**Appears in *Interspeech*, pp 2238–2241, Pittsburgh, PA, USA 2006.**

# Reconstructing Tongue Movements from Audio and Video

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

[1] Div. Sensor Technology, Swedish Defence Research Agency, SE–164 90 Stockholm, Sweden
[2] Centre for Speech Technology (CTT), CSC, KTH, SE–100 44 Stockholm, Sweden
[3] Human-Computer Interaction Group, CSC, KTH, SE–100 44 Stockholm, Sweden
{hedvig,engwall,balter}@kth.se

## Abstract

This paper presents an approach to articulatory inversion using audio and video of the user's face, requiring no special markers. The video is stabilized with respect to the face, and the mouth region cropped out. The mouth image is projected into a learned independent component subspace to obtain a low-dimensional representation of the mouth appearance. The inversion problem is treated as one of regression; a non-linear regressor using relevance vector machines is trained with a dataset of simultaneous images of a subject's face, acoustic features and positions of magnetic coils glued to the subjects's tongue. The results show the benefit of using both cues for inversion. We envisage the inversion method to be part of a pronunciation training system with articulatory feedback.

**Index Terms**: audio-visual to articulatory inversion.

## 1. Introduction

We are in the process of creating an automatic language training system, which gives the user articulatory feedback on the pronunciation [1]. One of the core challenges is the resynthesis of the user's articulatory movements from audio and video, i.e. the problem of audiovisual-to-articulatory inversion.

There are two approaches to inversion, a generative analysis-by-synthesis approach [2, 3, 4], and a discriminative approach [5, 6, 7, 8]. Like all generative approaches, inversion-by-synthesis is computationally demanding, and requires a generative model of the tongue and vocal tract. The downside of the discriminative approach is that it requires large volumes of labeled training data. In this paper, we follow the discriminative approach.

Previous attempts [3, 5, 6, 7, 8] at discriminative visual-to-articulatory or audio-visual-to-articulatory inversion have shown that important information on the tongue position may be gained from the speaker's face. All these studies however used 3D motion capture of the face, while in this paper, we investigate the possibilities of reconstructing the tongue shape from markerless video of the face. For the computer assisted pronunciation training system we envisage [1], this is necessary since the need for markers or blue lipstick would compromize the usability of the system.

Most existing methods for extracting information from face video rely on extracting the lip contours, either for lip tracking or for parameter extraction. The lip countours are modeled using snake-like methods [9, 10] or data driven principal component analysis (PCA) methods [11, 12, 13]. In contrast, we do not attempt to explicitly model the lip shape, for two reasons. Firstly, tracking the lips is difficult and computationally demanding to do robustly [9]. Secondly, important information, such as shading indicating lip protusion, and visibility of the teeth and tongue, is not present in the lip shape.

Instead of tracking the lips, we track the face as a whole [14], which is less deformable and easier to track. We then stabilize the lips in each image and extract the mouth region. The articulatory information in this region is represented in terms of the independent components of the lip image.

The approach is similar to methods used in [11, 13], with the execption that these studies stabilized the images by tracking the lips themselves. In [15] features in the face were tracked, but lip contours were then extracted and used as a basis for recognition. Saenko et al. [16] do not detect lip contours, but instead extract binary articulatory mouth features. Although robust for separation between a small set of words, the approach renders a quite coarse representation that might be unsuitable for inversion.

We explore the possibilities of reconstructing the tongue shape from the independent components of the lip image, from the acoustic signal, and from combinations thereof.

## 2. Data Acquisition

The midsagittal position of six electromagnetic articulography (EMA) coils on the tongue, jaw, upper incisor and upper lip were recorded simultaneously with the audio signal and video of the subject's face [17]. The data from the three coils on the tongue, Tg1–Tg3 (approximately 8, 20, 52 mm from the tip, respectively), and on the jaw were used in this study.

The subject was a female speaker of Swedish, judged as highly intelligible by hearing-impaired listeners. The corpus used in this study consisted of 63 symmetric VCV words with V=[a, ɪ, ʊ] and C=[p, t, k, b, d, ɡ, f, s, ɕ, ɧ, m, n, ŋ, l, r, ɳ, ʈ, ɖ, v, j]. Each word appeared only once in the training set.

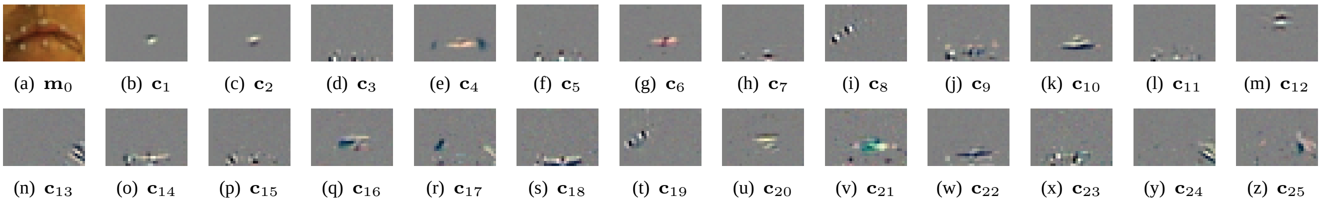
## 3. Data Processing

### 3.1. Speech Acoustics

The audio signal was originally sampled at 16 kHz. For correlation with the articulatory data, the audio signal was divided into frames of length 24 ms with a shift of 16.67 ms. Each acoustic frame was pre-emhasized and multiplied by a Hamming window. A covariance-based LPC algorithm [18] was then applied to generate 16 line spectrum pairs (LSP), which are closely related to the formant frequencies and the vocal tract shape [5, 6, 8].

To enable correlation with the PAL video stream, the speech signal was finally resampled with linear interpolation to 25 Hz, giving a sequence of 2101 17-dimensional vectors $\mathbf{a}_k$, consisting of the 16 LSP coefficient and the RMS amplitude in each frame.

[*] Formerly Hedvig Sidenbladh



(a) $\mathbf{m}_0$ (b) $\mathbf{c}_1$ (c) $\mathbf{c}_2$ (d) $\mathbf{c}_3$ (e) $\mathbf{c}_4$ (f) $\mathbf{c}_5$ (g) $\mathbf{c}_6$ (h) $\mathbf{c}_7$ (i) $\mathbf{c}_8$ (j) $\mathbf{c}_9$ (k) $\mathbf{c}_{10}$ (l) $\mathbf{c}_{11}$ (m) $\mathbf{c}_{12}$

(n) $\mathbf{c}_{13}$ (o) $\mathbf{c}_{14}$ (p) $\mathbf{c}_{15}$ (q) $\mathbf{c}_{16}$ (r) $\mathbf{c}_{17}$ (s) $\mathbf{c}_{18}$ (t) $\mathbf{c}_{19}$ (u) $\mathbf{c}_{20}$ (v) $\mathbf{c}_{21}$ (w) $\mathbf{c}_{22}$ (x) $\mathbf{c}_{23}$ (y) $\mathbf{c}_{24}$ (z) $\mathbf{c}_{25}$

Figure 1: ICA base for lip images. (a) Template $\mathbf{m}_0$. (b-z) The first 25 independent components $\mathbf{c}_i$ learned from a set of $N = 472$ difference images $\mathbf{x}_k = \mathbf{m}_k - \mathbf{m}_0$.

### 3.2. Video Data

The video had a frame-rate of 25 Hz, each frame $768 \times 576$ pixels. The subject was wearing white markers for a 3D motion capture system, but they were not employed in the lip parameter extraction.

The subject's mouth was stabilized in the image by tracking of the speaker's face [14]. After down-sampling, a 25 Hz, $33 \times 23$ pixel video of the mouth was obtained.

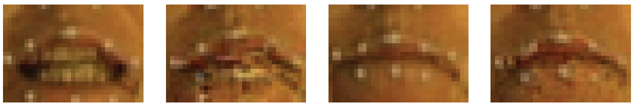
A low-dimensional representation of the mouth was learned according to the following. Consider a set of $N$ mouth images $\mathbf{m}_k$. Subtract a template image $\mathbf{m}_0$ with neutral lip pose (Figure 1a) from $\mathbf{m}_k$, the R, G and B bands subtracted separately. The difference image can be represented as a column vector $\mathbf{x}_k = \mathbf{m}_k - \mathbf{m}_0$ of size $d$, with $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. A projection of these vectors onto a base $C = [\mathbf{c}_1, \ldots, \mathbf{c}_n]$, where $n \leq N, n \leq d$ can be expressed as $X \approx CV$ where $V$ is a parameter matrix in the subspace defined by $C$. The base $C$ should be selected to represent the data set $X$ as well as possible.

This can be done in a number of ways, of which two are principal component analysis (PCA) [11, 12, 13] and independent component analysis (ICA) [19]. Using PCA, $C$ is selected so that the columns represent the $n$ largest principal components (eigenvectors) of the data set. In ICA, $C$ is instead selected as the $n$ most informative statistically independent components of the dataset – a more compact representation of the dataset than PCA. In our study we hence employed ICA (Figure 1).

All difference frames $\mathbf{x}_k$ in the training set were now projected onto the learned subspace $C$. With $n = 50$ and $d = 33 \times 23 \times 3$, we obtained a sequence of 2101 vectors $\mathbf{v}_k$ which were approximate representations of the mouth images $\mathbf{m}_k$ (Figure 2). The effect of different compression rates $\frac{n}{d}$ on the representation of visual articulatory features is discussed in a separate coming study.

#### 3.2.1. *The White Motion Capture Markers*

The subject was wearing reflective markers for the 3D motion capture system [17]. These markers were not used in this study, neither in the stabilization nor in the ICA learning. However, the markers clearly affected the component base (Figures 1b-z).



(a) Orig. /ɪ/ (b) ICA /ɪ/ (c) Orig. /p/ (d) ICA /p/

Figure 2: ICA representation. (a) Original frame $\mathbf{m}_k$, sound /ɪ/. (b) Reconstruction of the same frame $\mathbf{m}_0 + \Sigma_{i=1}^{n} v_{k,i}\mathbf{c}_i$, using $n = 50$. (c) Original frame $\mathbf{m}_k$, sound /p/. (d) Reconstruction of the same frame $\mathbf{m}_0 + \Sigma_{i=1}^{n} v_{k,i}\mathbf{c}_i$, using $n = 50$.

The question is if the markers improved or worsened the results. The reconstruction in Figure 2b indicates that the markers do *not* help in reconstruction; only four of the markers in Figure 2a are reconstructed properly in Figure 2b. Moreover, since our method does not rely on tracking of individual features around the mouth, but rather on a holistic representation of the mouth pattern, it whould even be possible that the markers cause the ICA method to fail to represent information about shadowing and teeth visibility, leading to a mouth representation with less expressive power.

We hence consider that the results presented below represent a fair estimation of what may be achieved with audiovisual inversion for an unmarked face, rather than a special case of marker tracking.

### 3.3. EMA Data

In previous studies using this data set [7, 8, 17], EMA data at a sampling rate of 60 Hz has been used, while we in the present study use EMA data resampled with linear interpolation to 25 Hz, to correspond to the frequency of the PAL video stream. This gives a sequence of 2101 8-dimensional vectors $\mathbf{t}_k$ (horizontal and vertical position of the four EMA coils in the midsagittal plane).

## 4. Inversion

For inversion, we want to learn functions $f^A$, $f^V$, $f^{AV}$, mapping respectively acoustic, video data and acoustic and video data to estimated EMA coil positions as $\hat{\mathbf{t}}_k^A = f^A(\mathbf{a}_k)$, $\hat{\mathbf{t}}_k^V = f^V(\mathbf{v}_k)$, $\hat{\mathbf{t}}_k^{AV} = f^{AV}(\mathbf{a}_k, \mathbf{v}_k)$. The set of training triples $(\mathbf{a}_k, \mathbf{v}_k, \mathbf{t}_k)$ can be used to learn these functions.

Previous similar inversion approaches [5, 6, 7, 8] have used linear or multi-linear regression to learn these functions. However, the relationship between the ICA parameters and the EMA coil positions can be expected to be higly non-linear. Thus, we employ a relevance vector machine (RVM) [20], which is a non-linear kernel-based regression technique.

### 4.1. Fusion of Audio and Video

There are two approaches to fusing the two modalities in the function $f^{AV}$, early and late fusion.

An early fusion approach is to simply concatenate the training vectors as $\hat{\mathbf{t}}_k^{AVearly} = f^{AV}([\alpha^A \mathbf{a}_k^T \ \mathbf{v}_k^T]^T)$ where $\alpha^A = \frac{\bar{\sigma}^V}{\bar{\sigma}^A}$ is a normalizing scale factor, $\bar{\sigma}^A$ and $\bar{\sigma}^V$ being the mean standard deviations in the audio and video datasets.

Late fusion instead performs regression separately for the two modalities, combining the results as $\hat{\mathbf{t}}_k^{AVlate} = \Gamma^A f^A(\mathbf{a}_k) + \Gamma^V f^V(\mathbf{v}_k)$ where $\Gamma^A$ and $\Gamma^V$ are a diagonal matrices whose respective elements are $\frac{(\rho_i^A)^2}{(\rho_i^A)^2 + (\rho_i^V)^2}$ and $\frac{(\rho_i^V)^2}{(\rho_i^A)^2 + (\rho_i^V)^2}$, $\boldsymbol{\rho}^A$ and $\boldsymbol{\rho}^V$ being the correlations between $\mathbf{t}_k$ and $\hat{\mathbf{t}}_k^A$ and $\hat{\mathbf{t}}_k^V$ respectively.

Both these approaches were evaluated in our study.

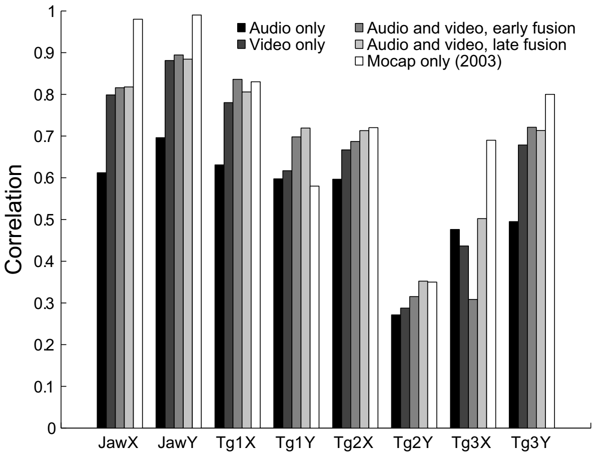


Figure 3: Correlation coefficients $\boldsymbol{\rho}^A$, $\boldsymbol{\rho}^V$, $\boldsymbol{\rho}^{AVearly}$, $\boldsymbol{\rho}^{AVlate}$ for jaw and tongue tip (Tg1), middle (Tg2) and back (Tg3) coil position (X vertical, Y horizontal). Correlations are also shown for 3D motion capture data of the face [7] ”Mocap only (2003)”.

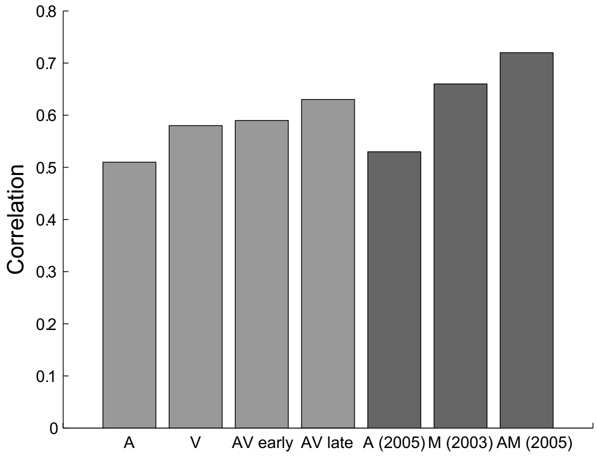


Figure 4: Mean correlation coefficients $\bar{\rho}^A$, $\bar{\rho}^V$, $\bar{\rho}^{AVearly}$, $\bar{\rho}^{AVlate}$ for the three EMA tongue coils. Mean correlations are also shown for audio [8] ”A (2005)”, 3D motion capture data of the face [7] ”M (2003)”, and a combination thereof [8] ”AM (2005)”.

## 5. Results

The training data was employed in a jackknife fashion: The training set was divided into 10 equally large parts. One part in turn was removed from the training data and used as test set, while the functions $f^A$, $f^V$, $f^{AV}$ were learned from the 9 others. This gave estimates $\hat{\mathbf{t}}_k^A$, $\hat{\mathbf{t}}_k^V$, $\hat{\mathbf{t}}_k^{AVearly}$, $\hat{\mathbf{t}}_k^{AVlate}$ for all training frames, with no overlap between training and test sets.

### 5.1. Correlation

For each frame $k$, the correlations $\boldsymbol{\rho}^A$, $\boldsymbol{\rho}^V$, $\boldsymbol{\rho}^{AVearly}$, $\boldsymbol{\rho}^{AVlate}$ between true EMA coil position $\mathbf{t}_k$ and the reconstructed positions $\hat{\mathbf{t}}_k^A$, $\hat{\mathbf{t}}_k^V$, $\hat{\mathbf{t}}_k^{AVearly}$, $\hat{\mathbf{t}}_k^{AVlate}$ were computed. Figure 3 shows the correlation coefficients for each coil parameter individually. The low correlation for parameter Tg2 Y is due to rapid tongue groove changes and spurious measurement errors for this coil.

Except for the horizontal position of the backmost tongue coil, the video makes a larger contribution than the acoustic signal, not only for the front coils, but even for positions further back, which are often considered impossible to lipread.

Compared to the results using 3D motion capture of the face, it is natural that the reconstruction of the jaw from the video images is not as perfect as from the 3D data, since both the horizontal and vertical jaw movement is given almost directly by the 3D data, while they must be estimated from the video. The horizontal movement is a hidden parameter, indicated only by changes in shading, while the vertical movement needs to be estimated from the shape and size of the mouth opening rather than from an absolute position.

It is noteworthy that the vertical tongue tip position is estimated better from video images than from 3D motion capture data, which may be explained by the fact that the tongue tip will actually be visible in some of the video images. For the remaining tongue coil coordinates, the estimation from video images is only marginally worse than that of the 3D motion capture, except for the back tongue coil. The better estimation of Tg3 from 3D motion capture data is probably due either to information given by markers on other parts of the face or the correlation between jaw and tongue movements: since the jaw position is almost perfectly estimated from the 3D motion capture, this data will have the upper hand for all frames for which there is no independent tongue movement with respect to the jaw.

Figure 4 shows that audio-visual speech inversion outperforms both acoustic- and visual-to-articulatory inversion, which is natural, since the two modalities are complementary. Figure 4 further indicates, in accordance with previous studies using 3D motion capture of the face [7, 8], that visual data contributes more than the acoustic data. The video images of the speaker’s lips can naturally not provide as much information for the inversion as 3D motion capture data of the entire face, but the improvement compared to the audio only case is nevertheless important.

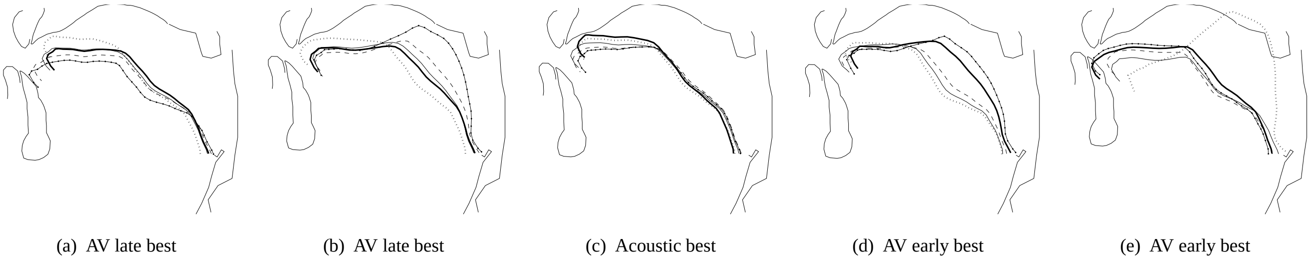
While the early fusion of audio and visual data is only marginally better than visual alone data, late fusion results in a substantially higher correlation. Interestingly, this concords with influential theories on human speech perception (e.g. [21, 22]) stating that humans process information within a modality independently and then fuse the processed, rather than the raw, data.

### 5.2. Articulatory reconstruction

To analyze the quality of the reconstructed coil positions in an articulatory context, the EMA coil positions were used to reconstruct the tongue shape in an articulatory model [23]. The conversion of EMA coil positions into the parameters controlling the model is described in [7]. It is based on a simultaneous optimization of the articulatory parameters in order to minimize the Euclidean distance between the three EMA coils and the tongue contour with the constraints that the tongue tip of the contour should correspond to that infered from the first tongue coil.

Figure 5 shows five different cases of reconstruction, with Figure 5a being the ”mean” case, i.e. that the audio-visual input is better than either modality alone. The most common is however that one of the modalities, either the visual (Figure 5b) or the acoustic (Figure 5c) is better than the other, thus contributing to a higher degree to the audio-visual reconstruction than the other. It would therefore be interesting to investigate online computation of the optimal fusion weights $\Gamma^A$ and $\Gamma^V$ for each new frame.

Comparing early and late fusion, Figures 5d-e give an illustra-



(a) AV late best
(b) AV late best
(c) Acoustic best
(d) AV early best
(e) AV early best

Figure 5: Reconstructed midsagittal tongue shapes, from original EMA data (thick solid line), acoustics only (dashed), video only (dotted), audio-visual early fusion (dash-dotted) and audio-visual late fusion (solid).

tion of the fact that the early fusion is often better when either both modalities fail or one of the modalities fail completely, whereas Figures 5a-c show that late fusion is better when at least one of the modalities is successful. Moreover, the early fusion fails less gracefully than the late fusion (Figure 5b), which also explains the over-all better correlation results for the late fusion.

## 6. Conclusions

We have in this study shown that automatic extraction of lip components from a video image can contribute substantially to a discriminative articulatory inversion. Indeed, the contribution given by the visual modality is higher than that from the acoustic, and even if the results obtained from video images of the face cannot quite match those of 3D motion capture, these results are more promising for pronunciation training applications, since the audio-visual inversion can be done without special measurement equipment on the user. Our results suggest that the best results may be achieved by processing the audio and video streams separately and subsequently fuse the inversion estimations.

## 7. Acknowledgements

This research is part of the two projects ARTUR, funded by the Swedish Research Council, and ASPI, funded by the EU $6^{th}$ framework program. The Centre for Speech Technology is supported by The Swedish Agency for Innovation Systems, KTH and participating Swedish companies and organizations.

## 8. 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*, in press.

[2] S. Maeda, Ed., *SpeechMaps, WP2 - From speech signal to vocal tract geometry*, vol. III, 1994.

[3] G. Bailly and P. Badin, “Seeing the tongue from outside,” in *ICSLP*, 2002, pp. 1913–1916.

[4] S. Ouni and Y. Laprie, “Introduction of constraints in an acoustic-to-articulatory inversion,” in *ICSLP*, 2002, pp. 2301–2304.

[5] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, “Quantitative association of vocal-tract and facial behaviour,” *Speech Communication*, vol. 26, pp. 23–43, 1998.

[6] J. Jiang, A. Alwan, L. Bernstein, P. Keating, and E. Auer, “On the correlation between facial movements, tongue movements and speech acoustics,” in *ICSLP*, 2000.

[7] O. Engwall and J. Beskow, “Resynthesis of 3D tongue movements from facial data,” in *Eurospeech*, 2003, pp. 49–54.

[8] O. Engwall, “Introducing visual cues in acoustic-to-articulatory inversion,” in *Interspeech*, 2005, pp. 3205–3208.

[9] R. Kaucic and A. Blake, “Accurate, real-time, unadorned lip tracking,” in *ICCV*, 1998, pp. 370–375.

[10] 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.

[11] C. Bregler and Y. Konig, ““Eigenlips” for robust speech recognition,” in *ICASSP*, 1994, pp. 669–672.

[12] 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.

[13] 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.

[14] H. Kjellström and O. Bälter, “Stabilization of face sequences with particle filters,” Tech. Rep. pending, KTH, Stockholm, Sweden, 2006.

[15] I. Shdaifat and R-R. Grigat, “A system for audio-visual speech recognition,” in *Interspeech*, 2005, pp. 1221–1224.

[16] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, “Visual speech recognition with loosely synchronized feature streams,” in *ICCV*, 2005, pp. 1424–1431.

[17] J. Beskow, O. Engwall, and B. Granström, “Resynthesis of facial and intraoral motion from simultaneous measurements,” in *ICPhS*, 2003, pp. 431–434.

[18] 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.

[19] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[20] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, vol. 2001, no. 1, pp. 211–244, 2001.

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

[22] Q. Summerfield, “Some preliminairies to a comprehensice account of audio-visual perception,” in *Hearing by eye: The psycohology of lip-reading*, pp. 3–51. Erlaub, London, 1987.

[23] O. Engwall, “Combining MRI, EMA & EPG in a three-dimensional tongue model,” *Speech Communication*, vol. 41, no. 2-3, pp. 303–329, 2003.