrix} \boldsymbol{e}_k \\ \Delta\boldsymbol{e}_k \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} \boldsymbol{I}_{3\times3} & \boldsymbol{I}_{3\times3} \\ \boldsymbol{0}_{3\times3} & \boldsymbol{I}_{3\times3} \end{bmatrix}, \quad \boldsymbol{G} = \boldsymbol{I}_{6\times6}, \quad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{I}_{3\times3} \\ \boldsymbol{0}_{3\times3} \end{bmatrix} \tag{4.7}$$

The 3×1 vector $\Delta\boldsymbol{e}_k$ denotes the first differences of the Euler angles. The model noise $\boldsymbol{u}_k$ and measurement noise $\boldsymbol{r}_k$ are assumed to be uncorrelated, zero-mean white Gaussian processes. The output of the Kalman filter gives the final feature vector for the head gestures.

4.1.2 Extraction of Prosody Features

The prosodic speech events can be described by the temporal variations of loudness/intensity and pitch as well as pauses between phrases, phoneme durations, timing, and rhythm. Among these, the most expressive one is the pitch, which is the rate of vocal-fold cycling. In this study, pitch frequency, P, and speech intensity, I, are considered as prosody features.

The pitch contour is extracted at a rate of 100 Hz from the speech signal using the autocorrelation method as described in [23]. The mean of the pitch contour is removed over the active utterances to emphasize local variations [49] and later it is low pass filtered to reduce discontinuities. The regions between utterances without a valid pitch are filled with zero mean unit variance Gaussian noise. The intensity features are also extracted over the active utterances. The squared sound intensities are weighted with a 32 ms Kaiser-20 window, and the speech signal intensity is calculated as the sum of these weighted samples. The 32 ms window is shifted by 10 ms for each frame to extract intensity values at 100 Hz frame rate. The intensity features are also mean removed over active utterances and between-utterance regions are filled with zero mean unit variance Gaussian noise. The first order derivative, ΔP_k, of the post-processed pitch frequency at frame k is calculated using the following regression formula:

$$\Delta P_k = \frac{\sum_{i=1}^{2} i(P_{k+i} - P_{k-i})}{2\sum_{i=1}^{2} i^2}. \tag{4.8}$$

Finally, the pitch frequency, its derivative and the intensity are concatenated to form the 3 dimensional prosody feature vector $\boldsymbol{f}_k^p$ at frame k:

$$\boldsymbol{f}_k^p = [P_k \ \Delta P_k \ I_k]^T \tag{4.9}$$

4.2 Head Gesture-Prosody Pattern Analysis

In this section, we propose a two stage HMM-based unsupervised analysis framework, where the first stage aims to separately extract elementary gesture and prosody patterns for a speaker, and the second stage determines a correlation model between these head gesture and prosody patterns. In the first stage analysis, recurring elementary gesture and prosody patterns are determined separately by unsupervised temporal clustering of individual gesture and prosody feature streams, respectively. The extracted elementary prosody and gesture patterns are analogous to phonemes and visemes in the speech and lip motion modeling. However, the elementary gesture and prosody patterns are not well established as in the case of phonemes and visemes, since the nature and strength of head gesture and prosody patterns may vary from person to person and in time. Hence, the need for unsupervised stage I analysis in order to extract these patterns for each speaker. Furthermore, the joint recurring nature of these patterns are also not well established as in the case of phoneme-viseme association; hence, the need for stage II analysis for joint modeling of correlations between head gesture and prosody patterns. In order to find a mapping between prosody and gesture patterns, unsupervised temporal segmentation of joint gesture and prosody pattern labels is performed, which defines the correlation between gesture and prosody pattern streams and relates co-occurring head gesture and prosody patterns.

We note that if a multi-stream HMM structure were directly employed for joint analysis of gesture and prosody feature streams, as commonly used for event detection [41], instead of the proposed two-stage analysis, the resulting joint gesture-prosody feature segments would not necessarily correspond to *independent* meaningful elementary gesture and prosody patterns. As a result, the synthesized gesture sequence might contain poorly defined gestural elements, which would degrade the quality of prosody-driven head gesture animation.

4.2.1 Stage-I: Extraction of Elementary Head Gesture and Prosody Patterns

The first stage analysis defines recurrent elementary head gesture and prosody patterns separately using unsupervised temporal clustering over individual feature streams. The gesture and prosody feature streams $\boldsymbol{F}^g$ and $\boldsymbol{F}^p$ are separately used to train two HMM structures Λ_g and Λ_p, which capture recurrent head gesture segments $\boldsymbol{\varepsilon}^g$ and prosody segments $\boldsymbol{\varepsilon}^p$. For ease of notation, we use a generic notation to represent the HMM structure which is identical for the gesture and prosody streams. The HMM structure Λ, which is used for unsupervised temporal segmentation, has M parallel branches and N states as shown in Fig. 4.2. The states labeled as s_s and s_e are non emitting start and end states of the parallel HMM structure. Fig. 4.2 clearly illustrates that the parallel HMM Λ is composed of M parallel left-to-right HMMs, $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, where each λ_m is composed of N states, $\{s_{m,1}, s_{m,2}, \ldots, s_{m,N}\}$. The state transition matrix A_{λ_m} of each λ_m is associated with a sub-diagonal matrix of A_Λ. The feature stream is a sequence of feature vectors, $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$, where $\boldsymbol{f}_t$ denotes the feature vector at frame t. Unsupervised temporal segmentation using HMM model Λ yields L number of segments $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$. The l-th temporal segment is associated with the following sequence of feature vectors,

$$\varepsilon_l = \{\boldsymbol{f}_{t_l}, \boldsymbol{f}_{t_l+1}, \ldots, \boldsymbol{f}_{t_{l+1}-1}\} \quad l = 1, 2, \ldots, L \tag{4.10}$$

where $\boldsymbol{f}_{t_1}$ is the first feature vector $\boldsymbol{f}_1$ and $\boldsymbol{f}_{t_{L+1}-1}$ is the last feature vector $\boldsymbol{f}_T$.

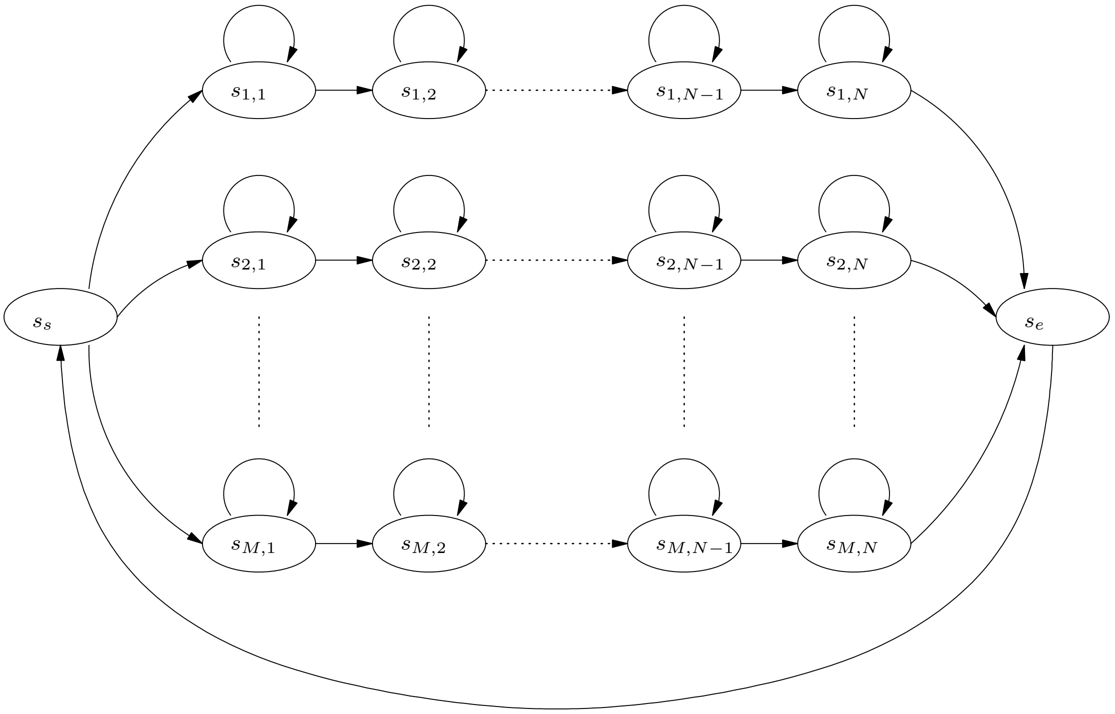


Figure 4.2: Parallel HMM structure

The segmentation of the feature stream is performed using Viterbi decoding to maximize the probability of model match, which is the probability of feature sequence $\boldsymbol{F}$ given the trained parallel HMM Λ,

$$\begin{aligned}\mathrm{P}(\boldsymbol{F}|\Lambda) &= \max_{t_l,m_l}\prod_{l=1}^{L}\mathrm{P}(\{\boldsymbol{f}_{t_l},\boldsymbol{f}_{t_l+1},\ldots,\boldsymbol{f}_{t_{l+1}-1}\}|\lambda_{m_l}) \\ &= \max_{\varepsilon_l,m_l}\prod_{l=1}^{L}\mathrm{P}(\varepsilon_l|\lambda_{m_l}) \qquad (4.11)\end{aligned}$$

where ε_l is the l-th temporal segment, which is modeled by the m_l-th branch of the parallel HMM Λ. One can show that λ_{m_l} is the best match for the feature sequence ε_l, that is,

$$m_l = \underset{m}{\operatorname{argmax}}\,\mathrm{P}(\varepsilon_l|\lambda_m) \qquad (4.12)$$

Since, the temporal segment ε_l from frame t_l to $(t_{l+1}-1)$ is associated with segment label m_l, we define the sequence of frame labels based on this association as,

$$\ell_t = m_l \qquad \text{for } t = t_l, t_l+1, \ldots, t_{l+1}-1 \qquad (4.13)$$

where ℓ_t is the label of the t-th frame and we have a label sequence $\boldsymbol{\ell} = \{\ell_1, \ell_2, \ldots, \ell_T\}$ corresponding to the feature sequence $\boldsymbol{F}$. The first stage analysis extracts the frame label sequences $\boldsymbol{\ell}^g$ and $\boldsymbol{\ell}^p$ given the head gesture and prosody feature streams $\boldsymbol{F}^g$ and $\boldsymbol{F}^p$. While mapping the gesture and prosody features to discrete frame labels, the mismatch between the frame rates of gesture and prosody is eliminated by downsampling the frame rate of prosody label stream to the rate of gesture label stream.

The parallel HMM structure has two important parameters to set before training the model Λ. The first parameter is the number of states in each branch, N. It should be selected by considering the minimum duration of temporal patterns. Selecting a small N may hamper modeling long term statistics for each branch of the parallel HMM. The extreme case $N = 1$ reduces to K-Means unsupervised clustering. We select the number of states in each branch of the head gesture HMM Λ_g as $N_{\Lambda_g} = 10$, corresponding to the minimum gesture pattern duration of 10 frames ($\frac{1}{3}$ sec assuming 30 video frames/sec). Note that, the gesture patterns can be longer than 10 frames since the HMM structure allows self-state transitions. On the other hand, the prosody patterns are expected to follow smooth pitch frequency movements over several syllables. Considering the average syllable durations and smoothness of the pitch contours, we set $N_{\Lambda_p} = 5$ in each branch of the prosody HMM model Λ_p.

The second parameter is the number of temporal patterns, M. Since the number of head gesture and prosody patterns is speaker dependent, we propose selection of M by using two fitness measures. The first fitness measure α, which is inversely related to in-class variance, is defined as the frame average of the log-probability of model match,

$$\alpha = \frac{1}{T} \log(\mathrm{P}(\boldsymbol{F}|\Lambda)) \tag{4.14}$$

The α measure is expected to saturate with increasing number of parallel branches in Λ, since the training database is expected to contain limited number of temporal patterns. However, small variations within temporal patterns are also expected, hence the number of branches M can be more than the actual number of temporal patterns in the training corpus. Consequently, the second fitness measure, which is the average statistical separation between two similar temporal patterns, increases with the decreasing number of temporal patterns. The second fitness measure β is considered as the average statistical separation between two similar temporal patterns, and it is defined as

$$\beta = \frac{1}{T}\sum_{l=1}^{L} \log(\frac{\mathrm{P}(\varepsilon_l|\lambda_{m_l})}{\mathrm{P}(\varepsilon_l|\lambda_{m_l^*})}), \tag{4.15}$$

where $\lambda_{m_l^*}$ is the second best match for the temporal segment ε_l, that is,

$$m_l^* = \underset{\forall m \neq m_l}{\operatorname{argmax}} \mathrm{P}(\varepsilon_l|\lambda_m) \tag{4.16}$$

While M is increasing, the HMM branch models λ_{m_l} and $\lambda_{m_l^*}$ are expected to be similar, which decreases the β measure. Therefore, the total number of temporal patterns, M, can be selected by jointly maximizing the α and β measures.

4.2.2 Stage-II: Joint Modeling of Prosody-Gesture Patterns

In the second stage, unsupervised segmentation of the joint gesture-prosody label stream is performed to detect recurrent joint label patterns. Note that, this task is similar to the task of stage I, except in the second stage we have a multi-stream discrete observation sequence. For this task, the parallel HMM structure in Fig. 4.2 is used with discrete multi-stream HMM branches. In multi-stream HMMs, all streams share the same state transition structure however emission probabilities are determined independently for each stream.

The joint gesture-prosody frame label stream, denoted by $\boldsymbol{\ell}^{gp}$, is defined such that for every frame k, $\ell_k^{gp} = [\ell_k^g,\ \ell_k^p]^T$. We represent the discrete multi-stream parallel HMM structure with Γ_{gp} and its m-th branch with γ_m^{gp}. The discrete HMM Γ_{gp} is trained over the joint gesture-prosody frame label stream. Each branch γ_m^{gp}, associated with a joint gesture-prosody temporal label pattern, is then described with a state transition matrix $\boldsymbol{A}_{\gamma_m^{gp}}$, a discrete observation probability distribution $\boldsymbol{B}_{\gamma_m^{gp}}$ and an initial state probability matrix $\boldsymbol{\Pi}_{\gamma_m^{gp}}$,

$$\gamma_m^{gp} = (\boldsymbol{A}_{\gamma_m^{gp}}, \boldsymbol{B}_{\gamma_m^{gp}}, \boldsymbol{\Pi}_{\gamma_m^{gp}}) \tag{4.17}$$

The discrete observation probability distribution $\boldsymbol{B}_{\gamma_m^{gp}}$ defines the probability of observing a gesture-prosody frame label at state s and frame k,

$$P(\ell_k^{gp}|s) = P(\ell_k^g|s)^{\kappa_g} P(\ell_k^p|s)^{\kappa_p} \tag{4.18}$$

where the exponents, κ_g and κ_p, are the stream weights and they are selected equal to each other as 1. Note that, the multi-stream discrete HMM γ_m^{gp} models can be split into individual single-stream discrete HMM models $\gamma_m^g = (\boldsymbol{A}_{\gamma_m^{gp}}, \boldsymbol{B}_{\gamma_m^g}, \boldsymbol{\Pi}_{\gamma_m^{gp}})$ and $\gamma_m^p = (\boldsymbol{A}_{\gamma_m^{gp}}, \boldsymbol{B}_{\gamma_m^p}, \boldsymbol{\Pi}_{\gamma_m^{gp}})$ respectively for gesture and prosody streams. The single stream HMM models share the same state transition and initial state probability matrices while they differ with the discrete observation probability distributions. The individual observation distributions are associated with $P(\ell_k^g|s)$ and $P(\ell_k^p|s)$ for gesture and prosody models, respectively. Unsupervised temporal segmentation of joint label streams is demonstrated by the following example.

Example: Let us have two label streams $\boldsymbol{\ell}^a$ and $\boldsymbol{\ell}^b$, where each label can assume values 1, 2, or 3. When temporal segmentation of the joint label stream is performed using the HMM structure Γ_{gp} with $M = 2$ patterns and $N = 3$ number of states for each pattern, we obtain the result shown in Fig. 4.3. One can observe that the recurrent joint label patterns are captured and the asynchrony between individual label streams is modelled by the first and the last states of the HMM branches.

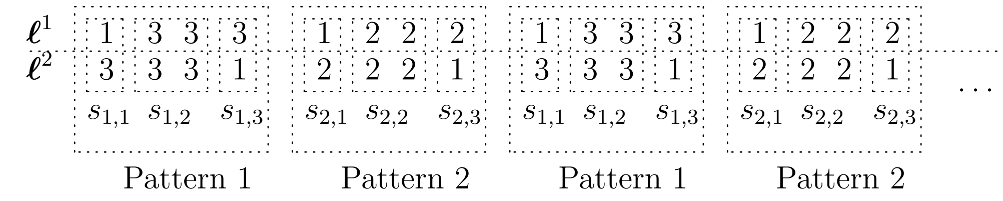
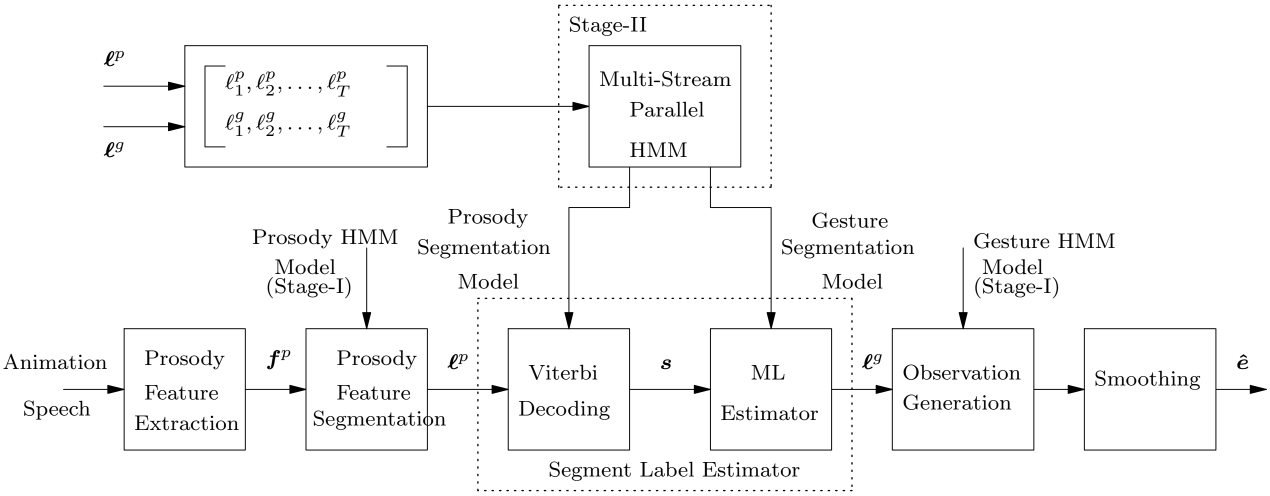


Figure 4.3: Example for unsupervised joint label segmentation

The number of states $N_{\Gamma_{gp}}$ for each branch of Γ_{gp} should be selected according to the number of head gesture and prosody patterns determined by the stage I analysis, since Γ_{gp} models the recurrent joint gesture-prosody label pairs. Similarly, the number of branches $M_{\Gamma_{gp}}$ in Γ_{gp} should be selected by considering the two fitness measures α and β as defined in (4.14) and (4.15). The selection of $N_{\Gamma_{gp}}$ and $M_{\Gamma_{gp}}$ is further discussed in Section**??**.

4.3 Prosody-Driven Gesture Synthesis

In this section, we address prosody-driven gesture synthesis using the proposed gesture-prosody pattern model. A detailed block diagram of the proposed prosody-driven gesture synthesis system is shown in Fig. 4.4. The system takes speech as input and produces a sequence of head gesture features, i.e., Euler angle vectors, which are naturally correlated with the input speech. The details of the sub-blocks are described in the following.



$\boldsymbol{\ell}^p$
$\boldsymbol{\ell}^g$
$\ell_1^p, \ell_2^p, \ldots, \ell_T^p$
$\ell_1^g, \ell_2^g, \ldots, \ell_T^g$
Stage-II
Multi-Stream Parallel HMM
Prosody HMM Model (Stage-I)
Prosody Segmentation Model
Gesture Segmentation Model
Gesture HMM Model (Stage-I)
Animation Speech
Prosody Feature Extraction
$\boldsymbol{f}^p$
Prosody Feature Segmentation
$\boldsymbol{\ell}^p$
Viterbi Decoding
$\boldsymbol{s}$
ML Estimator
$\boldsymbol{\ell}^g$
Observation Generation
Smoothing
$\hat{\boldsymbol{e}}$
Segment Label Estimator

Figure 4.4: The proposed prosody-driven gesture synthesis system.

Prosody Feature Extraction

The prosody features, $\boldsymbol{F}^p$, are extracted from the input speech signal as described in Section 4.1.2.

Prosody Feature Segmentation

Temporal segmentation of prosody feature sequence $\boldsymbol{F}^p$ is performed using the HMM model Λ_p, which is trained in the stage I analysis in Section 4.2.1. During the temporal segmentation, the conditional probability $P(\boldsymbol{F}^p|\Lambda_p)$ is maximized using Viterbi decoding to extract the temporal prosody segment sequence, $\boldsymbol{\varepsilon}^p$, and the sequence of prosody frame labels, $\boldsymbol{\ell}^p$.

Gesture Segment Label Estimation

The aim of this step is to predict the sequence of gesture frame labels, $\boldsymbol{\ell}^g$, given the prosody frame labels $\boldsymbol{\ell}^p$. To this effect, temporal segmentation of the prosody frame labels, $\boldsymbol{\ell}^p$ is performed using the HMM model Γ_p, which is extracted by splitting the jointly trained gesture-prosody HMM model Γ_{gp}. As a result of this temporal prosody label segmentation, a state sequence $\boldsymbol{s}^p = \{s_1^p, s_2^p, \ldots, s_K^p\}$ associated with $\boldsymbol{\ell}^p = \{\ell_1^p, \ell_2^p, \ldots, \ell_K^p\}$ is extracted. Then, the gesture frame label sequence $\boldsymbol{\ell}^g$ is predicted by maximizing the probability of observing gesture label on the state sequence path $\boldsymbol{s}^p$ over the gesture HMM model Γ_g, such that,

$$\ell_k^g = \arg\max_m P(m|s_k^p, \Gamma_g) \tag{4.19}$$

where k is the frame index, m runs over all possible M gesture patterns and the conditional probability $P(m|s_k^p, \Gamma^g)$ is defined by the discrete observation probability distribution $\boldsymbol{B}_{\gamma_m^g}$.

Generation of Euler Angles

This step computes the gesture segment sequence $\boldsymbol{\varepsilon}^g$, consisting of the Euler angle features, given the gesture frame label sequence $\boldsymbol{\ell}^g$. First, we find the segment frame boundaries, $\{t_l\}_{l=1}^L$, by merging the same gesture frame labels in the sequence $\boldsymbol{\ell}^g$. Then, the Euler angle features for the l-th segment, $\varepsilon_l^g = \{\boldsymbol{f}_{t_l}^g, \boldsymbol{f}_{t_l+1}^g, \ldots, \boldsymbol{f}_{t_{l+1}-1}^g\}$, are generated from the HMM $\lambda_{\ell_{t_l}}^g$, which is the ℓ_{t_l}-th branch of the parallel HMM model Λ_g (computed in stage I).

Note that, the segment duration for the l-th segment is extended as $d_l = (t_{l+1} + \Delta - (t_l - \Delta))$ frames, where Δ is the number of overlapping frames at the segment boundaries to smooth segment-to-segment transitions. The state sequence $\boldsymbol{s}_l^g$ or equivalently the state occupancy durations for the l-th segment is calculated using the diagonal terms of the d_l-step state transition matrix of the HMM $\lambda_{\ell_{t_l}}^g$. Having the state sequence $\boldsymbol{s}_l^g$ and the continuous observation probability $P(\boldsymbol{f}^g|\boldsymbol{s}_l^g)$, which are modeled using a Gaussian distribution, the Euler angle features are generated along the state sequence associated with the distribution $P(\boldsymbol{f}^g|\boldsymbol{s}_l^g)$. The segment boundaries have $2\Delta + 1$ number of frame overlaps, where the overlapped and averaged features generate smoother segment-to-segment transitions.

Smoothing of Euler Angles

As the final step of the gesture synthesis, the Euler angles are smoothed using median filtering followed by a Gaussian low pass filter to remove motion jerkiness. The median filtering is performed over 11 visual frames and the Gaussian smoothing is performed over 15 visual frames. Fig. 4.5 depicts the samples generated from the HMM, and outputs of the median and Gaussian filters. The figure clearly shows that the median filter removes jitters within a state and the Gaussian low pass filter smooths the state-to-state transitions.

There are two main advantages of using HMMs for gesture synthesis. The first is the random variations in the synthesized gesture patterns for each segment. This variation yields

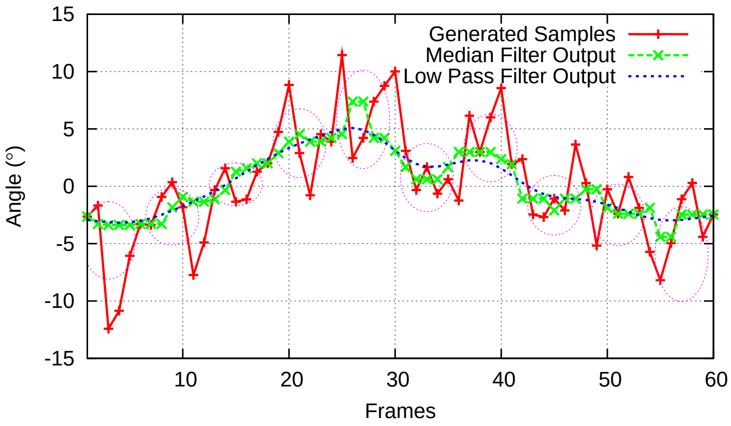


Figure 4.5: The effect of filtering in the synthesis of Euler angle θ. The dashed circles represent the states of a single branch HMM model. The vertical position and size of each circle are adjusted considering the mean and variance of the Euler angles associated with each state.

more natural looking synthesis results than using a fixed gesture dictionary, since humans produce slightly varying gestures at different occasions for the same semantics. The second advantage is generating gestures with varying durations in accordance with prosody of the speaker.

4.4 Evaluation and Results

In this section, we present experimental results and evaluation of the proposed system. Section 4.4.1 describes the audio-visual database, which is used in the experimental evaluation to generate objective and subjective results. The evaluation of the gesture-prosody pattern analysis is presented in Section 4.4.2, and the objective and subjective performance results for synthesis are presented in Section 4.4.3.

4.4.1 Database and Experimental Setup

We have conducted experiments using the MVGL-MASAL gesture-speech database. The database includes four recordings of a single subject telling stories in Turkish. Each story is approximately 7 minutes long and the total duration of the database is 27 min and 45 seconds. The audio-visual data is synchronously captured from the stereo camera and sound card. The stereo video includes only upper body gestures with 30 frames per second whereas the audio is recorded with 16 kHz sampling rate and 16 bits per sample. The detailed specification of the stereo camera can be found on [50]. The database is partitioned into two parts such that three stories are used for training of the models and one story is used for testing. For objective evaluation of the synthesis, the Euler angles extracted from the test sequence are considered as the ground truth for the synthesized head motion.

4.4.2 Analysis Results

The head gesture and prosody correlation analysis includes unsupervised temporal segmentation of the individual feature streams as well as the joint gesture-prosody label stream. The objective and subjective evaluation of these tasks are presented in the following.

Segmentation of Head Gesture Patterns

The parallel HMM Λ_g is trained with features extracted from the training video using Expectation-Maximization (EM) algorithm. The resulting HMM structure provides a probabilistic cluster model for unsupervised segmentation of head gestures into recurring elementary patterns.

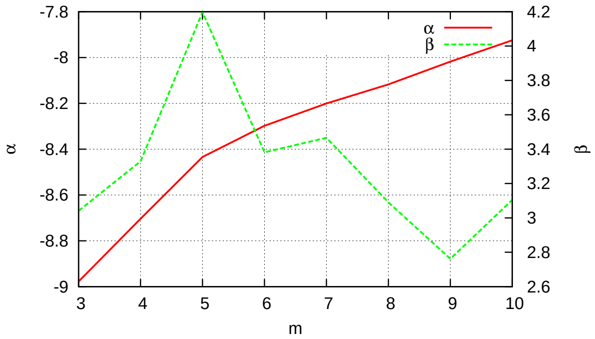


Figure 4.6: The α and β fitness measures for varying number of head gesture patterns

The number of branches, or equivalently the number of gesture patterns, M_{Λ_g} is a critical model parameter. In order to set M_{Λ_g}, the two fitness measures α and β, as respectively defined in (4.14) and (4.15), are calculated for varying number of gesture patterns and plotted in Fig. 4.6. The α measure, which yields the probability of model match, increases with increasing number of patterns as expected. The β measure, which yields statistical separation between patterns, has a maximum at $M_{\Lambda_g} = 5$. Hence, we set the number of gesture patterns M_{Λ_g} to 5, since the α and β measures are clearly jointly maximized at this value.

Consequently, when the training head gesture sequence is segmented using Λ_g, the segments belonging to the same gestural patterns are observed to be visually alike. The mean Euler angle vectors and the typical thumbnails for the five gesture patterns are depicted in Fig. 4.7.

Segmentation of Prosody Patterns

The speech prosody feature sequence is extracted from the audio part of the training database. As defined in stage I, the HMM model Λ_p is trained with prosodic features to obtain unsupervised temporal segmentation of the audio stream.

The two fitness measures α and β are calculated for varying number of prosody patterns using HMM model Λ_p and plotted in Fig. 4.8. The α measure, which yields the probability of model match, increases and the β measure, which yields statistical separation between patterns, decreases with increasing number of patterns as expected. The number of prosody patterns M_{Λ_p} is set to 5, since the α and β measures are jointly maximized around this value.

The means and standard deviations of the normalized pitch frequency trajectories for the five prosody patterns are depicted in Fig. 4.9. Note that, the first pitch trajectory (upper-left) is associated with the no-pitch segments that we filled with zero mean and unit variance Gaussian noise. The noise filling is necessary for successful modeling of those segments with continuous density HMMs. The other four prosody patterns can be classified using the prosodic transcription conventions introduced by the American English Tones and Break Indices (ToBI) standard [51]. The two prosody patterns on the upper right are both falling boundary tones (L%); the pattern on the lower left is a falling boundary tone, which makes a peak before the last syllable (HL%), and the pattern on the lower right is a rising-falling boundary tone, which rises within the last syllable (LHL%). We should note that these prosody patterns are obtained using unsupervised clustering over the training database, and they do not define a complete prosodic transcription convention for Turkish.

Segmentation of Joint Gesture-Prosody Patterns

In the first stage analysis, we obtain two independent HMM structures, Λ_g and Λ_p, respectively for recurrent head gesture and prosody patterns. We then extract two independent and parallel streams of head gesture and prosody pattern labels via temporal segmentation using these HMM structures. In the second stage, the discrete multi-stream HMM structure Γ_{gp} is trained using EM over the joint gesture-prosody pattern label stream to perform unsupervised segmentation. The number of states for each branch of Γ_{gp} is selected as $N_{\Gamma_{gp}} = 4$ to model possible label pair transitions. These four states model four different gesture-prosody label pair combinations within a joint gesture-prosody label pattern. Note that the extreme case, $N_{\Gamma_{gp}} = 1$, can only model a single co-occurrence pattern of gesture-prosody labels.

The two fitness measures α and β for Γ_{gp}, and also the number of gesture patterns in Λ_g, are considered for selection of the number of joint gesture-prosody label patterns $M_{\Gamma_{gp}}$. The number of joint patterns $M_{\Gamma_{gp}}$ is expected to be larger than or equal to the number of gesture patterns M_{Λ_g}, since in a robust synthesis process all the gesture patterns need to be generated for some temporal prosody label pattern. Hence, for the selection of $M_{\Gamma_{gp}}$, we present the two fitness measures α and β together with the normalized Euclidean distance measure ϵ_n as defined in (4.20) for varying number of joint gesture-prosody label patterns in Fig. 4.10. The $M_{\Gamma_{gp}}$ parameter is selected as 6, since this $M_{\Gamma_{gp}}$ value is greater than the M_{Λ_g} value, the ϵ_n distance has a minimum at $M_{\Gamma_{gp}} = 6$ and the α and β measures are jointly higher for $M_{\Gamma_{gp}} = 6$.

We observe that each branch of HMM Γ_g, which is the gesture stream of HMM Γ_{gp}, represents a single gesture pattern. Hence, each single gesture pattern is associated with one or more temporal prosody label patterns in the joint correlation model. Note that, the association between temporal prosody label patterns and single gesture pattern is valueable for the prosody-driven head gesture synthesis task.

4.4.3 Synthesis Results

Prosody-driven head gesture synthesis generates an Euler angle sequence, which is naturally correlated to a given test speech signal. The details of the synthesis process is given in Section 4.3. In this section, we present objective and subjective evaluations of the prosody-driven head gesture synthesis process. The evaluations are performed over the test database, which is defined in Section 4.4.1.

The objective evaluations compare the difference between original and synthesized Euler angles. Furthermore, A-B comparison type subjective evaluations are performed using the talking head avatar of *Momentum Inc.* [52], where the Euler angles that we deliver are used to drive head gestures/motion of the speech-driven talking head animation. The subjective tests are used to measure opinions on the naturalness of the synthesized head gestures using the speech-driven talking head animations.

We have also considered using an Input-Output Hidden Markov Model (IOHMM) structure [34, 53] for joint analysis of head gestures and prosody. In that case, the IOHMM structure replaces the HMM Γ_{gp} and builds gesture-prosody segment label mapping to predict the gesture segment labels from the prosody. The states in the IOHMM are fully connected and the number of states is selected to be same to the number of states in the Γ_{gp} model, which is 24. The IOHMM implementation in the Torch Machine Learning Library [54] is used in our experiments.

Objective Results

The objective evaluations compare the distance between original and synthesized Euler angles. In our evaluations we used three different distance measures. Let the original and synthesized Euler angles at frame k are represented with $\boldsymbol{e}_k$ and $\hat{\boldsymbol{e}}_k$, respectively. The first distance measure ϵ_n is a normalized Euclidean distance measure, which penalize Euler angles in wrong directions [32],

$$\epsilon_n = \sum_{k=1}^{K} \frac{(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)^T(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)}{(\hat{\boldsymbol{e}}_k + \boldsymbol{e}_k)^T(\hat{\boldsymbol{e}}_k + \boldsymbol{e}_k)} \tag{4.20}$$

Table 4.1: The distance measures between the original and the two sets of synthesized Euler angles: from the proposed Γ_{gp} and IOHMM models

Model	Γ_{gp}	IOHMM
ϵ_n	0.880	0.897
ϵ_m	1.798	1.857
ϵ_e	12.518	13.287

The second measure ϵ_m is the Mahalanobis distance, which is the Euclidean distance weighted with the inverse covariance matrix, $\mathbf{\Sigma}^{-1}$, of the original Euler angles $\boldsymbol{e}_k$,

$$\epsilon_m = \frac{1}{K}\sum_{k=1}^{K}\sqrt{(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)^T\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)} \tag{4.21}$$

The third distance measure is the Euclidian distance, $\epsilon_e = \frac{1}{K}\sum_{k=1}^{K}\sqrt{(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)^T(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)}$.

The original Euler angles from the visual part of the test database are extracted to be used as the ground truth in the objective evaluations. Two sets of synthesized Euler angles are generated using the audio part of the test database. The first set is the proposed head gesture synthesis system based on the Γ_{gp} model. The second set is generated with the same head gesture synthesis system as defined in Section 4.3 by replacing the second stage joint gesture-prosody correlation model Γ_{gp} by IOHMM. The three distance measures ϵ_n, ϵ_m and ϵ_e between the original and synthesized Euler angles are given in Table 4.1. Note that, all the three distance measures yield better distances for the proposed joint gesture-prosody correlation model Γ_{gp}.

Subjective Results

Subjective A-B comparisons are performed using the speech-driven talking head animations to measure opinions on the naturalness of the synthesized head gestures. The subjects are asked to evaluate the naturalness of the speech-driven synthesized head gestures for an A-B test pair on a scale of $(-2,\ -1,\ 0,\ 1,\ 2)$, where the scale corresponds to (A much better, A

Table 4.2: The Subjective A-B Comparison Results

A-B pair	Preference Score
Original - Γ_{gp}	-0.23
Original - IOHMM	-0.83
Γ_{gp} - IOHMM	-0.56
Identical pairs	0.04

better, no preference, B better, B much better).

The whole test database is manually partitioned into meaningful 15 segments, where each segment is approximately 12 seconds. For each evaluation 8 segments out of 15 are randomly selected. Three sets of A-B comparison pairs, each including these 8 segments, are considered for the speech-driven talking head animations using the original and two sets of synthesized Euler angles. Furthermore, three random startup A-B test pairs and another three test pairs with identical synthesis algorithms are also included to the subjective test set. Hence, the total number of A-B pairs in a test is 30. Apart from the three random start-up A-B pairs, all the pairs are randomized across conditions and pairwise. The subjective tests are performed over 15 subjects. The average preference scores for the three comparison sets are presented in Table 4.2. Note that, the scores of the three random start-up pairs are ignored in calculating the final preference scores. The subjective A-B comparisons, as expected, indicate a preference for the talking head animations using original Euler angles. The animations that are derived with the proposed joint gesture-prosody correlation model Γ_{gp} are preferred over the animations using IOHMM correlation model with an average preference score of -0.63. Also note that, the preference of the animations using the original Euler angles is stronger for the IOHMM driven animations than the proposed Γ_{gp} driven animations.

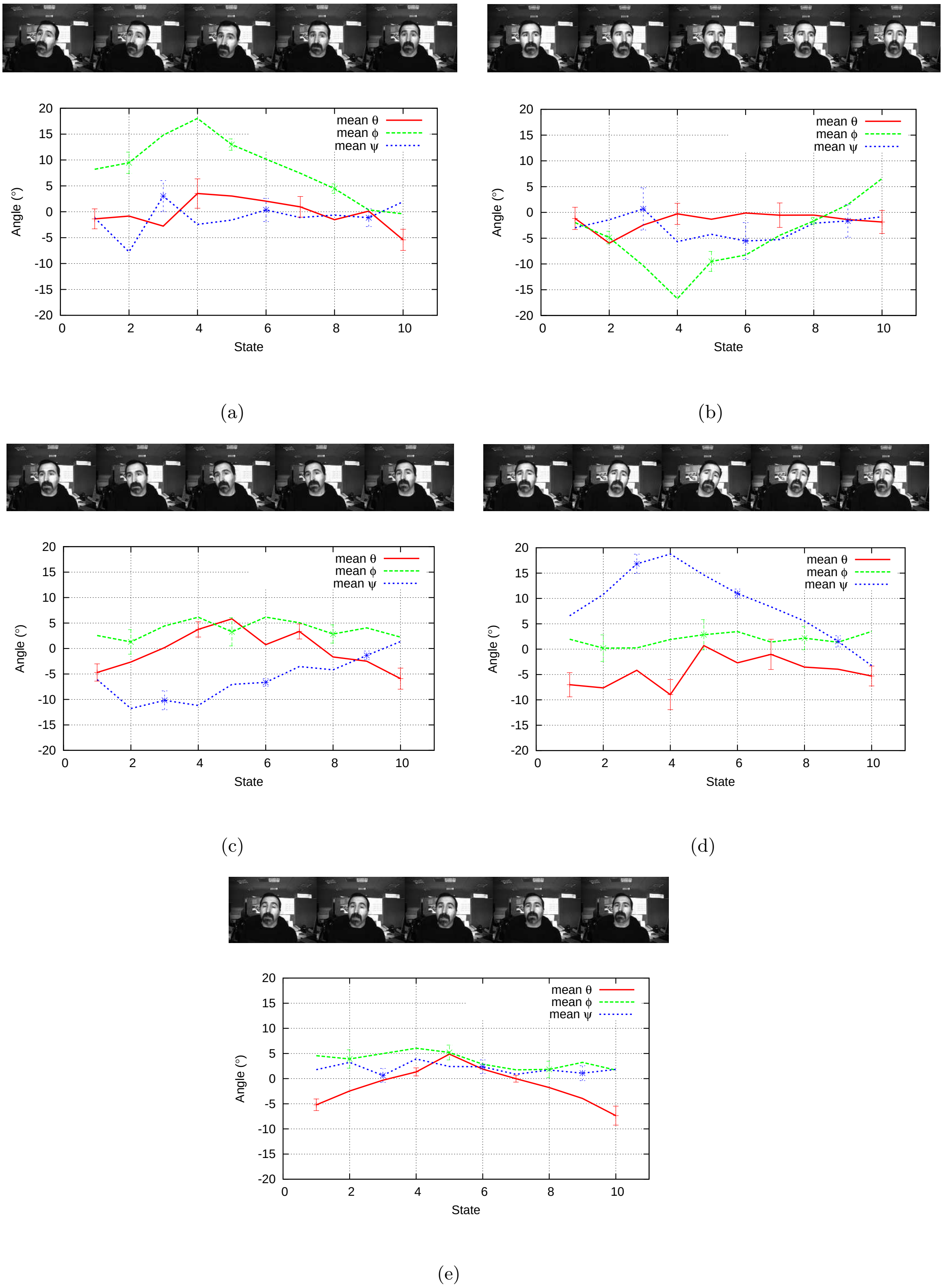
Samples of the audio-visual sequences for the prosody-driven talking head animations are available online [55]. These samples are selected to demonstrate three possible related applications. The first one is the speaker dependent prosody-driven gesture synthesis application, where gesture-prosody correlation model of a speaker is used to animate the same speaker with her/his speech. The second application is head gesture transplant, where gesture-prosody correlation model of speaker A is used to animate speaker B from speaker A's speech. Furthermore, the prosody transplant is considered as the third application, where gesture-prosody correlation model of speaker A is used to animate speaker A from speaker B's speech. In the demonstration of the prosody transplant we used speech input from audio-book recordings in English, where the gesture-prosody correlation model is performed over the story telling recordings in Turkish. Although one should expect differences in prosody patterns across different languages, the naturalness of the animations is observed to be acceptable. We also note that, the talking speed of these two speakers are different, where the native Turkish speaker has a faster rate than the native English speaker. As expected from the proposed correlation model, we observe slower head gesture animations for the native English speaker.

4.5 Discussions

We proposed a new two-stage joint head gesture and speech prosody analysis framework, where in the first stage elementary gesture and prosody patterns are extracted using unsupervised segmentation for a speaker, and in the second stage a correlation model between head gesture and prosody patterns is developed. The proposed two-stage analysis framework offers the following advantages: i) Meaningful elementary gesture and prosody patterns are defined for a speaker in the first stage. ii) A mapping between these elementary prosody and head gesture patterns is obtained with the unsupervised segmentation of joint gesture-prosody label stream. iii) The HMM-based analysis and synthesis yields flexibility in modeling structural and durational variations within gestural and prosodic patterns. iv) Automatic generation of the elementary gesture patterns produces natural looking prosody-driven head gesture synthesis.

In addition to successful demonstration of speaker dependent speech-driven head gesture synthesis system, different applications, such as head gesture transplant and prosody transplant, are also demonstrated. After extracting a gesture-prosody correlation model for speaker A, head gesture transplant animates speaker B from speaker A's speech, and prosody transplant animates speaker A from speaker B's speech. In the prosody transplant demonstration, gesture-prosody correlation model is trained with audio-visual recordings in Turkish, and prosody-driven gesture synthesis is performed with speech input recordings in English. The naturalness of the prosody transplant is found to be acceptable. Also in this demonstration, we observe slower head gesture animations for the native English speaker whose talking speed is slower.

The proposed HMM based two-stage head gesture and speech prosody analysis system can be utilized to model the correlation between any other losely correlated modalities, such as facial expressions and speech prosody, arm gestures and speech semantics, etc. Furthermore, the proposed speaker dependent speech-driven head gesture synthesis system can be tailored to model speaker's emotion and mood. Furthermore, we note that prosody patterns obtained using the proposed stage I analysis over a multi-speaker phonetically rich Turkish (or any other language) training database, can be used to define a complete ToBI-like prosodic transcription convention for Turkish (or any other language) intonation.



(a)

(b)

(c)

(d)

(e)

Figure 4.7: The mean Euler angles with standard deviations and typical thumbnails for the five gesture patterns: (a) Turn Left, (b) Turn Right, (c) Tilt Left, (d) Tilt Right, and (e) Nod

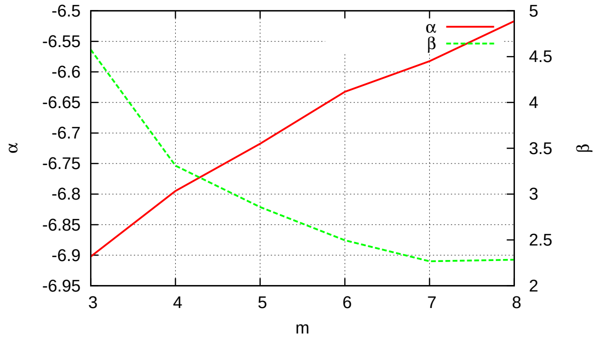


Figure 4.8: The α and β fitness measures for varying number of prosody patterns

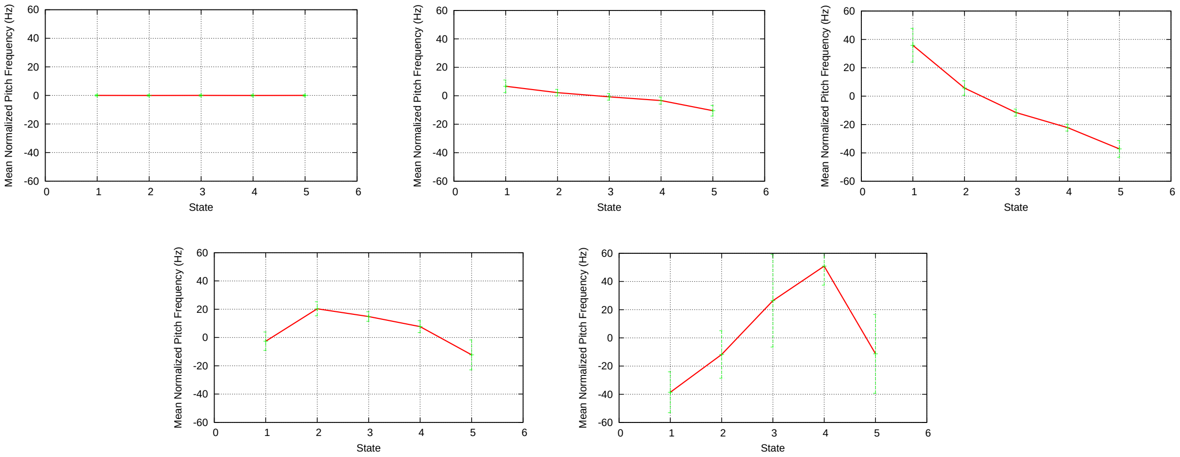


Figure 4.9: The means and standard deviations of the normalized pitch frequency trajectories for the five prosody patterns

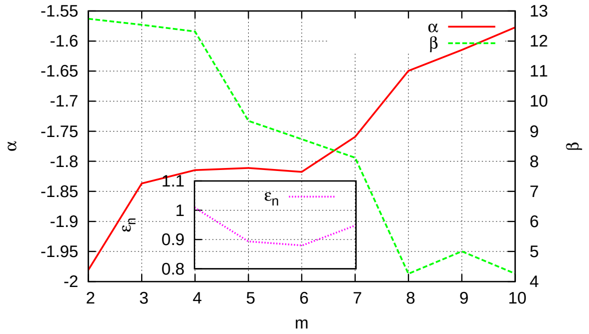


Figure 4.10: The α and β fitness measures and the normalized Euclidean distance measure ϵ_n for varying number of joint gesture-prosody label patterns

Chapter 5

CONCLUSIONS

In this thesis, correlation analysis of gesture and speech modality is investigated under different multi-modal applications. First, the application of CCA on the lip texture and speech modality for the speaker identification task is investigated. Then, the co-occurrence based correlation analysis of gesture and speech modality for the synthesis of head and arm gestures accompanying speech is studied. In this application, the correlation analysis is performed over pre-designated elements for gesture and speech modality. Using the clues obtained in gesture-speech correlation analysis, we focus on the head gesture and speech prosody correlation modeling to automatically synthesize head gestures from speech prosody.

In the CCA analysis, high precision feature synchronization is applied based on the maximization of the CCA-based correlation measure between the modalities. The synchronization improved the EER performance of the early integration since the correlation between the modalities are maximized. However, since the modalities contain both correlated and uncorrelated components neither late integration nor the early integration is optimal. The decomposition of both modalities into correlated and uncorrelated components using CCA and performing the optimal fusion strategy related with each component outperforms the other fusion strategies.

In the second multi-modal application, co-occurrence based correlation analysis is applied over the pre-designated gesture and speech events. The speech events are accents and directional keywords and gestural events are directional arm gestures and head nods-tilts which are determined manually by examining a training sequence. Correlation analysis is applied over the gesture-speech elements and experiments show that the directional keywords and accents are related with directional arm gestures and head nods-tilts respectively. The synthesis of gesture elements accompanying speech events made the animation more natural since the information given in speech is complimented with the gestural information.

Even though the gestural events improved the naturalness of the animation, confining the whole gesture set to a limited number of gestures is not the optimal way for synthesizing natural gestures. In addition, definition of the gesture element set for each subject is required since there not a common gesture element set that is valid for almost everyone as in the case of visemes.

The two-stage head gesture and speech prosody analysis deals with the HMM based unsupervised detection of elementary gesture and prosody patterns for a speaker in the first stage analysis. Then, in the second stage analysis, a correlation model between head gesture and prosody patterns is developed. The advantage of using HMM based analysis and synthesis is the flexibility in modeling structural and durational variations within gestural and prosodic patterns. We also observed that the automatic generation of the elementary gesture patterns produces natural looking prosody-driven head gesture synthesis. Furthermore, the correlation between any other losely correlated modalities (i.e. facial expressions and speech prosody, arm gestures and speech semantics) can be modeled using the proposed HMM based two-stage head gesture and speech prosody analysis system.

Appendix A

CANONICAL CORRELATION ANALYSIS

The canonical correlation analysis (CCA) is a linear statistical analysis technique, that provides a way of measuring how much and in what directions are two given multidimensional variables correlated. It was first proposed in [56], and then found applications in various fields [57],[58].

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two jointly Gaussian, zero-mean multidimensional variables with dimensions N_x and N_y, respectively. CCA seeks two linear transformations $\boldsymbol{H}_x$ and $\boldsymbol{H}_y$, one for each multidimensional variable, that maximize the mutual information between the transformed variables $\boldsymbol{x}'$ and $\boldsymbol{y}'$

$$\begin{aligned} \boldsymbol{x}' &= \boldsymbol{H}_x \boldsymbol{x} \\ \boldsymbol{y}' &= \boldsymbol{H}_y \boldsymbol{y}, \end{aligned} \tag{A.1}$$

where the multidimensional variables are represented with column vectors. We will refer to the pair $(\boldsymbol{x}', \boldsymbol{y}')$ as the CCA transform of $\boldsymbol{x}$ and $\boldsymbol{y}$. The transformations $\boldsymbol{H}_x$ and $\boldsymbol{H}_y$ are represented by matrices of dimensions $N \times N_x$ and $N \times N_y$, respectively, where $N \leq \min(N_x, N_y)$:

$$\boldsymbol{H}_x = \begin{bmatrix} \boldsymbol{h}_{x1}^T \\ \boldsymbol{h}_{x2}^T \\ \vdots \\ \boldsymbol{h}_{xN}^T \end{bmatrix}, \boldsymbol{H}_y = \begin{bmatrix} \boldsymbol{h}_{y1}^T \\ \boldsymbol{h}_{y2}^T \\ \vdots \\ \boldsymbol{h}_{yN}^T \end{bmatrix} \tag{A.2}$$

The rows of each of these matrices, $\{\boldsymbol{h}_{xi}\}$ and $\{\boldsymbol{h}_{yi}\}$, $i = 1, 2, ..., N$, form a basis for the corresponding transform space and are referred to as CCA basis vectors. The first pair of these basis vectors, $(\boldsymbol{h}_{x1}, \boldsymbol{h}_{y1})$, is given by the directions along which the projections are maximally correlated:

$$(\boldsymbol{h}_{x1}, \boldsymbol{h}_{y1}) = \arg \max_{(\boldsymbol{h}_x, \boldsymbol{h}_y)} \mathrm{Corr}(\boldsymbol{h}_x^T \boldsymbol{x}, \boldsymbol{h}_y^T \boldsymbol{y}) \tag{A.3}$$

subject to the constraints $\boldsymbol{h}_x^T \boldsymbol{C}_x \boldsymbol{h}_x = 1$ and $\boldsymbol{h}_y^T \boldsymbol{C}_y \boldsymbol{h}_y = 1$ where $\boldsymbol{C}_x$ and $\boldsymbol{C}_y$ are the covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. The projections, $x'_1 = \boldsymbol{h}_{x1}^T \boldsymbol{x}$ and $y'_1 = \boldsymbol{h}_{y1}^T \boldsymbol{y}$, are the first pair of canonical components. The second pair of CCA basis vectors can then be extracted using the residuals left after removing the components along the first pair of basis vectors from the original variables. This is equivalent to maximizing the same correlation, but this time subject to the constraint that the projections are to be uncorrelated with the first pair of canonical components. The same procedure can be iterated to extract the remaining canonical pairs.

The CCA basis vectors are usually computed by solving an equivalent eigenvalue problem. The joint covariance matrix of the two random variables $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E(\begin{bmatrix} \boldsymbol{x}\boldsymbol{x}^T & \boldsymbol{x}\boldsymbol{y}^T \\ \boldsymbol{y}\boldsymbol{x}^T & \boldsymbol{y}\boldsymbol{y}^T \end{bmatrix}) \tag{A.4}$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-set covariance matrices, $\mathbf{C}_{xy}$ is the between-set covariance matrix and $E(\cdot)$ is the expected value function. These covariance matrices can be estimated using a sufficiently representative set of realizations of the random variables. The problem of CCA then becomes solving the following eigenvalue equations,

$$\begin{aligned} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{h}_x &= \gamma^2 \mathbf{h}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{h}_y &= \gamma^2 \mathbf{h}_y, \end{aligned} \tag{A.5}$$

where the eigenvectors correspond to the normalized CCA basis vectors and each associated eigenvalue γ_i, $i = 1, 2, ..., N$, is the canonical correlation between the components of the corresponding canonical pair, x'_i and y'_i:

$$\gamma_i = E(x'_i y'_i). \tag{A.6}$$

Since the two solutions of (A.5) are related by

$$\mathbf{C}_{xy}\mathbf{h}_y = \gamma \lambda_x \mathbf{C}_{xx}\mathbf{h}_x, \tag{A.7}$$

where

$$\lambda_x = \sqrt{\frac{\mathbf{h}_y^T \mathbf{C}_{yy} \mathbf{h}_y}{\mathbf{h}_x^T \mathbf{C}_{xx} \mathbf{h}_x}}, \tag{A.8}$$

it suffices to solve only one of the eigenvalue equations.

As a result, the CCA transform diagonalizes the between-set covariance matrix,

$$\boldsymbol{C}_{x'y'} = \boldsymbol{H}_x \boldsymbol{C}_{xy} \boldsymbol{H}_y^T \tag{A.9}$$

so that the diagonal entries of the resulting covariance $\boldsymbol{C}_{x'y'}$ correspond to the canonical correlations, γ_i. Similarly, the non-diagonal entries, which are all zero, are the cross-correlations,

$$E(x_i' y_j') = 0 \quad \text{for all } i \neq j. \tag{A.10}$$

Moreover, since the pairs of canonical components are uncorrelated with each other, we also have

$$E(x_i' x_j') = E(y_i' y_j') = 0 \quad \text{for all } i \neq j. \tag{A.11}$$

Appendix B

RIGID MOTION PARAMETER ESTIMATION BY UNITARY CONSTRAINED OPTIMIZATION

This appendix summarizes the method used for estimating the rotation matrix, $\boldsymbol{R}$, and translation vector, $\boldsymbol{t}$, that describe the rigid motion between the world point coordinate matrices $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ (see Section 4.1.1).

Let $\boldsymbol{m}_k$ denote the mean of the column vectors in the matrix $\boldsymbol{W}_k$ such that

$$\boldsymbol{m}_k = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_k^i \tag{B.1}$$

and $\boldsymbol{m}_{k_r}$ be defined similarly. Then, the translation $\boldsymbol{t}$ between $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ is given by

$$\boldsymbol{t} = \boldsymbol{m}_k - \boldsymbol{m}_{k_r} \tag{B.2}$$

Furthermore, let $\boldsymbol{W}'_{k_r}$ and $\boldsymbol{W}'_k$ represent the mean-removed coordinate matrices such that

$$\boldsymbol{W}'_k = \boldsymbol{W}_k - \boldsymbol{m}_k \mathbf{1}^T, \text{ and } \boldsymbol{W}'_{k_r} = \boldsymbol{W}_{k_r} - \boldsymbol{m}_{k_r}\mathbf{1}^T \tag{B.3}$$

Then, the rotation matrix $\boldsymbol{R}$ can be found by minimizing the cost function

$$f(\boldsymbol{R}) = \|\boldsymbol{E}\|_F^2 = \mathrm{tr}(\boldsymbol{E}\boldsymbol{E}^T) \tag{B.4}$$

where $\|\cdot\|_F$ and $\mathrm{tr}(\cdot)$ denote the Frobenius-norm and the matrix trace, respectively, and

$$\boldsymbol{E} = \boldsymbol{W}'_k - \boldsymbol{R}\boldsymbol{W}'_{k_r} \tag{B.5}$$

The minimization of the cost function $f(\boldsymbol{R}), f : \mathbb{R}^{3\times 3} \to \mathbb{R}$, is a non-linear optimization problem, under the unitary constraint $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$, i.e. over the orthogonal group $O(3)$, which can be solved by the algorithm proposed in [59], where Manton proposed a modified Newton method for optimization on the complex Stiefel manifold which defines the space related with the unitary constraint.

We simplified this method to minimize the cost function $f(\boldsymbol{R})$ for a square and real matrix $\boldsymbol{R}$ subject to the constraint $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$ as follows:

1. Choose initial $\boldsymbol{R}$ such that $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$.

   We initialize the rotation matrix $\boldsymbol{R}$ using a linearized version of the optimization problem. For small rotations, $\boldsymbol{R}$ can be approximated in terms of a parameter vector $\boldsymbol{u} = [u_x, u_y, u_z]^T$ such that [60]

   $$\boldsymbol{R} \approx \boldsymbol{I} + \boldsymbol{S} = \boldsymbol{I} + \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix} \tag{B.6}$$

   Equating the residual defined in (B.5) to zero, we obtain the following equation to solve for $\boldsymbol{S}$:

   $$\boldsymbol{W}'_k - \boldsymbol{W}'_{k_r} = \boldsymbol{S}\boldsymbol{W}'_{k_r} \tag{B.7}$$

   which can be expressed in terms of $\boldsymbol{u}$ as

   $$\text{vec}(\boldsymbol{W}'_k - \boldsymbol{W}'_{k_r}) = \boldsymbol{K}\boldsymbol{u} = \begin{bmatrix} \boldsymbol{K}_1 \\ \vdots \\ \boldsymbol{K}_N \end{bmatrix} \boldsymbol{u}, \quad \boldsymbol{K}_n = \begin{bmatrix} 0 & Z_n & -Y_n \\ -Z_n & 0 & X_n \\ -X_n & Y_n & 0 \end{bmatrix} \tag{B.8}$$

   where each 3×3 sub-matrix $\boldsymbol{K}_n$ is constructed using the nth point (X_n, Y_n, Z_n) from $\boldsymbol{W}'_{k_r}$. The least squares solution of (B.8) can then be used to find $\boldsymbol{u}$ and to construct $\boldsymbol{S}$. The initial guess for $\boldsymbol{R}$ can finally be obtained by projection onto the unitary space $\boldsymbol{R} = \pi(\boldsymbol{I} + \boldsymbol{S})$ (described in step 5 below).

2. Compute the derivative $\boldsymbol{D}_R$ and the Hessian $\boldsymbol{H}_R$ of f at the point $\boldsymbol{R}$ given by

   $$\boldsymbol{D}_R = -2\boldsymbol{E}\boldsymbol{W}'^T_{k_r} \tag{B.9}$$

   $$\boldsymbol{H}_R = -2((\boldsymbol{W}'_{k_r}\boldsymbol{W}'^T_{k_r}) \otimes \boldsymbol{I}_{3\times3}) \tag{B.10}$$

3. If $\sqrt{tr(\boldsymbol{D}_R^T\boldsymbol{D}_R - \boldsymbol{R}^T\boldsymbol{D}_R\boldsymbol{R}^T\boldsymbol{D}_R)} < \epsilon$, then stop.

4. Compute the Newton step size $\boldsymbol{Z} := \boldsymbol{Z}^{(cp)}$.

   The Newton step size is defined as the value of $\boldsymbol{Z}$, $\boldsymbol{Z} \in \mathbb{R}^{3\times 3}$, confined to the tangent space V where $V = \{\boldsymbol{RA} : \boldsymbol{A} = -\boldsymbol{A}^T\}$, at which the quadratic approximation $g(\boldsymbol{Z})$ has its critical point:

$$g(\boldsymbol{Z}) \approx f(\boldsymbol{R}) + \mathrm{tr}(\boldsymbol{Z}^T\boldsymbol{D}) + (1/2)\,\mathrm{vec}(\boldsymbol{Z})^T\boldsymbol{H}\,\mathrm{vec}(\boldsymbol{Z}) \tag{B.11}$$

   where

$$\boldsymbol{D} = \boldsymbol{D}_R, \quad \boldsymbol{H} = \boldsymbol{H}_R - (1/2)[(\boldsymbol{R}^T\boldsymbol{D}_R + \boldsymbol{D}_R^T\boldsymbol{R})^T \otimes \boldsymbol{I}] \tag{B.12}$$

   The tangent space V is defined as a subset of $\mathbb{R}^{3\times 3}$ such that $\boldsymbol{Z} = \boldsymbol{RA}$ where $\boldsymbol{A}$ is skew-symmetric. The critical point $\boldsymbol{Z}^{(cp)} \in V$, i.e. the Newton step size, satisfies the following linear constraint:

$$tr(\boldsymbol{Z}^T D) + \left[\mathrm{vec}(\boldsymbol{Z})^T\boldsymbol{H}\right]\mathrm{vec}(\boldsymbol{Z}^{(cp)}) = 0 \tag{B.13}$$

   By writing $\boldsymbol{Z}$ as $\boldsymbol{Z} = \sum_{i=1}^{3}\alpha_i\boldsymbol{RA}_i$, where $\boldsymbol{A}_i \quad (i = 1, 2, 3)$ is an arbitrary basis for skew-symmetric matrix, the critical point $\boldsymbol{Z}^{(cp)}$ can be found by solving the following linear equation for $\alpha_i \quad (i = 1, 2, 3)$

$$\frac{\partial g(\boldsymbol{Z})}{\partial \alpha_i} = \mathrm{tr}((\boldsymbol{RA}_i)^T\boldsymbol{D}) + \mathrm{vec}(\boldsymbol{RA}_i)^T\boldsymbol{H}\,\mathrm{vec}(\boldsymbol{Z}) = 0 \tag{B.14}$$

   Note that the above equation can be put into a matrix form:

$$\boldsymbol{K}^T\boldsymbol{HK\alpha} = \boldsymbol{K}^T\,\mathrm{vec}(\boldsymbol{D}) \tag{B.15}$$

   where

$$\boldsymbol{K} = (\boldsymbol{I}\otimes\boldsymbol{R})[\mathrm{vec}(\boldsymbol{A}_1), \mathrm{vec}(\boldsymbol{A}_2), \mathrm{vec}(\boldsymbol{A}_3)], \quad \boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]^T \tag{B.16}$$

5. Set $\boldsymbol{R}' := \pi(\boldsymbol{R} + \boldsymbol{Z})$.

   The projection $\pi(\boldsymbol{R})$, $\pi : \mathbb{R}^{3\times 3} \to St$, onto the Stiefel manifold, $St = \{\boldsymbol{R} \in \boldsymbol{R}^{3\times 3} : \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}\}$, is defined as

$$\pi(\boldsymbol{R}) = \underset{\boldsymbol{Q}\in St}{\mathrm{argmin}}\,\|\boldsymbol{R} - \boldsymbol{Q}\|^2. \tag{B.17}$$

   If the singular value decomposition of $\boldsymbol{R}$ is $\boldsymbol{U\Sigma V}^T$, then the projection is simply given by [59]

$$\pi(\boldsymbol{R}) = \boldsymbol{UV}^T \tag{B.18}$$

6. If $f(\boldsymbol{R} \leq f(\boldsymbol{R}'))$ then abort.

7. Set $\boldsymbol{R} := \boldsymbol{R}'$. Go to Step 2.

BIBLIOGRAPHY

[1] T. Chen, “Audio-visual speech processing,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech,” *Proc. of the IEEE*, vol. 91, no. 9, September 2003.

[3] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, “Gesture-speech correlation analysis and speech driven gesture synthesis,” in *Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME 2006)*, 2006.

[4] R.W. Frischholz and U. Dieckmann, “BioID: A multimodal biometric identification system,” *Journal of IEEE Computer*, vol. 33, no. 2, pp. 64–68, February 2000.

[5] E. Erzin, Y. Yemez, and A.M. Tekalp, “Multimodal speaker identification using an adaptive classifier cascade based on modality reliability,” *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, October 2005.

[6] T. Wark and S. Sridharan, “Adaptive fusion of speech and lip information for robust speaker identification,” *Digital Signal Processing*, vol. 11, no. 3, pp. 169–186, July 2001.

[7] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, “Multimodal speaker identification using canonical correlation analysis,” *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2006 (ICASSP ’06)*, vol. I, pp. 613–616, 2006.

[8] T. Chen, “Audiovisual speech processing,” *IEEE Signal Processing Mag.*, vol. 18, pp. 9–21, 2001.

[9] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.

[10] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 1, pp. 406–413, June 2004.

[11] Quan-Sen Sun, Sheng-Gen Zeng, Pheng-Ann Heng, and De-Sen Xia, "Feature fusion method based on canonical correlation analysis and handwritten character recognition," in *IEEE Int. Conf. on Control, Automation, Robotics and Vision Conference, 2004 (ICARCV)*, 2004, vol. 2, pp. 1547–1552.

[12] K. Choukri, G. Chollet, and Y. Grenier, "Spectral transformation through canonical correlation analysis for speaker adaptation in asr," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1986 (ICASSP '86)*, 1986, pp. 2659–2662.

[13] S.Nakamura, K.Kumatani, and S.Tamura, "Multi-modal temporal asynchronicity modeling by product hmms for robust audio-visual speech recognition," in *Proc. of the Int. Conf. on Multimodal Interfaces 2002 (ICMI '02)*, 2002, pp. 305–309.

[14] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. Neural Information Processing Systems*, 2000, pp. 814–820.

[15] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, "Lip feature extraction based on audio-visual correlation," *Proc. of the European Signal Processing Conference (EUSIPCO'05)*, 2005.

[16] J.-M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, December 1995.

[17] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp, "Discriminative Lip-Motion Features for Biometric Speaker Identification," *Proc. of the Int. Conf. on Image Processing 2004 (ICIP 2004)*, pp. 2023–2026, October 2004.

[18] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, 1976.

[19] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody based co-analysis for continuous recognition of coverbal gestures," *Proc. ICMI'02*, 2002.

[20] F. Quek, D. McNeill, R. Ansari, X-F. Ma, R. Bryll, S. Duncan, and K.E. McCullough, "Gesture cues for conversational interaction in monocular video," *Proc. Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems'99*, 1999.

[21] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," *Proc. of the European Signal Processing Conference 2002 (EUSIPCO'02)*, vol. 1, pp. 75–78, 2002.

[22] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, "Combined gesture-speech analysis and synthesis," *Proc. of the eNTERFACE'05 The SIMILAR Workshop on Multimodal Interfaces*, August 2005.

[23] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. of the Inst. of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[24] Jie Yang and Yangsheng Xu, "Hidden markov model for gesture recognition," Tech. Rep. CMU-RI-TR-94-10, Robotics Institute, Carnegie Mellon University, May 1994.

[25] S. Duncan, F. Parrill, and D. Loehr, "Discourse factors in gesture and speech prosody," *Conf. of the International Society for Gesture Studies (ISGS)*, 2005.

[26] M.A. Epstein, "Voice quality and prosody in english," *Proceedings of the XVth International Congress of Phonetic Sciences*, 2003.

[27] M.E. Sargin, "Output of the animation engine," 2005, Available in `http://home.ku.edu.tr/~msargin/icme06/`.

[28] S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '89)*, pp. 1795–1798, 1989.

[29] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," *Proc. ACM SIGGRAPH '97*, pp. 353–360, 1997.

[30] F.J. Huang and T. Chen, "Real-time lip-synch face animation driven by human voice," *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pp. 352–357, 1998.

[31] E. Yamamoto, S. Nakamura, and K. ShiKano, "Lip movement synthesis from speech based on hidden markov models," *Speech Communication*, pp. 105–115, 1998.

[32] M. Brand, "Voice puppetry," *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 21–28, 1999.

[33] P. S. Aleksic and A. K. Katsaggelos, "Speech-to-video synthesis using facial animation parameters," *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 14, no. 5, pp. 682–692, 2004.

[34] Y. Li and H.-Y. Shum, "Learning dynamic audio-visual mapping with inputoutput hidden markov models," *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 542–549, 2006.

[35] J. Xue, J. Borgstrom, J. Jiang, L.E. Bernstein, and A. Alwan, "Acoustically-driven talking face synthesis using dynamic bayesian networks," in *Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME 2006)*, 2006, pp. 1165–1168.

[36] K.G. Munhall, Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," in *PSYCHOLOGICAL SCIENCE*, 2004, vol. 15, pp. 133–137.

[37] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K.E. McCullough, "Gesture cues for conversational interaction in monocular video," *ICCV99 Wksp on RATFGRTS*, pp. 64–69, 1999.

[38] Takaaki Kuratate, Kevin G. Munhall, Philip E. Rubin, Eric Vatikiotis-Bateson, and Hani Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*, 1999, pp. 1279–1282.

[39] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: facial movements accompanying speech," *Proc. of IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 381–386, 2002.

[40] Zhigang Deng, Carlos Busso, Shri Narayanan, and Ulrich Neumann, "Audio-based head motion synthesis for avatar-based telepresence systems," in *ACM SIGMM 2004 Workshop on Effective Telepresence (ETP'04)*, 2004, pp. 24–30.

[41] M.R. Naphade and T.S. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proc. of the Int. Conf. on Image Processing 2002 (ICIP 2002)*, 2002, pp. II: 13–16.

[42] P. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE CVPR*, pp. 511–518, 2001.

[43] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," *Proc. of the Int. Conf. on Image Processing 2002 (ICIP'02)*, vol. 1, pp. 900–903, September 2002.

[44] J. Y. Bouguet, "Pyramidal implementation of the lucas kanade feature trackerdescription of the algorithm," *Intel Corporation, Microprocessor Research Labs, OpenCVDocuments*, 1999.

[45] MZ Brown, D. Burschka, and GD Hager, "Advances in computational stereo," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 8, pp. 993–1008, 2003.

[46] P. Fua, "Combining stereo and monocular information to compute dense depth maps

that preserve depth discontinuities," *12th. International Joint Conference on Artificial Intelligence*, pp. 1292–1298.

[47] D.A. Varshalovich, A.N. Moskalev, and V.K. Khersonskii, *Description of Rotation in Terms of the Euler Angles*, Quantum Theory of Angular Momentum. World Scientific, 1988.

[48] K. Shoemake, "Animating rotation with quaternion curves," *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

[49] S. Ananthakrishnan and S. Narayanan, "An Automatic Prosody Recognizer using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model," *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1, 2005.

[50] ," Point Grey Research Inc. `http://www.ptgrey.com/`.

[51] J.Hirschberg and G.Ward, "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english," *Journal of Phonetics*, vol. 20, pp. 241–251, 1992.

[52] ," Momentum Inc. Speech-Driven Talking Head Avatar is available at `http://www.momentum-dmt.com/`.

[53] Y. Bengio and P. Frasconi, "Input-output HMMs for sequence processing," *Neural Networks, IEEE Transactions on*, vol. 7, no. 5, pp. 1231–1249, 1996.

[54] R. Collobert, S. Bengio, and J. Mariethoz, "Torch: a modular machine learning software library," *IDIAP Research Report*, vol. 2, pp. 46, 2002.

[55] ," Prosody-Driven Head Gesture Animation demonstrations are available at `http://mvgl.ku.edu.tr/prosodygesture/`.

[56] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[57] M. Borga, "Learning multidimensional signal processing," *PhD Thesis, Linkoping University, Sweden*, vol. Dissertation No 531, 1998.

[58] D. R. Hardoon, S. Szedmak, and J. S. Taylor, "Canonical correlation analysis: An overview with application learning," *Technical Report, Department of Computer Science, University of London*, , no. CSD-TR-03-02, 2003.

[59] Jonathan H. Manton, "Optimisation algorithms exploiting unitary constraints," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635–650, March 2002.

[60] D. Demirdjian and T. Darrell, "Motion estimation from disparity images," in *Proc. of the Eighth IEEE Int. Conf. on Computer Vision*, 2001, vol. 1, pp. 213–218.

VITA

MEHMET EMRE SARGIN was born in Izmir, Turkey on Augusy 25, 1982. He received his B.Sc. degree in Electrical and Electronics Engineering from Middle East Technical University, Ankara, Turkey, in 2004. From August 2004 to July 2006, he worked as a teaching and research assistant in Koç University, Istanbul, Turkey. At Koç University, he focused on Audio-Visual Signal Processing for Speaker Recognition and Synthesis, which has been supported by TUBITAK and European FP6 Network of Excellence SIMILAR project. He attended the SIMILAR NoE Summer Workshop on Multi-Modal Interfaces, Mons, Belgium on July-August 2005 where he worked as the project leader. He also worked as a visiting researcher at Polytechnic University of Catalonia (UPC), Barcelona, Spain on September-October 2005 for SIMILAR project.

He has submitted 2 journal publications and published several papers about Correlation Analysis for Biometric Speaker Identification and Speech Driven Gesture Synthesis in the following conferences: SIU2005 (Kayseri, Turkey), eNTERFACE2005 (Mons, Belgium), EUSIPCO2005 (Antalya, Turkey), SIU2006 (Antalya, Turkey), ICASSP2006 (Toulouse, France), ICME2006 (Toronto, Canada).