Review of the bachelor's thesis of Hana Šarbortová:

Audio-Visual Speech Activity Detector

Reviewed by Xavier Alameda-Pineda, PhD, on January the 20th, 2015.

Hana Šarbortová presents her bachelor's thesis with the title *Audio-Visual Speech Activity Detector*. In the submitted report, the problem of detecting the active speaker is addressed within the non-parametric framework of Canonical Correlation Analysis (CCA). Upon examination of recent literature on the topic, the choice is motivated by three main reasons: robustness to noise, ability to fuse features of different dimensionality and automatic computation of the statistical resemblance, i.e., the correlation. Traditional well-known auditory features and geometrical visual features are used, together with their derivatives, to provide a dynamic description of the auditory and visual information. The auditory and visual features are fused later on by the aforementioned CCA methodology, which in addition computes a statistical resemblance or correlation. The experimental use of this correlation is two-fold: on the one side to classify the sequences between synchronous and asynchronous and on the other side to estimate the delay between the audio waveform and the image flow.

Hana Šarbortová wrote a clear exposition of the state-of-the-art. Even if a couple of points are missing, it is impressive how the closest research studies are described. Indeed the amount of literature on audio-visual signal processing addressing problems related to the detection of speaker is vast. Moreover, the relations, dependencies and differences between the related tasks (diarization, video indexing, tracking, sound source localization and audio-visual fusion, to cite a few) are extremely complex. Therefore, navigating through the existing literature and stating the prominent features, limitations and differences between methods is, in my opinion, a matter of study at a Master's level. Despite all the complexity and the challenges associated to the field of audio-visual speaker detection, the exposition of the state-of-the-art is clear and well-structured.

Following the intuition of traditional speech processing text-books and well-known visual descriptors, the designed system relies on Mel-Frequency Cepstral Coefficients (MFCC) and on normalized cues of the mouth aperture. These features, together with their first and second order derivatives, form the auditory and visual descriptors respectively. The procedures to extract MFCC from the auditory waveform and mouth aperture features (and their normalization using projective geometry) from the images is fully detailed. Likewise, the framework of CCA is explained providing insights on the advantages and disadvantages of the chosen methodology.

An extensive set of experiments is presented and discussed. First, the lip-only speech activity detector is evaluated. The dynamics of the lips' features serve to classify each visual frame as speaking or not speaking. More precisely, the mean of the standard deviation of the derivatives is used. The effect of the length of the two sliding windows (for the mean and the standard deviation) is tested, providing key insights for the next experimental protocol on detecting audio-visual synchrony from CCA. One of the prominent features of the approach described in this thesis is that it is independent of the sequence. That is to say that the CCA projection can be computed on one (training) sequence and applied in another (testing) sequence. Figure 13 of the report shows the performance in these precise conditions for different values of the integrating window. Even if it is true that the projections are still valid when changing the sequence and thus the approach generalizes well, it is also true (as shown in the same figure) that the window choice is crucial and can have a very negative effect if the length is not appropriate. Interestingly, figure 15 shows the histogram of the correlation value for synchronous and asynchronous sequences and for different values of the window length. This proves the intuition given before, since the histograms are clearly separated for higher values of the integrating window. Precision-recall curves also point to this direction. All these encouraging results motivate a final and more complex experiment with the Cardiff conversational data set. The aim of the experiment is two-fold: classifying synchronous and asynchronous sequences and estimating the asynchrony between audio and video. While the latter is extremely satisfying, the former shows some limitation of the proposed approach. Indeed, in some cases the video-only detector outperforms the proposed audio-visual detector. However, in most cases, the theoretical limit performance of the CCA-based approach (i.e., using ground truth) is far better than the video-only detector. These are fantastic news, that, together with the reported results of sections 4.1 and 4.2 encourage to pursue the research in this direction.

Having read the manuscript submitted and at the light of the results, the quality of writing and the ability of Hana Šarbortová to analyze the state-of-the-art, discuss the obtained results and synthesize adequate conclusions and enthusiastic future work guidelines, I strongly recommend to mark the aforementioned bachelor's thesis with **A-excellent**. I am available for any further comments or questions the evaluation committee may have about this review.

Yours sincerely,
Xavier Alameda-Pineda, PhD.
Dipartimento di Ingegneria e Scienza dell'Informazione
Via Sommarive, 5 - 38123 Povo, Trento, Italy