Abstract

The contribution of this paper is to propose a novel approach of evaluating the performance of a noise robust audio-visual speaker identification system in challenging environment. Though the traditional HMM based audio-visual speaker identification system is very sensitive to the speech parameter variation, the proposed hybrid feature and decision fusion based audio-visual speaker identification is found to be stance and performs well for improving the robustness and naturalness of human-computer-interaction. Linear Prediction Cepstral Coefficients and Mel Frequency Cepstral Coefficients are used to extract the audio features and Active Appearance Model and Active Shape Model have been used to extract the appearance and shape based features for the facial image. Principal Component Analysis method has been used to reduce the dimensionality of large feature vector and to normalize, the vector normalization algorithm has been used. Features and decision both are fused in two different levels and finally four
different classifier outputs are combined in parallel fashion to achieve the identification result.
The performances of all these uni-modal and multi-modal system performance have been
evaluated and compared with each other on VALID audio-visual multi-modal database,
containing both vocal and visual biometric modalities.

Reference

- D. G. Stork and M. E. Hennecke, Eds., Speechreading by Humans and Machines. Berlin,
  Germany: Springer, 1996.
- R. Campbell, B. Dodd, and D. Burnham, Eds., Hearing by Eye II. Hove, United Kingdom:
- S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech
- G. Potamianos, J. Luettin, and C. Neti, “Hierarchical discriminant features for
- Reynolds, D.A., “Experimental evaluation of features for robust speaker identification,”
- Sharma, S., Ellis, D., Kajarekar, S., Jain, P. & Hermansky, H., “Feature extraction using
- C. C. Chibelushi, F. Deravi, and J. S. D. Mason, “A review of speech-based bimodal
- D. N. Zotkin, R. Duraiswami, and L. S. Davis, “Joint audio-visual tracking using particle
- P. De Cuetos, C. Neti, and A. Senior, “Audio-visual intent to speak detection for human
  June 5–9, 2000, pp. 1325–1328.
  speech sources: A new approach exploiting the audio-visual coherence of speech stimuli,”
  to exploit the audio/video correlation,” Proc. Conf. Audio-Visual Speech Processing, Terrigal,
- J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, “Integratio of multimodal features for
  video scene classification based on HMM,” in Proc. Works. Multimedia Signal Processing,
- E. Cosatto and H. P. Graf, “Photo-realistic talking-heads from image samples,” IEEE
- Gerasimos Potamianos, Chalapath Neti, and Sabine Deligne, “Joint Audio-Visual
  Speech Processing for Recognition and Enhancement,” Auditory-Visual Speech Processing
- Simon Doclo and Marc Moonen, “On the Output SNR of the Speech-Distortion Weighted
  Multichannel Wiener Filter”, IEEE SIGNAL PROCESSING LETTERS, VOL. 12, NO. 12,
  DECEMBER 2005.

Index Terms

Computer Science
Bio-informatics

Key words
Hybrid Feature and Decision Fusion
Audio-Visual Speaker Identification
Cepstral Base Audio Features
Appearance and Shape Based Facial Features
Likelihood Ratio Based Score Fusion
Discrete Hidden Markov Model