Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

fast lip tracking for speech/nonspeech detection

Authors: Stefan Schacht, Oleg Fallman, Dietrich Klakow
spoken language systems
Saarland university
66041 Saarbrücken

An efficient speech/nonspeech detection is an important part of any speech recognition system. It allows a good estimation of the background noise, which can be used for noise cancellation techniques like spectral subtraction. Furthermore it avoids the activity of the speech recognizer on unwanted segments of the audio stream. Recently speech recognition has gained popularity on mobile devices, e.g. smartphones. These devices are often used in crowded environments. So the unwanted audio segments often contain speech. This makes a decision, which is based only on audio data, difficult. Since most modern smartphones have a front facing camera we propose

a speech/nonspeech detection algorithm that also uses video informations. First we use a modified version of the well known Viola Jones algorithm for face detection to locate the users face in the video stream. Our modifications exploit the typical usage of a mobile device. First we search only for one face in the video stream. The normal distance from the users face to the camera also limits the size of the frame segment which might be classified as a face. If the previous video frame contained a face we start the search in the current frame at the same location. After the localization of the users face we use linear discriminant analysis and polynomial approximation to track the contour of the lips. We combine these informations with the analysis of the audio stream to obtain a robust speech/nonspeech decision. We present an example implementation of this algorithm for a mobile device and evaluation results on a large collection of videos, which were recorded during the SmartWeb project.