Putting ‘Visual’ in Audio-Visual Speech Recognition System

Alex Zhang, Michal Zak

Abstract

The presentation covers the methodology and preliminary results for our audio-visual speech recognition system. The initial approach based solely on processing audio features using cepstral coefficients proved to be unreliable in low Signal-to-Noise-Ratio conditions. We addressed this issue by processing visual lip features in addition to audio features.

Since there is no industry standard way of extracting visual lip features, we came up with our own implementation utilising marker tracking. This was done by drawing green dots around the mouth region and performing image processing to determine the best position of the markers. Since not all markers are present all the time due to various video artefacts, we came up with techniques to reconstruct them.

Having a lip shape mask enabled us to extract several key features that are distinct to how speech was produced. The difficult task was to ascertain which features are actually useful and how to combine them together to achieve satisfying accuracy. Throughout the process we discovered the importance of matching lighting conditions across the data sets and came across unexpected test results related to the quirks of the HTK system.

Our final solution utilized a combined stream of features including the amount of teeth pixels visible and the low frequency representation of the isolated mouth region.