Preface

It has been widely accepted that speech perception is a multimodal process and involves information from more than one sensory modality. The famous McGurk effect [McGurk and MacDonald, Nature 264(5588): 746–748, 1976] shows that visual articulatory information is integrated into our perception of speech automatically and unconsciously. For example, a visual /ga/ combined with an auditory /ba/ is often heard as /da/. This effect is shown to be very robust and knowledge about it seems to have very little effect on one's perception of it.

Interest in machine lip reading starts to emerge in the mid-1980s (Petajan was probably the first to investigate the problem of machine lipreading (E.D. Petajan 1984), ,when it was shown that visual lip information extracted from the speaker's lip can enhance the performance of automatic speech recognition system, especially in noisy environment. Recently, it has also been shown that dynamics of the speaker's lip during speech articulation provides useful biometric information for speaker recognition.

Machine lip reading or visual speaker recognition, generally involves three major steps: lip segmentation, feature extraction, and classifier design. Although significant research effort and many technological advances have been made recently, machine lip reading is still far from practical deployment. Unlike the relatively mature field of automatic speech recognition, there are still many unsolved theoretical and algorithmic issues in machine lip reading. For example, the problems of lighting, shadow, pose, facial hair, camera resolution, and so forth, make reliable segmentation and extraction of lip feature a difficult task. The problem is further compounded by the difficult and variable environment visual speech recognition systems tend to operate in. There is also relatively little theoretical study on the amount of phonetic/ linguistic information that can be extracted from the speaker's lip for speech perception.

In this book, we introduce the readers to the various aspects of this fascinating research area, which include lip segmentation from video sequence, lip feature extraction and modeling, feature fusion, and classifier design for visual speech recognition and speaker verification. This book collects together recent state-of-the-art research in these areas. There are altogether 17 chapters, from 44 authors in 14 countries & regions (Australia, Canada, China, France, Germany, Greece, Hong Kong, Japan, Mexico, New Zealand, Singapore, Sweden, Turkey, and the United States). Many of the contributing authors are well-known researchers in the field. This book would be of great interest to researchers and graduate students working in the fields of audiovisual speech and speaker recognition.

The 17 chapters in the book are organized into four sections: *Section I: Introduction & Survey, Section II: Lip Modeling, Segmentation, and Feature Extraction, Section III: Visual Speech Recognition, and Section IV: Visual Speaker Recognition*. Section I contains four survey/tutorial type chapters (Chapter I to IV) that describe recent progress in the field. They serve to introduce readers to this emerging field of research and summarize the state-of-the-art techniques that are available today. Section II contains four chapters (chapter V to VIII), that deal with lip segmentation, modeling, and feature extraction. Section III (Chapter IX to XV) contains chapters that look at issues related specifically to visual speech recognition. For example, chapter X investigates the use of multiple views of the speaker for visual speech recognition, chapter XI and XII concentrates on classifier design and training, chapter XIV looks at the possibility of obtaining prosodic information from the lip, and chapter XV discusses perceptual studies that quantify the information content in visible speech. Section IV contains two chapters (chapter XVI and XVII) that describe the use of visual lip feature for biometric applications. Below we give a brief description of each chapter.

Chapter I, "Audio-Visual and Visual-only Speech and Speaker Recognition- issues about theory, system design, and implementation" provides a tutorial coverage of the research in audio-visual speech and speaker recognition. It describes the major research issues and techniques in feature extraction, feature fusion, classifier design, and performance evaluation, and lists the major audiovisual speech databases used for performance evaluation. The authors survey several current audiovisual speaker/speech recognition systems and discussed challenges and future research directions.

Chapter II, "Lip Feature Extraction and Feature Evaluation in the Context of Speech and Speaker Recognition" surveys the different dynamic lip features and their use in visual speech and speaker recognition. In this chapter, the focus is more on the approaches for detection and tracking of important visual lip features. Together, the two survey chapters serve to introduce readers to the exciting field of audiovisual speech and speaker recognition.

Chapter III, "Lip modeling and segmentation" is a detailed survey of state-of-the-art in lip modeling and segmentation. The authors discussed about the different color spaces that provide good separation of lip and non-lip region. They describe the two major approaches for lip segmentation, that is, contour-based and region-based. Within each category, they further categorize the methods as deterministic, statistical, supervised, or un-supervised. Different techniques to extract the lip, such as active shape model, snake, parametric model, deformable template, as well as their optimization, are discussed. The chapter further discusses different performance evaluation techniques and concludes by discussing some possible applications that would benefit from advances in lip segmentation.

Chapter IV "Visual Speech and Gesture Coding using the MPEG-4 Face and Body Animation Standard" introduces the MPEG-4 Face and Body Animation (FBA) standard for representing visual speech data as part of a whole virtual human specification. The super low bit-rate FBA codec included with the standard enables thin clients to access processing and communication services over any network including enhanced visual communication, animated entertainment, man-machine dialog, and audio/visual speech recognition. In the chapter, the author described the deployment of the MPEG-4 FBA standard in face animation, body animation, and visual speech processing. The computing architectures that support various applications are also outlined. This chapter would be of great interest to readers interested in the topic of Human Computer Interaction.

Chapter V "Lip Region Segmentation with Complex Background" describes a lip segmentation method that is able to handle complex non-lip region such as the presence of beard or shadow. The method employs a Multi-class, Shape-guided FCM (MS-FCM) clustering algorithm to separate the lip pixels from the non-lip pixels. A spatial penalty term, based on the lip shape information is introduced in the clustering algorithm, which boosts the lip membership for pixels inside the lip region while penalizes the value for pixels outside the lip region. With the spatial penalty term, lip and non-lip pixels with similar color but located in different regions can be differentiated.

Chapter VI "Lip Contour Extraction from Video Sequences under Natural Lighting Conditions" presents an algorithm for lip contour tracking under natural lighting conditions. The algorithm extracts the inner and outer lip borders from color video sequences. To extract the lip contour, the video images are processed in three steps. In step one the mouth area is segmented using color and movement information from the face skin. In step two, the mouth corners are detected. The mouth corners are used to initialize the active contours. Step three extracts the lip contour using active contours. The chapter gives detail description about the logarithmic hue-like color space transformation that are used to separate the lip and non-lip pixels, the hierarchical spatiotemporal color segmentation algorithm integrating hue and motion information, and finally, the theoretical derivation used in the optimization of the active contour. The algorithm has been successfully applied to several video sequences with no specific model of the speaker and variable illumination conditions.

Chapter VII "3D Lip Shape SPH Based Evolution Using Prior 2D Dynamic Lip Features Extraction and Static 3D Lip Measurements" describes a 3D lip modeling and animation technique whose dynamic behavior is governed by Smooth Particles Hydrodynamics. The 3D lip model is constructed from facial data acquired by a 3D scanner and 2D lip contours extracted from video-sequences of the subject. The influence of muscle contraction around the lip is considered in the model. The authors described in detail the 3D model construction and the derivation of the controlling forces that drive the model, and presented some simulation results to show the feasibility of their approach.

Chapter VIII "How to Use Manual Labelers in the Evaluation of Lip Analysis Systems?" examines the issues involved in evaluating and calibrating labeled lip features which serve as ground truth from human operators. The authors showed that subjective error in manual labeling can be quite significant, and this can adversely affect the validity of an algorithm's performance evaluation and comparative studies between algorithms. The chapter describes an iterative method based on Expectation-Maximization to statistically infer the ground truth from manually labeled data.

Chapter IX "Visual Speech Processing and Recognition" describes an algorithm that performs limited vocabulary recognition of the first four digits in English based on visual speech features. Three aspects of visual speech processing and recognition, namely, mouth region segmentation, lip contour extraction, and visual speech recognition are dealt with. For the mouth region segmentation, a modified Fuzzy C-means method with the addition of spatial constraints is introduced. For the lip contour extraction, the image gradient information is used to create a color map of edge magnitude and edge direction. Edge following is then applied on the color map to extract the lip contour. For the visual speech recognition, a SVM dynamic network is proposed. SVM classifiers are used to obtain the posterior probabilities and the SVMs are then integrated into a Viterbi decoding lattice for each visemic word. The authors showed that the SVM dynamic network has superior performance compared to some existing techniques.

Chapter X "Visual Speech Recognition across Multiple Views" investigates the use of multiple views of the speaker for visual speech recognition of connected digit strings. Most works on visual speech recognition assume that the speaker's face is captured in a frontal pose. However, in many applications, this assumption is not realistic. In this chapter, the authors considered the frontal and the profile views captured synchronously in a multi-camera setting. An appearance-based visual front-end that extracts features for frontal and profile videos is first developed. The extracted features then undergo normalization across views to achieve feature-space invariance. This normalization allows recognizing visual speech using a single pose-invariant statistical model, regardless of camera view.

Chapter XI "Hidden Markov Model Based Visemes Recognition. Part I: AdaBoost Approach" describes an AdaBoost-HMM classifier for visemes recognition. The authors applied AdaBoost technique to HMM modeling to construct a multi-HMM classifier that improves the recognition rate. Whereas conventional single HMM identifies the ideal samples with good accuracy but fail to handle the hard or outlier samples, Adaboosting allows new HMMs in a multi-HMM classifier to bias towards the hard samples, thus ensuring coverage of the hard samples with a more complex decision boundary. The method is applied to identify context-independent and context-dependent visual speech units. The technique was compared to conventional HMM for visemes recognition and has shown improved performance.

Chapter XII "Hidden Markov Model Based Visemes Recognition. Part II: Discriminative Approaches" describes an alternative approach to classifier training for visemes recognition. The focus is on emphasizing the minor differences between pairs of confusable training samples during HMM classifier training. The authors proposed two training approaches to maximize discrimination: Maximum Separable Distance (MSD) training and Two-channel HMM training. Both training approaches adopt a criterion function called separable distance to improve the discriminative power of an HMM classifier. The methods are applied to identify confusable visemes and their results indicate that higher recognition accuracy can be attained using these approaches than using conventional HMM.

Chapter XIII "Motion Features for Visual Speech Recognition" studies the motion features that are effective for visual speech recognition. A review of two motion feature extraction techniques, namely, optical flow method and image subtraction method is given. The authors then present their recent work on the motion history image (MHI) technique. The MHI method captures the lip dynamics through temporal integration of image sequence into a 2-D spatio-temporal template. Feature vectors based on DCT coefficients or Zernike moments are then computed from the MHI image and are used for visual speech recognition.

Chapter XIV "Recognizing Prosody from the Lips: Is it Possible to Extract Prosodic Focus from Lip Features?" This chapter investigates the feasibility of extracting prosodic information from visual lip features. Prosodic information plays a critical role in spoken communication, and reflects not only the emotional state of the speaker, but also carries crucial linguistic information, such as whether an utterance is a statement, a question, or a command, or whether there is an emphasis, contrast or focus. The authors used two lip feature measurement techniques to evaluate the lip pattern of prosodic focus in French. Lip opening and spreading and lip protrusion gestures are tracked and the lip features analyzed for prosodic focus in a natural dialogue situation.

Chapter XV "Visual Speech Perception, Optical Phonetics, and Synthetic Speech" reviews perceptual studies that quantify the information content in visible speech, demonstrating that visible speech is a rich and detailed source of phonetic information. The authors discussed the relations between optical phonetic signals and phonetic perception and demonstrated the existence of a strong second-order isomorphism between optical signals and visual perception. They further discussed how this second-order isomorphism of perceptual dissimilarities and optical dissimilarities can be exploited beneficially in the development of a visual speech synthesizer, and suggested that the perceptually relevant phonetic details in visible speech should be synthesized in order to create meaningful synthetic visual speech.

Chapter XVI "Multimodal Speaker Identification using Discriminative Lip Motion Features" describes a multimodal speaker identification system that integrates audio, lip texture, and lip motion modalities. The authors proposed a two-stage, spatial-temporal discrimination analysis framework that involves the spatial Bayesian feature selection and the temporal LDA to obtain the best lip motion feature representation for speaker identification. Two types of lip motion features, that is grid-based image motion features and lip shape features, are examined and compared. A multimodality recognition system involving audio, lip texture, and lip motion information is demonstrated to be feasible.

Chapter XVII "Lip Motion Features for Biometric Person Recognition" describes the use of lip motion as a single biometric modality as well as a modality integrated with audio speech for speaker identity recognition and digit recognition. The lip motion is modeled as the distribution of apparent line velocities in the movement of brightness patterns in an image. The authors described in detail how the lip motion features can be extracted reliably from a video sequence. Speaker recognition results based on single digit recognition using the XM2VTS database containing the video and audio data of 295 people are presented. They also described how the system can be used in a text prompted mode to verify the liveness of the user utilizing digit recognition.

Visual speech/speaker recognition is an emerging field of research that has many interesting applications in human computer interaction, security, and digital entertainment. This book provides a timely collection of latest research in this area. We believe that the chapters provide an extensive coverage of the field and would prove to be a valuable reference to current and future researchers working in this fascinating area.

Editors
Alan Wee-Chung LIEW
Shilin WANG