

Audio-Visual Speech Processing: Progress and Challenges

Gerasimos Potamianos

Human Language Technologies Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
Email: gpotam@us.ibm.com

Abstract

This keynote focuses on using visual channel information to improve automatic speech processing for human computer interaction. Two main issues are discussed: the extraction and representation of visual speech, as well as its fusion with traditional acoustic information. The talk mostly considers applying these techniques to automatic speech recognition, however additional areas of interest are also mentioned, for example audio-visual speech detection, enhancement, and synthesis, as well as speaker recognition. The state-of-the-art and remaining challenges in these areas are also discussed.

Keywords: Audio-Visual Speech Processing; Speech Recognition; Speech Enhancement; Speaker Recognition; Speech Synthesis.

Brief Overview of the Keynote

Speech is viewed as an integral part of human-computer interaction (HCI), conveying not only user linguistic information, but also emotion, identity, location, and computer feedback. However, although great progress has been achieved over the past decades, computer processing of speech still lags significantly compared to human performance levels. For example, automatic speech recognition (ASR) lacks robustness to channel mismatch and noise; text-to-speech (TTS) systems continue to lag in naturalness, expressiveness, and, somewhat less, in intelligibility; and typical real-life interaction scenarios, where emotion and non-acoustic cues are used to convey a message, prove insurmountably challenging to traditional computer systems that rely on the audio signal alone. In contrast, humans easily master complex communication tasks by utilizing additional channels of information, most notably the visual sensory channel.

Of central importance to human communication is the visual information present in the face, with the lower face playing an integral role in the production of human speech and of its perception, both being audio-visual in nature. This has motivated significant research over the past quarter century on automatic processing of visual speech and its integration with audio for a number of speech processing applications (Chen 2001). In particular, automatic recognition of visual speech, also known as automatic speechreading, and its fusion with audio-only systems, that gives

rise to audio-visual ASR, has attracted much of this interest (Potamianos et al. 2003). In addition, the need for improved naturalness, expressiveness, and intelligibility of synthesized speech, has steered research work towards augmenting TTS systems by synthesized visual speech (Cosatto and Graf 2001). Further, a number of recently proposed techniques utilize visual-only or joint audio-visual signal processing for speech enhancement, speech activity detection, and source localization, identity recognition from face appearance or visual speech (Chibelushi et al. 2000), and visual recognition and synthesis of human facial emotional expressions. In all cases, the visual modality can significantly improve audio-only systems.

In order to automatically process and incorporate the visual information into the above speech-based HCI technologies, a number of steps are required that are surprisingly similar across them. Central to all technologies is the feature representation of visual speech and its robust extraction. In addition, appropriate integration of the audio and visual representations is required, in order to ensure improved performance of the bimodal systems over audio-only baselines. In a number of technologies, this integration occurs by exploiting audio-visual signal correlation, whereas in others, feature or decision (classifier) fusion techniques are employed. These topics are discussed in detail in this talk, with emphasis on their application to audio-visual ASR. The current state-of-the-art in the area and what is viewed as the remaining challenges to be met are also presented.

Acknowledgements

A number of colleagues at IBM have contributed to the presented work: Stephen M. Chu, Jonathan Connell, Sabine Deligne, Norman Haas, Jing Huang, Giridharan Iyengar, Vit Libal, Etienne Marcheret, Chalapathy Neti, Hariett Nock, Larry Sansone, Andrew W. Senior, and Roberto Sicconi. Furthermore, the following have collaborated with the group through summer internships or postdoctoral fellowships: Ashutosh Garg, Roland Goecke, Guillaume Gravier, Jintao Jiang, Patrick Lucey, and Patricia Scanlon. Finally, collaboration with Petar S. Aleksic and Aggelos K. Katsaggelos (Northwestern U.) on the subject of this talk is also acknowledged.

References

- Chen, T. (2001), "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Processing Mag.*, Vol. 18, No. 1, pp. 9–21.
- Chibelushi, C.C., Deravi, F., & Mason, J.S.D. (2002), "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, Vol. 4, No. 1, pp. 23–37.
- Cosatto, E. & Graf, H.P. (2000), "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, Vol. 2, No. 3, pp. 152–163.
- Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A.W. (2003), "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, Vol. 91, No. 9, pp. 1306–1326.