Speechreading

László Czap

University of Miskolc, Department of Automation
3515 Miskolc, Egyetemváros
czap@mazsola.iit.uni-miskolc.hu

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 paper 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 a 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 favorite 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 a similarity study is proposed that does not need tracking of the lips.

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