Obtaining Person-independent Feature Space for Lip-reading

Jacek C. Wojdeł and Leon J. M. Rothkrantz
Knowledge Based Systems,
Delft University of Technology

Abstract

Obtaining a person-independent representation of the lip-movements is crucial in developing a multi-modal speech recognizer. The geometrical models used in most lip-tracking techniques can remove some of the features such as skin texture or color, but even then the resulting data remains highly person-specific. It is certainly beneficial in multi-modal person identification systems, but undesirable in case of speech recognizers. Appropriate normalization of the data and it's projection in the principal components space can reduce the amount of person-specific features even further.

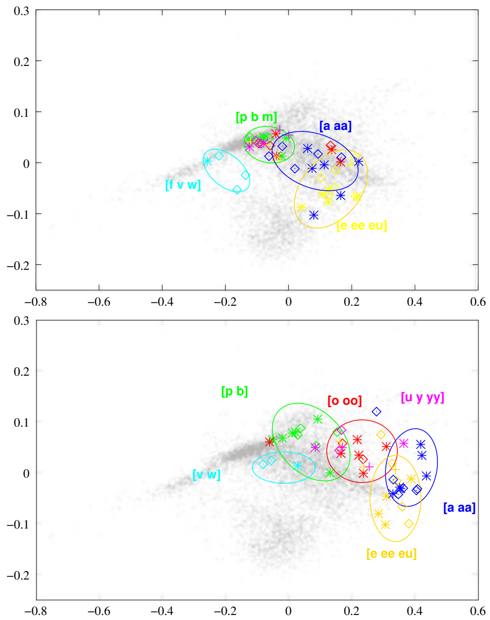
During our research we've concluded that although using Principal Component Analysis (PCA) of the multi-person dataset reveals some interesting features, the inter-person variation is too big to allow for robust speech recognition (see Fig. 1). There are however substantial similarities in the lip-shape variations when analyzing only a single-person data sets. For example, the first PC seems always to relate to the stretching of the mouth, while the second PC varies with the opening of the mouth. Starting from this observation we propose to use an adaptive PCA that updates the projection coefficients with respect to the data available for the specific person.

We use a linear combination of the projection acquired from a large multi-person dataset with the projection computed on the basis of the specific person's recordings. The weight factor of the generic projection diminishes with the amount of person-specific data gathered. Our experiments show that the projection matrix stabilizes after roughly 90 seconds of recordings, at which point, the generic projection matrix can be discarded altogether.

In our experiments we used a technique that estimates the visible width of the lip and its distance from the center of the mouth at different angles (see [2, 3]. The measurements were then normalized using the width of the bounding box, so that the obtained data does not depend on the distance between the subject and the camera. A PCA is done on the data coming from each person separately. Although the resulting projections are different, the clusters corresponding to the phonemes appear in the same place of the target space (see Fig. 2).

Adaptive person specific PCA is not constrained to our specific model. The same approach can be used to process the data obtained using other lip-tracking techniques. As long as the proper normalization measures are taken, we may expect the primary components of the extracted data to relate to the corresponding changes in the lip-shape.

Using a person specific, adaptive projection into the feature space has both advantages and disadvantages. One of the positive things is that it allows for more robust lip-reading using a person independent recognition system. The amount of data (recordings) that is needed to obtain a correct projection is much smaller than the amount of data needed to adapt the system itself. At the same time however, for some applications the adaptation time might be unacceptable. Over one minute of speech data that is needed to stabilize the PCA is usually longer that the whole user interaction at some information booth. It may also be expected that in case of a long-term person specific applications (such as dictation systems) it might be more beneficial to retrain the recognition system rather than only a feature-space projection. The proposed approach should therefore be used in cases where human-computer interaction for a specific person is expected to be relatively long but not permanent.



(a)

(b)

Figure 1: Several phonemes mapped on a 2D plane using PCA for the whole dataset (grey points). The clustering of the phoneme groups (colored points) can be seen for both subjects (a and b). There is however not much correspondence between subjects.

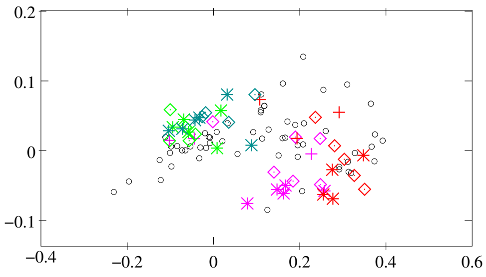


Figure 2: The same phonemes mapped using different, person-specific PCAs. The phoneme frames map to the same region of the 2D space. Two phoneme groups for two different subjects are highlighted: green – *[p b m]*, purple/red – *[e ee eu]*

References

[1] Jacek C. Wojdel and Leon J. M. Rothkrantz. Robust video processing for lipreading applications. In *EUROMEDIA'2001*, pages 195–199, Valencia, Spain, 2001.

[2] Jacek C. Wojdel and Leon J. M. Rothkrantz. Using aerial and geometric features in automatic lip-reading. In *Proceedings of Eurospeech 2001 – Scandinavia*, September 2001.