Appears in *Interspeech*, pp 702–705, Antwerp, Belgium 2007.

Audio-visual classification of Swedish phonemes for pronunciation training

Hedvig Kjellström [1], *Olov Engwall* [2], *Sherif Abdou* [3], *Olle Bälter* [4]

[1] Computational Vision and Active Perception Laboratory, CSC, KTH, Stockholm, Sweden
[2] Centre for Speech Technology (CTT), CSC, KTH, Stockholm, Sweden
[3] Department of IT, Faculty of Computers and Information, Cairo University, Giza, Egypt
[4] Human-Computer Interaction Group, CSC, KTH, Stockholm, Sweden
{hedvig,engwall,balter}@kth.se, s.abdou@fci-cu.edu.eg

Abstract

We present a method for audio-visual classification of Swedish phonemes, to be used in computer-assisted pronunciation training. The probabilistic kernel-based method is applied to the audio signal and/or either a principal or an independent component (PCA or ICA) representation of the mouth region in video images. We investigate which representation (PCA or ICA) that may be most suitable and the number of components required in the base, in order to be able to automatically detect pronunciation errors in Swedish from audio-visual input. Experiments performed on one speaker show that the visual information help avoiding classification errors that would lead to gravely erroneous feedback to the user; that it is better to perform phoneme classification on audio and video seperately and then fuse the results, rather than combining them before classification; and that PCA outperforms ICA for few components.

Index Terms: audiovisual phoneme classification, pronunciation error detection, PCA, ICA

1. Introduction

The needs and potential for Computer Assisted Pronunciation Training (CAPT) are great, but if a breakthrough is to be achieved, the CAPT systems must become much more apt than currently at helping the user correct the error, rather than merely pointing out that something is wrong. We are developing a computer-animated articulation tutor, ARTUR [1], who should assist hearing- or language-impaired children and second language learners with their pronunciation of Swedish. The aim is to detect pronunciation errors and give audio-visual help on how they may be corrected. In order to achieve this, the system must firstly have a knowledge about the important features of each phoneme, and, secondly, gain information about how the user produced it.

Swedish vowel roundedness is particularly difficult for foreign speakers. Not only do rounded front vowels occur (as in e.g., French, but contrary to most other languages), but it is also one of the few languages that has two types of phonemic distinct rounded vowel types: endolabial (or compressed) and exolabial (or protruded), in addition to the unrounded vowels. This leads to frequent mispronunciations dependent on the speaker's first language, such as the near-close, near-front compressed vowel [ʉ̟ː] being pronounced as [uː] or [yː], and [ʏ] as [ɪ] or [ʊ]. Among the consonants, at least the palatovelar fricative [ɧ] is troublesome and often mispronounced as one of [ʃ, x, χ].

Automatic detection of such pronunciation errors is difficult, but the task becomes more feasible if video data is added, since the phonemes are acoustically close but visually distinct. In this paper, we hence investigate audio-visual phoneme classification for pronunciation training. This signifies that the method should be able to identify deviations in contrastive features, such as vowel roundedness and fricative place of articulation.

2. Audio-visual phoneme recognition

The most common approach to visual speech recognition is to track or extract the lip contours, which are modeled using snakes [2, 3] or data-driven PCA methods [4, 5]. This approach is successful because much of the articulatory information is present in the lip shape, but some information, such as the visibility of the tongue tip, and shadows above and below the mouth indicating lip protrusion, is lost in this type of representation.

Since lip protrusion is of particular interest in Swedish, we instead track the upper part of the face, extract the mouth region in the stabilized image and represent its articulatory information implicitly in terms of image pixel values [6].

The advantages of tracking the face and use all pixel values, rather than lip contours, are that information about both horizontal and vertical lip movements relative to the face is preserved, and that it is easier and less computationally demanding to track the face robustly [2, 7], since it is less deformable than the lips.

From a set of training images of the mouth, basis functions representing the most prominent variations were learned with principal or independent component analysis (PCA or ICA).

To test the quality of the PCA and ICA representations for pronunciation error detection purposes, we perform audio-visual classification of Swedish phonemes from the speech signal and/or lip images, represented by principal or independent components (PC or IC). We in particular investigate the number of PC and IC required to classify vowel roundedness correctly.

3. Data Acquisition

A video of the face of a female speaker of Swedish was recorded together with the audio signal. Simultaneously, registrations were made of the 2D positions of electromagnetic articulography coils on the tongue and jaw and the 3D positions of 28 infra-red reflectors on the face [8], but that data was not used in this study.

The corpus consisted of 37 asymmetric CVC words for the vowel classification, and 63 symmetric VCV words for the consonant classification. The C_1VC_2 words were combinations of C_1_C_2=[p_k, k_p, k_r] and V=[iː, ɪ, eː, e, ɛː, ɛ, æː, æ, yː, ʏ, ʉ̟ː, ɵ, øː, ø, œ, œː, oː, ɔ, uː, ʊ, a, ɑː], and the VCV words were combinations of V=[ɪ, ʊ, a] and C=[p, t, k, b, d, ɡ, f, s, ɕ, ɧ,

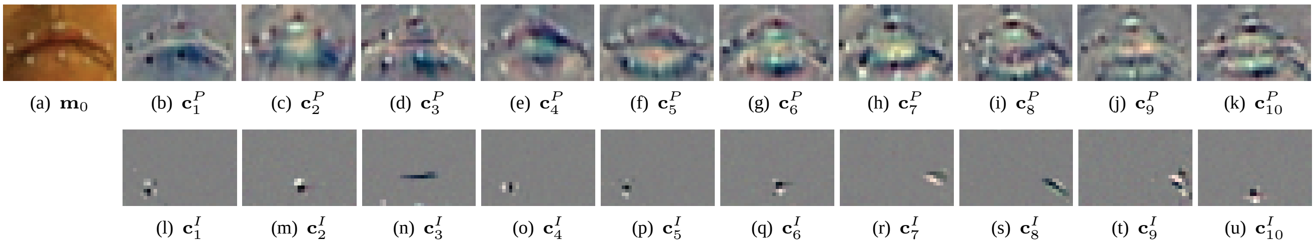


Figure 1: (a) Template $\mathbf{m}_0$. (b-k) The first 10 principal components $\mathbf{c}_{1-10}^P$. (l-u) The first 10 independent components $\mathbf{c}_{1-10}^I$.

m, n, ŋ, l, r, ɳ, ʈ, ɖ, v, j]. Each word appeared once in the set.

4. Data Processing

4.1. Video Data

The video frame-rate was 25 Hz and the image size 768 × 576 pixels. After image stabilization, a 33 × 23 pixel image of the mouth region was extracted in the frames representing the central phoneme in each CVC and VCV word, identified using an HMM-based forced-alignment.

A low-dimensional representation of this region was learned with PCA or ICA. The task of the component analysis is to select a base $C = [\mathbf{c}_1, \dots, \mathbf{c}_n]$ that represents the data set $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ as well as possible, using $\mathbf{x}_k = \sum_{i=1}^n v_{i,k}\mathbf{c}_i$. $V = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ is a parameter matrix in the subspace C. In this case, $\mathbf{x}_k = \mathbf{m}_k - \mathbf{m}_0$ is the difference (the R, G, B bands substracted seperately) between image $\mathbf{m}_k$ and a template image $\mathbf{m}_0$ with neutral lip pose, Fig. 1(a). Any new lip image can be approximated as a linear combination of $\mathbf{m}_0$ and the components in the base.

Using PCA, C is selected so that the columns represent the n largest principal components of the data set, c.f. Fig. 1(b-k), while ICA instead selects the n most informative statistically independent components, c.f. Fig. 1(l-u).

The reflective markers that the subject wore to allow for 3D motion capture [8] were not used in this study, neither in the stabilization nor in the learning, but they nevertheless clearly affected the ICA base, Fig. 1(l-u). The effect of the markers was discussed in [9], concluding that they do not improve the results, and possibly even worsen them. This is in accordance with the reconstructed images in Figs. 2(e-g), where the markers in the original, Fig. 2(a), are reconstructed in the wrong positions or not at all. We have further tested ICA of an unmarked face, with similar results for the appearance of the reconstructed images. Since the method relies on a holistic representation of the mouth pattern, rather than tracking of individual features, the presence of the markers are not crucial. We therefore consider that the results presented below are comparable to what may be achieved for an unmarked face.

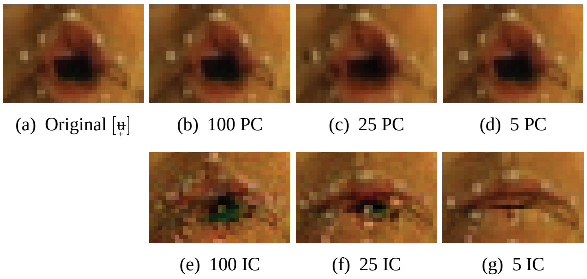


Figure 2: (a) Original frame. (b-d) PCA reconstruction of the same frame. (e-h) ICA reconstruction of the same frame.

4.2. Speech Signal

The audio signal was originally sampled at 16 kHz, but was divided into frames of length 57.6 ms with a shift of 40 ms to correspond to the video frame rate. Each acoustic frame was pre-emhasized and multiplied by a Hamming window, before applying a covariance-based LPC algorithm [10] to generate 16 line spectrum pairs (LSP). The acoustic data hence consisted of vectors $\mathbf{a}_k$ with the 16 LSP coefficients and the RMS amplitude in the frames representing the central phoneme in each word.

5. Classification training

The phoneme classification was evaluated on separate frames, without any contextual information, vocabulary or grammar defined, using the acoustic and video data. In order to analyze the effect of the different image representations (ICA or PCA and number of components) in more detail, a viseme classification was also performed, using the video data only.

A jackknife procedure was employed for training and testing. The data for each phoneme or viseme was divided into four equally large parts. One part in turn was removed from the training data and used as test set, while the three others and the data from all other classes constituted the training set. The classification result for each class was then averaged over the four permutations.

Consonants and vowels were trained separately, using the VCV words for consonants and the C$_1$VC$_2$ words for vowels.

The phoneme category π_k is estimated from the acoustic $\mathbf{a}_k$ and video $\mathbf{v}_k$ signals using a probabilistic maximum likelihood classifier, as $\pi_k = \mathrm{argmax}_\pi p(\mathbf{a}_k, \mathbf{v}_k \,|\, \pi)$. The functions $p(\mathbf{a}_k, \mathbf{v}_k \,|\, \pi)$ are kernel based [11] and describe the likelihood of observing $\mathbf{a}_k$ and $\mathbf{v}_k$ given that the speaker uttered phoneme π. In general terms, $p(\mathbf{t}_k \,|\, \pi) = \frac{1}{m}\sum_{i=1}^m N(\mathbf{t}_k \,|\, \mathbf{t}_i^\pi, \sigma_0)$ where the vectors $[\mathbf{t}_1^\pi, \dots, \mathbf{t}_m^\pi]$ are the training examples for phoneme π and $N(\cdot \,|\, \mathbf{t}, \sigma)$ is a Gaussian with mean $\mathbf{t}$ and stddev σ. In the case of audio-visual phoneme classification, $\mathbf{t}_k$ includes both $\mathbf{a}_k$ and $\mathbf{v}_k$, combined using either early or late fusion.

In early fusion training vectors are concatenating before classification as $p(\mathbf{a}_k, \mathbf{v}_k \,|\, \pi) = \frac{1}{m}\sum_{i=1}^m N([\alpha(\mathbf{a}_k)^T \ (\mathbf{v}_k)^T]^T \,|\, [\alpha(\mathbf{a}_i^\pi)^T \ (\mathbf{v}_i^\pi)^T]^T, \sigma_0)$ where $\alpha = \frac{\bar{\sigma}^V}{\bar{\sigma}^A}$ is a normalizing scale factor, $\bar{\sigma}^A$ and $\bar{\sigma}^V$ being the mean standard deviations in the audio and video datasets, and $\sigma_0 = 0.03\sqrt{n^A}\bar{\sigma}^A + 0.15\sqrt{n^V}\bar{\sigma}^V$. The scale factors in σ_0 are chosen empirically to maximize classification results. Since σ_0 is selected to maximize the performance on *test data*, and test and training sets are non-overlapping, overlearning is avoided.

With late fusion, classification is performed seperately for the two modalities and the results are combined, assuming that the data from the two modalities are statistically independent, as $p(\mathbf{a}_k, \mathbf{v}_k \,|\, \pi) = p(\mathbf{a}_k \,|\, \pi)\ p(\mathbf{v}_k \,|\, \pi) =$ $\frac{1}{m} \sum_{i=1}^{m} N(\mathbf{a}_k \,|\, \mathbf{a}_i^{\pi}, \sigma_0)\ \frac{1}{m} \sum_{j=1}^{m} N(\mathbf{v}_k \,|\, \mathbf{v}_j^{\pi}, \sigma_0)$.

The viseme classification is performed similarly. The projection $\mathbf{v}_k$ of the image $\mathbf{m}_k$ is classified into viseme category ϕ_k as $\phi_k = \mathrm{argmax}_{\phi} p(\mathbf{v}_k \,|\, \phi)$ where $p(\mathbf{v}_k \,|\, \phi)$ is the likelihood of observing $\mathbf{v}_k$ given viseme ϕ. σ_0 is set empirically to $\sigma_0 = 0.2\sqrt{n}\bar{\sigma}$.

The Swedish viseme classes [12] are bilabial [p, b, m], labiodental [f, v], alveodental [t, d, n, r, s, l], palatal [ɕ, j, ɧ], and velar [k, ɡ, ŋ] consonants, and front unrounded [iː, ɪ, eː, e, ɛː, ɛ, æː, æ], front rounded [yː, ʏ, ʉ̟ː, ɵ, øː, ø, œ, œː], back unrounded [ɑː, a] and back rounded [oː, ɔ, uː, ʊ] vowels. [ɵ] and [a] are quite central, but are grouped with the corresponding long vowels.

6. Results

The phoneme recognition was made from audio alone, from video alone and from audio-visual input, combining the two sources either before classification or after separate classifications (c.f. Section 5). The overall classification rates in Fig. 3 are similar within each condition for ICA or PCA, if enough components are used. 25 components are enough for the PCA base, but more are needed for ICA to maintain the video only classifiaction rate.

While the early fusion of audio and visual data is only marginally better than visual alone data (for >10 PC, >25 IC), late fusion results in a substantially higher correct classification rate. Interestingly, this coincides with theories on human speech perception (e.g., [13]) stating that information is processed within each modality independently and then fused. When the confusion matrices are considered, it becomes clear that the early fusion gives rise to unreasonable confusions, considering that visual information is available, with unrounded vowels misclassified as rounded and vice versa, e.g., [iː]↔[ʉ̟ː] and [yː]→[æː]. Late fusion is hence better at taking information from both modalities into account, most notably for [f], [s] and [l].

The two modalities contribute in varying degrees to the recognition of different phonemes. The audio signal is more important for [p, m, t, d, ɡ], which is natural, since the phonemes within each of the bilabial, alveolar and velar groups are identical or extremely similar in the visual input, but quite distinct in the audio, because of the difference in voicing or nasality. For video only, there are indeed frequent confusions within the bilabial and labiodental viseme groups. The video is, on the other hand, better at separating [f] and [s] and identifying [iː] and [yː]. The complementary information is most important for [eː, œː, æː, l, j], with clearly better results for the audiovisual case.

For the use as a detector of mispronunced features of Swedish, the main benefit of adding visual input is that many misclassifications between unrounded and rounded vowels and between significantly different places of consonant articulation can be avoided. Errors such as [eː, æː]→[œː], [uː]→[iː], [yː]→[æː], [p]↔[k], [v, l]→[ɧ] or [ŋ]→[j] are eliminated, or drastically reduced. In cases when the audio-visual classification still makes errors, they tend to be less serious, when the classification is used to generate articulatory feedback, e.g. confusions [yː]↔[œː] instead of with [æː]; [uː] classified as [ʉ̟ː] rather than [iː]; [f] as [v] rather than [p, ɕ, ɧ]; or [b] as [p] rather than [s].

Such errors should certainly be avoided as far as possible, but the consequences for the feedback instructions are less serious since vowel roundedness or consonant place of articulation are correct. The overall classification rate of 74% may therefore be acceptable for pronunciation training purposes.

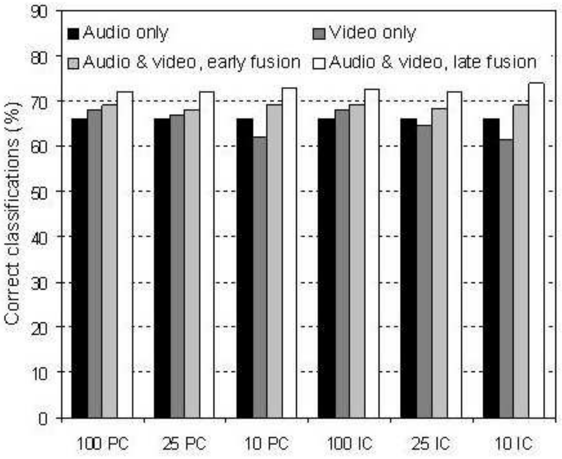


Figure 3: Phoneme classification rates using PCA or ICA.

The overall viseme classification accuracy from video data only (100 IC, PC), 80% for consonants (Figs. 4a-h) and 87% for vowels (Figs. 4a-h), outperforms human speechreaders (68% for consonants and 79% for vowels in [12]).

As already indicated by Fig. 3, more than 25 IC are needed, or else vowel groups are confused and labiodentals are classified as bilabials. At 10 IC, even velars and palatals are classified as bilabials. The cause of the confusion between vowel visemes is illustrated in Fig. 2. When too few IC are used, the reconstructed image becomes quite neutral and the lip rounding of [ʉ̟ː] in the original image disappears.

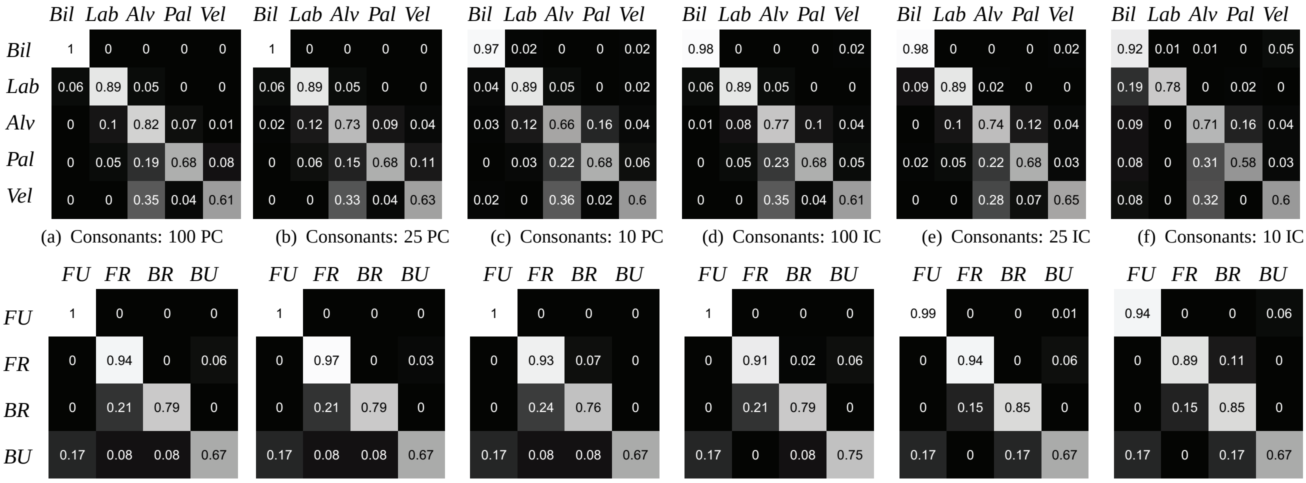
For the PCA it is less crucial how many components that are used, the classification score remains more or less the same, except for alveolars and rounded vowels. When only 10 PC are used, the alveolars are more commonly classified as palatals, possibly because the visibility of the tongue tip is not reconstructed.

The relatively high viseme classification score for palatals and velars is promising for mispronunciation detection of, e.g., [ɧ], which is problematic for many foreign speakers of Swedish.

7. Discussion & Conclusions

Our experiments with a probabilistic phoneme classification have shown that the addition of visual input avoids many of the misclassifications between unrounded and rounded vowels and between consonants with very different places of articulation that are made if acoustic only data is used. Audio-visual error detection is hence very appealing for CAPT, since erroneous feedback on lip rounding or consonant place of articulation may be avoided.

The performance was similar for PCA and ICA, with PCA being better for few (25 or less) components. In previous studies on visual speech recognition and face expression recognition [6, 14], ICA has outperformed PCA. One reason for the weak performance of the ICA, compared to previous studies, is that the lip rounding is better reconstructed with PCA. As vowel roundedness is more important in Swedish than in many other languages, it is essential to correctly represent lip rounding. The white markers may also have affected the bases, but preliminary tests with an unmarked face indicate that they are not the main



	Bil	*Lab*	*Alv*	*Pal*	*Vel*
Bil	1	0	0	0	0
Lab	0.06	0.89	0.05	0	0
Alv	0	0.1	0.82	0.07	0.01
Pal	0	0.05	0.19	0.68	0.08
Vel	0	0	0.35	0.04	0.61

(a) Consonants: 100 PC

	Bil	*Lab*	*Alv*	*Pal*	*Vel*
Bil	1	0	0	0	0
Lab	0.06	0.89	0.05	0	0
Alv	0.02	0.12	0.73	0.09	0.04
Pal	0	0.06	0.15	0.68	0.11
Vel	0	0	0.33	0.04	0.63

(b) Consonants: 25 PC

	Bil	*Lab*	*Alv*	*Pal*	*Vel*
Bil	0.97	0.02	0	0	0.02
Lab	0.04	0.89	0.05	0	0.02
Alv	0.03	0.12	0.66	0.16	0.04
Pal	0	0.03	0.22	0.68	0.06
Vel	0.02	0	0.36	0.02	0.6

(c) Consonants: 10 PC

	Bil	*Lab*	*Alv*	*Pal*	*Vel*
Bil	0.98	0	0	0	0.02
Lab	0.06	0.89	0.05	0	0
Alv	0.01	0.08	0.77	0.1	0.04
Pal	0	0.05	0.23	0.68	0.05
Vel	0	0	0.35	0.04	0.61

(d) Consonants: 100 IC

	Bil	*Lab*	*Alv*	*Pal*	*Vel*
Bil	0.98	0	0	0	0.02
Lab	0.09	0.89	0.02	0	0
Alv	0	0.1	0.74	0.12	0.04
Pal	0.02	0.05	0.22	0.68	0.03
Vel	0	0	0.28	0.07	0.65

(e) Consonants: 25 IC

	Bil	*Lab*	*Alv*	*Pal*	*Vel*
Bil	0.92	0.01	0.01	0	0.05
Lab	0.19	0.78	0	0.02	0
Alv	0.09	0	0.71	0.16	0.04
Pal	0.08	0	0.31	0.58	0.03
Vel	0.08	0	0.32	0	0.6

(f) Consonants: 10 IC

	FU	*FR*	*BR*	*BU*
FU	1	0	0	0
FR	0	0.94	0	0.06
BR	0	0.21	0.79	0
BU	0.17	0.08	0.08	0.67

(g) Vowels: 100 PC

	FU	*FR*	*BR*	*BU*
FU	1	0	0	0
FR	0	0.97	0	0.03
BR	0	0.21	0.79	0
BU	0.17	0.08	0.08	0.67

(h) Vowels: 25 PC

	FU	*FR*	*BR*	*BU*
FU	1	0	0	0
FR	0	0.93	0.07	0
BR	0	0.24	0.76	0
BU	0.17	0.08	0.08	0.67

(i) Vowels: 10 PC

	FU	*FR*	*BR*	*BU*
FU	1	0	0	0
FR	0	0.91	0.02	0.06
BR	0	0.21	0.79	0
BU	0.17	0	0.08	0.75

(j) Vowels: 100 IC

	FU	*FR*	*BR*	*BU*
FU	0.99	0	0	0.01
FR	0	0.94	0	0.06
BR	0	0.15	0.85	0
BU	0.17	0	0.17	0.67

(k) Vowels: 25 IC

	FU	*FR*	*BR*	*BU*
FU	0.94	0	0	0.06
FR	0	0.89	0.11	0
BR	0	0.15	0.85	0
BU	0.17	0	0.17	0.67

(l) Vowels: 10 IC

Figure 4: Confusion matrices for viseme classification of (a)-(f) consonants and (g)-(l) vowels using either PCA (a)-(c), (g)-(i) or ICA (d)-(f), (j)-(l) for different numbers of PC and IC. Rows represent the true viseme class i, columns the assigned class j, and the numbers in element (i, j) the fraction of visemes i classified as j. The classes are *Bil*: [p, b, m], *Lab*: [f, v], *Alv*: [t, d, n, r, s, l], *Pal*: [ɕ, j, ɧ], *Vel*: [k, g, ŋ], *FU*: [iː, ɪ, eː, e, ɛː, ɛ, æː, æ], *FR*: [yː, ʏ, ʉ̟ː, ɵ, øː, ø, œ, œː], *BR*: [oː, ɔ, uː, ʊ], *BU*: [a, ɑː].

problem. Either ICA with more components or PCA should hence be used for audio-visual detection of Swedish pronunciation errors.

Further, late fusion of separate acoustic and visual classifications should be used rather than early in order to take advantages of the complementarity of the modalities.

Much work remains before the method can be successfully incorporated in pronunciation training. The most crucial is to achieve speaker independence or adaptation, so that the detection can be used for any speaker, without previous training on that speaker. It is then probable that ICA will be the better option, since ICA seems to be able to separate variations due to speaker identity from those due to articulation [14], if learned from a training set with several speakers. Alternatively, speaker adaptation may be achieved by mapping a neutral template image of the reference speaker onto the target speaker and recalculate the components based on the template differences. It then remains to be shown that the components extracted from one speaker are able to describe the articulations of another speaker after adaptation, despite inter-speaker variability.

8. Acknowledgements

This research is performed within the two projects ARTUR – the ARticulation TUtoR, funded by the Swedish Research Council, and ADEPT – Audiovisual Detection of Errors in Pronunciation Training, a collaboration within the Swedish Research Links program, funded by the Swedish Research Council and the Swedish International Development Cooperation Agency.

9. References

[1] O. Engwall, O. Bälter, A-M. Öster, and H. Kjellström, “Designing the human-machine interface of the computer-based speech training system ARTUR based on early user tests,” *Behaviour and Information Technology*, vol. 25, pp. 353–365, 2006.

[2] R. Kaucic and A. Blake, “Accurate, real-time, unadorned lip tracking,” in *ICCV*, 1998.

[3] R. Seymour, J. Ming, and D. Stewart, “A new posterior based audio-visual integration method for robust speech recognition,” in *Interspeech*, 2005, pp. 1229–1232.

[4] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[5] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” *PAMI*, vol. 24, no. 2, pp. 198–213, 2002.

[6] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Classifying facial actions,” *PAMI*, vol. 21, no. 10, pp. 974–989, 1999.

[7] I. Shdaifat and R-R. Grigat, “A system for audio-visual speech recognition,” in *Interspeech*, 2005.

[8] J. Beskow, O. Engwall, and B. Granström, “Resynthesis of facial and intraoral motion from simultaneous measurements,” in *ICPhS*, 2003.

[9] H. Kjellström, O. Engwall, and O. Bälter, “Reconstructing tongue movements from audio and video,” in *Interspeech*, 2006, pp. 2238–2241.

[10] N. Sugamura and F. Itakura, “Speech analysis and synthesis methods developed at ECL in NTT,” *Speech Communication*, vol. 5, no. 2, pp. 199–215, 1986.

[11] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.

[12] J. Marthony, “On speechreading of Swedish consonants and vowels,” *STL-QPSR*, pp. 011–033, 1974.

[13] D.W. Massaro, *Speech reading by Ear and Eye*, Erlaub, Hillsdale, 1987.

[14] T-K. Kim, H. Kim, W. Hwang, and J. Kittler, “Independent component analysis in a local residue space for face recognition,” *Pattern Recognition*, vol. 2004, no. 37, pp. 1873–1885, 2004.