Chapter XVI
Multimodal Speaker Identification Using Discriminative Lip Motion Features

H. Ertan Çetingül
John Hopkins University, USA

Engin Erzin
Koç University, Turkey

Yücel Yemez
Koç University, Turkey

A. Murat Tekalp
Koç University, Turkey

ABSTRACT

This chapter presents a multimodal speaker identification system that integrates audio, lip texture, and lip motion modalities, and the authors propose to use the “explicit” lip motion information that best represent the modality for the given problem. The work is presented in two stages: First, they consider several lip motion feature candidates such as dense motion features on the lip region, motion features on the outer lip contour, and lip shape features. Meanwhile, the authors introduce their main contribution, which is a novel two-stage, spatial-temporal discrimination analysis framework designed to obtain the best lip motion features. For speaker identification, the best lip motion features result in the highest discrimination among speakers. Next, they investigate the benefits of the inclusion of the best lip motion features for multimodal recognition. Audio, lip texture, and lip motion modalities are fused...
Multimodal Speaker Identification Using Discriminative Lip Motion Features

by the reliability weighted summation (RWS) decision rule, and hidden Markov model (HMM)-based modeling is performed for both unimodal and multimodal recognition. Experimental results indicate that discriminative grid-based lip motion features are proved to be more valuable and provide additional performance gains in speaker identification.

INTRODUCTION

Audio is probably the most natural modality to recognize the speech content and a valuable source to identify a speaker. However, especially under noisy conditions, audio-only speaker/speech recognition systems are far from being perfect. Video also contains important biometric information such as face/ lip appearance, lip shape, and lip movement that is correlated with audio. Due to this correlation, it is natural to expect that speech content can be partially revealed through lip reading; and lip movement patterns also contain information about the identity of a speaker. Nevertheless, performance problems are also observed in video-only speaker/speech recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions may have detrimental effects (Turk & Pentland, 1991; Zhang, 1997). Hence robust solutions should employ multiple modalities, i.e., audio and various lip modalities, in a unified scheme.

Indeed in speaker/speech recognition, state-of-the art systems employ both audio and lip information in a unified framework (see Chen (2001) and references therein). However, most of the audio-visual biometric systems combine a simple visual modality with a sophisticated audio modality. Systems employing enhanced visual information are quite limited due to several reasons. On one hand, lip feature extraction and tracking are complex tasks, as it has been shown by few studies in the literature (see Çetingül (2006b) and references therein). On the other hand, the exploitation of this cue has been limited to the use of three alternative representations for lip information: i) lip texture, ii) lip shape (geometry), and iii) lip motion features. The first represents the lip movement implicitly along with appearance information that might sometimes carry useful discrimination information; but in some other cases the appearance may degrade the recognition performance since it is sensitive to acquisition conditions. The second, i.e., lip shape, usually requires tracking the lip contour and fitting contour model parameters and/or computing geometric features such as horizontal/vertical openings, contour perimeter, lip area, etc. This option seems as the most powerful one to model the lip movement, especially for lip reading, since it is easier to match mouth openings-closings with the corresponding phonemes. However, lip tracking and contour fitting are challenging tasks since contour tracking algorithms are sensitive to lighting conditions and image quality. The last option is the use of explicit lip motion, which are potentially easy to compute and robust to lighting variations.

Following the generation of different audio-visual modalities, the design of a multimodal recognition system requires addressing three basic issues: i) which modalities to fuse, ii) how to represent each modality with a discriminative and low-dimensional set of features, and iii) how to fuse existing modalities. For the first issue, speech content and voice can be interpreted as two different though correlated information existing in audio signals. Likewise, video signal can be split into different modalities, such as face/lip texture, lip geometry, and lip motion. The second issue, feature selection, also includes modeling of the classifiers through which each class is represented with a statistical model or a representative feature set. Curse of dimensionality, computational efficiency, robustness, invariance and discrimination capability are the most important criteria in the selection of the feature set and the recognition methodol-
Related Content

Natural Language Processing as Feature Extraction Method for Building Better Predictive Models
Goran Klepac and Marko Veli (2015). Modern Computational Models of Semantic Discovery in Natural Language (pp. 141-166).
www.igi-global.com/chapter/natural-language-processing-as-feature-extraction-method-for-building-better-predictive-models/133878?camid=4v1a

Lip Motion Features for Biometric Person Recognition
www.igi-global.com/chapter/lip-motion-features-biometric-person/31079?camid=4v1a

The Bengali Literary Collection of Rabindranath Tagore: Search and Study of Lexical Richness
www.igi-global.com/chapter/bengali-literary-collection-rabindranath-tagore/78480?camid=4v1a

Bangla Speech Analysis, Synthesis, and Vowel Nasality
www.igi-global.com/chapter/bangla-speech-analysis-synthesis-vowel/78477?camid=4v1a