

# Audio-Visual Speech Recognition and Synthesis

László Czap

## Summary

Automatic speechreading systems through their use of visual information to support the acoustic signal have been shown to yield better recognition performance than purely acoustic systems, especially when background noise is present. In this thesis an answer is sought to the most important questions of speechreading: Which features can represent visual information well? How can they be extracted? An intelligibility study was carried out to see which parts of the face give the most support to speechreading. The whole face, mouth or lips were visible dubbed with noisy voice. Visual support to speech perception of the image ellipse model is compared to that of the parts of the natural face.

It is generally agreed that most visual information is carried by the lips. The inner lips are especially important and remarkable improvement comes from the visibility of teeth and tongue. Geometric features and the intensity factor of the oral cavity are discussed as a means of visual speech representation.

Much of the research in speechreading systems is focused on the crucial problem of feature extraction. How can it best transform a sequence of images into feature values that facilitate recognition? The process should be fast, robust, and yield as much information as possible carried by the fewest number of features, removing redundant and linguistically irrelevant information. Whereas there is no one favourite way of representing visual speech, there are impressive methods that all require tracking the inner and outer contours of the lips. A novel feature extraction method based on image similarity study is proposed that does not need tracking of the lips. Images for an articulation library have been selected iteratively from characteristic lips shapes and visibility of teeth and tongue.

The proposed methods were evaluated for speech recognition on a 600 word audio-visual database of one subject, and on a 9400 word acoustic database of the same subject. Efficiency of the geometric and pixel based features are compared on a continuous speech recognition task. Pixel based features can represent the visual speech better than the geometric ones.

Semi-syllables and diphones were compared as subword candidates for basic linguistic elements of automatic speech recognition for an agglutinating language. The diphone based recognition highly outperformed the semi syllable based one on the audio-visual and the acoustic database as well. The conclusion is that context sensitive elements can yield better performance.

Based on the analysis of articulation a talking head has been developed. Classification of visemes – the visual equivalents of phonemes – was carried out for Hungarian vowels and consonants, yielding key frames for interpolation. A major difficulty in operating a 3D head model is to control the dynamic features of articulation. A three level dominance model has been proposed. Each feature is considered as dominant, flexible or uncertain one.