COMPARISON BETWEEN DIFFERENT FEATURE EXTRACTION TECHNIQUES FOR AUDIO-VISUAL SPEECH RECOGNITION

Alin G. Chi_u[1], Leon J.M. Rothkrantz[1], Pascal Wiggers[1], Jacek C. Wojdel[2]

[1] Man-Machine Interaction Group, Delft University of Technology, The Netherlands. Emails:{A.G.Chitu; L.J.M.Rothkrantz; P.Wiggers}@ewi.tudelft.nl
[2] Quantum Chemistry of Materials Research Group, University of Barcelona, Spain. Email: J.C.Wojdel@ub.edu

Over the years much work has been done in the domain of automatic speech recognition. The progress made is significant for small, medium and even large vocabulary systems. However, the good results are only valid when the testing environment offers conditions similar to the ones present when the training database was created. The level of accuracy of the current Automatic Speech Recognition (ASR) suffers greatly when the background acoustical noise increases, namely when the Signal to Noise Ratio (SNR) decreases. This is especially the case when the system is deployed in public spaces or is used for crises situations management where the background noise is expected to be extremely large.

The video information is not affected by acoustical noise which makes it an ideal candidate for data fusion in speech recognition benefit. In the paper [1] the authors have shown that most of the techniques used for extraction of static visual features result in equivalent features or at least the most informative features exhibit this property. We argue that one of the main problems of existing methods is that the resulting features contain little or no information about the motion of the speaker's lips. Therefore, in this paper we will analyze the importance of motion detection for speech recognition. For this we will first present the Lip Geometry Estimation (LGE) method for static feature extraction. This method combines an appearance based approach with a statistical approach to accurately extract the shape of the mouth. The method was introduced in [2] and explored in detail in [3]. Further more, we introduce a second method based on a novel approach that captures the relevant motion information with respect to speech recognition by performing optical flow analysis on the contour of the speaker's mouth. For completion, a middle way approach is also analyzed. This third method considers recovering the motion information by computing the first derivatives of the static visual features.

All methods were tested and compared with a continuous speech recognizer for Dutch. The evaluation of these methods is done under different noise conditions. We show that the audio-visual recognition based on the true motion features, namely obtained by performing optical flow analysis, outperforms the other methods in low SNR conditions.

[1] L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Comparison between different feature extraction techniques in lipreading applications", in Specom'2006, SpIIRAS Petersburg, 2006.
[2] J. C. Wojdel and L. J. M. Rothkrantz, "Visually based speech onset/offset detection", in Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia 2000), (Antwerp, Belgium), pp. 156–160, 2000.
[3] L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Fusing Data Streams in Continuous Audio-Visual Speech Recognition", in Text, Speech and Dialogue: 8th International Conference, TSD 2005, vol. 3658, (Karlovy Vary, Czech Republic), pp. 33–44, Springer Berlin / Heidelberg, September 2005.