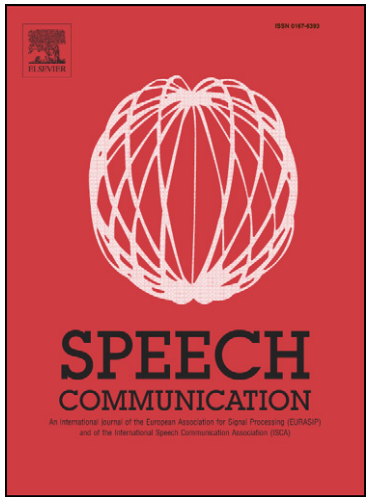
Audiovisual-to-Articulatory Inversion

Hedvig Kjellström, Olov Engwall



Audiovisual-to-Articulatory Inversion

Hedvig Kjellström, Olov Engwall

PII: S0167-6393(08)00127-1
DOI: 10.1016/j.specom.2008.07.005
Reference: SPECOM 1743

To appear in: *Speech Communication*

Received Date: 8 February 2008
Revised Date: 24 July 2008
Accepted Date: 24 July 2008

Please cite this article as: Kjellström, H., Engwall, O., Audiovisual-to-Articulatory Inversion, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.07.005

Audiovisual-to-Articulatory Inversion

Hedvig Kjellström [1]

Computer Vision and Active Perception Lab, School of Computer Science and Communication, KTH (Royal Institute of Technology), SE–100 44 Stockholm, Sweden

Olov Engwall

Centre for Speech Technology (CTT), School of Computer Science and Communication, KTH (Royal Institute of Technology), SE–100 44 Stockholm, Sweden

Abstract

It has been shown that acoustic-to-articulatory inversion, i.e. estimation of the articulatory configuration from the corresponding acoustic signal, can be greatly improved by adding visual features extracted from the speaker's face. In order to make the inversion method usable in a realistic application, these features should be possible to obtain from a monocular frontal face video, where the speaker is not required to wear any special markers. In this study, we investigate the importance of visual cues for inversion. Experiments with motion capture data of the face show that important articulatory information can be extracted using only a few face measures that mimic the information that could be gained from a video-based method. We also show that the depth cue for these measures is not critical, which means that the relevant information can be extracted from a frontal video. A real video-based face feature extraction method is further presented, leading to similar improvements in inversion quality. Rather than tracking points on the face, it represents the appearance of the mouth area using independent component images. These findings are important for applications that need a simple audiovisual-to-articulatory inversion technique, e.g. articulatory phonetics training for second language learners or hearing-impaired persons.

Key words: speech inversion, articulatory inversion, computer vision

[1] Formerly Hedvig Sidenbladh. Corresponding author, phone +46-8-790 6906.
Email addresses: `hedvig@kth.se` (Hedvig Kjellström), `engwall@kth.se` (Olov Engwall).
URLs: `www.csc.kth.se/~hedvig` (Hedvig Kjellström), `www.csc.kth.se/~engwall` (Olov Engwall).

List of Figures

List of Tables

1 Introduction

Acoustic-to-articulatory inversion, i.e. estimation of the articulatory configuration from the corresponding acoustic signal, can be considered as one of the greatest challenges in speech technology research. The applications of a working inversion method are many. A reliable and robust inversion method would provide a major breakthrough in computer assisted language learning, as automatic articulatory feedback could be given on the student's pronunciation (Engwall et al., 2006). Furthermore, it could be beneficial in speech coding and recognition, by transforming complex acoustic feature vectors to slowly varying articulatory parameters.

However, important problems need to be solved en route towards robust inversion. The greatest obstacle in acoustic-to-articulatory inversion is probably the non-unicity of the inverse solutions; the same sound could have been generated by not one, but a range of articulatory configurations. It has been shown (Yehia et al., 1998; Jiang et al., 2002) that important information regarding the tongue position may be gained from 3D marker data of the speaker's face. Moreover, the face motion is to a degree conditionally independent of the acoustic information given the articulation, which means that the face gives additional information about the articulatory configuration, not encoded in the speech signal. A way to address the non-unicity problem is thus to infer the articulatory configuration from both speech and visual information, i.e. to perform *audiovisual-to-articulatory inversion*.

Articulatory features, such as jaw position, mouth opening and lip rounding and protrusion, should ideally be extracted from a video of the subject's face, not relying on either markers on the face or calibrated stereo. However, if monovideo is to be a viable solution, the improvement of the inversion results compared to acoustic only case needs to approach that of marker tracking. At KTH, we have investigated different variants of visual-to-articulatory and audiovisual-to-articulatory inversion (Engwall and Beskow, 2003; Engwall, 2005; Engwall, 2006; Kjellström et al., 2006). In this article, we present the common general approach, and compare the performance of different types of visual data. We also study different methods for mapping the visual and acoustic signal to the articulatory configuration. Taking the results in previous studies (Yehia et al., 1998; Jiang et al., 2002) as a starting point, we want to investigate the following five issues:

(1) Does information from the face aid the inversion, even if only a small number of points are tracked?
(2) Is 3D data required for these points or is a 2D frontal view sufficient?
(3) Is it beneficial to employ a non-linear mapping between the audiovisual and the articulatory data as opposed to a linear mapping?

(4) How does a video-based face feature extraction algorithm perform compared to marker tracking?
(5) How should the acoustic and visual information be combined to achieve the best inversion results?

There are two approaches to inversion: a generative analysis-by-synthesis approach (Maeda, 1994; Bailly and Badin, 2002; Ouni and Laprie, 2002), and a discriminative approach (Yehia et al., 1998; Jiang et al., 2002).

The inversion problem has mostly been addressed by the generative approach, i.e., an articulatory model is used to create a feature description of all possible sounds (Maeda, 1994; Ouni and Laprie, 2002). Inversion from acoustic data to an articulatory description is then performed by matching the acoustic data to the synthesized output of the model, using a lookup in an articulatory codebook. Like all generative approaches, inversion-by-synthesis is computationally demanding, and requires a generative model of the tongue and vocal tract. This approach put high demands on the quality of the synthesis (or else discrepancies might result from the synthesis rather than the inversion), and is currently mainly applicable to vowels.

In the discriminative approach, which we follow, the quantitative association of visio-acoustic and articulatory data of the tongue is determined statistically (Yehia et al., 1998; Jiang et al., 2002). In this case, articulatory parameters are related to the speech and/or visual signals using regression. Both Yehia et al. (1998) and Jiang et al. (2002) found rather high correlation between acoustic, facial and vocal tract data. While both these studies used linear regressors, we investigate non-linear alternatives as well. The downside of the discriminative approach is that it requires large volumes of labeled training data.

For our experiments, simultaneous visual (both video and 3D marker positions), acoustic and articulatory data are collected. The data acquisition is detailed in Section 2.

We first investigate how different elements of face motion contribute to the inversion. To investigate to what extent depth information contributes to the inversion, we perform experiments (Section 5.1) with three sets of marker data; the full 3D positions, sparse "stereo" data mimicking what could be recovered from unmarked stereo video of the lips, and "monocular" 2D data that would be possible to extract from unmarked monocular lip images. The three datasets are more closely described in Section 3.3. In short, we find that depth information does not contribute significantly to the inversion, and that even though the full dataset gives the best results, important improvements are also obtained using a few measures from the lip configuration only.

The results hence suggest that it should be possible to improve articulatory

inversion using an approach not relying on 3D reconstruction of face features. We therefore also evaluate an approach for extraction of face information without explicit marker tracking (Section 5.3).

A number of methods have been presented for extracting non-rigid motions of face articulations from video. One approach is to reconstruct the surface of the face in 3D. The estimation of a non-rigid shape in 3D from a monocular video sequence is of course underconstrained, but the 3D reconstruction can be constrained using assumptions about temporal and spatial smoothness (Ahlberg, 2001; Torresani and Hertzmann, 2004). Additional robustness is gained from propagation of information both forward and backward in time (Torresani and Hertzmann, 2004), which however is prohibitive in a real-time application such as ours.

An approach more commonly used for speech inversion and recognition is to not explicitly model the depth information, but instead use a 2D model. Many methods rely on extracting the lip contours (Bregler and Konig, 1994; Kaucic and Blake, 1998; Dupont and Luettin, 2000; Matthews et al., 2002; Seymour et al., 2005). The lip countours are modeled using snake-like methods (Kaucic and Blake, 1998; Seymour et al., 2005) or data driven PCA methods (Bregler and Konig, 1994; Dupont and Luettin, 2000; Matthews et al., 2002). Alternatively, the shape and appearance of the whole face can be modeled using, e.g., Active Appearance Models (AAM) (Cootes et al., 2001). When employed for articulatory inversion (Katsamanis et al., 2008), an AAM improved the inversion results on the dataset presented in Section 2 by 25%.

A third option is to not explicitly estimate the shape at all, and just model the appearance. Saenko et al. (2005) do not detect lip contours, but instead employ a cascade of support vector machines. Each stage in the cascade partitions lip images according to speaking/non-speaking, closed/narrow/medium/wide, rounded/unrounded, etc. This approach is very robust and enables separation between a small set of spoken commands without the use of acoustic information. However, the approach renders a quite coarse representation which is unsuitable for visual-to-articulatory inversion.

We also take this approach of not explicitly modeling the lip or face shape. Non-rigid tracking of an articulated face inevitably introduces some errors, due to image noise and necessary simplifications in the model compared to the real face. We wish to investigate how well we can do without modeling shape at all, i.e., avoiding the tracking of face articulation altogether. The articulatory information is represented in terms of the independent components (Hyvärinen et al., 2001) of the lip image. A similar approach using PCA is used by Bregler and Konig (1994) and Matthews et al. (2002). Details of our lip tracking approach can be found in Section 3.4.

Mappings from the visual and acoustic features to the articulatory features are learned using either linear or non-linear regression (Section 4). These two types of regression methods are compared in the experiment in Section 5.2. We also explore two different approaches for combining acoustic and visual data, early and late fusion (Section 5.3). The inversion is found to improve between 20% and 44% when adding the extracted face parameters to the acoustic information.

2 Data Acquisition

2.1 Subject and Corpus

The subject was a female speaker of Swedish, judged as highly intelligible by hearing-impaired listeners.

The corpus consisted of 135 symmetric $VC_1\{C_2C_3\}V$ words (henceforth referred to as VCV words) where V=[a, ɪ, ʊ] and C=[p, t, k, b, d, g, f, s, ɕ, ʃ, m, n, ŋ, l, r, ɳ, ʈ, ɖ, v, j, h, jk, rk, pl, bl, kl, gl, fl, pr, br, kr, gr, kt, nt, tr, dr, fr, st, sp, sk, sl, str, spr, skr, skl] and 178 short sentences (4-5 words, 7-9 syllables). Each VCV word appeared only once in the training set, but some consonants vary in voicing only ([p] *vs.* [b], [t] *vs.* [d] and [k] *vs.* [g]). In the sentences some words appear several times, but each sentence is unique. The sentences have a simple structure (subject, predicate, object) and "everyday content", such as "Kappan hänger i garderoben" (The coat hangs in the wardrobe) or "Laget förlorade matchen" (The team lost the game). They followed a rationale for audiovisual speech perception tests (MacLeod and Summerfield, 1990) and were adapted to Swedish by G. Öhngren.

2.2 Measurement Setup

The midsagittal position of points on the tongue, jaw and upper lip were recorded simultaneously with the audio signal, 3D points on the subject's face and video of the subject's face.

The video had a frame-rate of 25 Hz, and each frame had a resolution of 768×576 pixels (Figure 1a).

Figure 1b shows the positions of the 25 facial markers used in the motion capture. The 3D positions of the 4 mm large markers were tracked at 60 Hz using four infrared cameras in the MacReflex system from Qualisys, which has a spatial accuracy of well below 1 mm. An additional three fixed markers

(on the headmount in Figure 1a) were used to adjust for head movements. In the reduced datasets (Section 3.3), the positions of only five markers are maintained: J = jaw, RC = right lip corner, LC = left lip corner, LL = lower lip edge and UL = upper lip edge (Figure 1b). Arrows indicate if horizontal or vertical movement was used in the monovision dataset.

Tongue points were measured with the electromagnetic articulography (EMA) system Movetrack (Branderud, 1985), with 4x1.5 mm coils glued onto the tongue, jaw and lips, as shown in Figure 1c. The coils on the upper lip and upper incisor were used to align the motion capture data with the EMA data, to create a common 3D dataset. The three coils on the tongue (T1–T3, approximately 8, 20, 52 mm from the tip, respectively) and the coil on the jaw (JW) were used in the inversion experiments. Details of the recording procedure can be found in (Beskow et al., 2003).

3 Data Processing

3.1 Speech Acoustics

The audio signal was originally sampled at 16 kHz, and a resampling to match the different types of visual data was hence needed.

For correlation with the 60 Hz motion capture data, the audio signal was divided into frames of length 24 ms with a frame shift of 16.67 ms. Each acoustic frame was pre-emhasized and multiplied by a Hamming window. A covariance-based LPC algorithm (Sugamura and Itakura, 1986) was then applied to generate 16 line spectrum pairs (LSP). LSP coefficients were used, following (Yehia et al., 1998) and (Jiang et al., 2002), as they are closely related to the formant frequencies and the vocal tract shape. In total, after resampling, there were 5090 frames of speech in the VCV corpus.

To enable correlation with the 25 Hz PAL video stream, a lower resoluted version of the VCV corpus described above was also created, by resampling with linear interpolation. It should be noted that the 25 Hz VCV dataset thus was a factor 2.4 smaller than the 60 Hz VCV dataset, with 2101 frames.

A frame k of the speech signal is henceforth denoted $\mathbf{a}_k$. It is a vector of length 17, consisting of the 16 LSP coefficient and the RMS amplitude at timestep k.

3.2 EMA Data and Articulatory Parameters

The EMA data was resampled to 60 Hz and spatially aligned with the facial markers, to create a coherent data set of face and tongue movements. A 25 Hz version of the EMA data was also created to cohere with the PAL video stream, by resampling with linear interpolation. This 25 Hz dataset was also a factor 2.4 smaller than the corresponding 60 Hz dataset.

The EMA positions $\mathbf{t}_k$ can be mapped to the corresponding articulatory parameters of a tongue model f^T (Engwall, 2003). The 3D tongue model was derived from a statistical analysis of 3D MR images of one Swedish subject producing a corpus of 13 vowels in isolation and 10 consonants in three symmetric VCV contexts. Parameter values for jaw height (JH), dorsum raise (TD), body raise (TB), tip raise (TT), tip advance (TA) and tongue width (TW) are set based on the the horizontal and vertical positions of the three tongue coils (T1–T3) and jaw coil (JW) in the midsagittal plane at each timestep k:

$$\boldsymbol{\tau}_k = f^T(\mathbf{t}_k) \tag{1}$$

where $\boldsymbol{\tau}_k$ is a vector of the 7 articulatory parameters.

The parameter values are estimated using a global optimisation to minimize the difference in position between the measured EMA coils and corresponding virtual markers in the model. The optimisation relies on a goodness of fit between real and virtual marker positions, while preserving the tongue volume and keeping the parameter values within the allowed range defined by the MRI data. The fitting procedure is described in detail in (Beskow et al., 2003).

The benefit of estimating the articulatory parameters $\boldsymbol{\tau}_k$ rather than the EMA coil positions $\mathbf{t}_k$ from the visual and acoustic signals is that the parameters have an articulatory relevance; a qualitative investigation may hence be made on the amount and type of articulatory information (place and degree of constriction, movement of different articulators etc.) that can be recovered from the audiovisual data. Jiang et al. (2002) grouped different CV syllables based on the consonant place and manner of articulation and found some statistical differences between the groups, but did not investigate the significance on the articulatory level. The resulting parameter trajectories are shown in Figure 2, where it can be seen that the parameter values are to some extent clustered due to vowel context. This is particularly true for the jaw opening (JH), where all three contexts are distinct. For the body raise (TB), the [a] context separates from the other two, for the dorsum (TD) and tongue tip (TT) raise, the [ʊ] context is different and for tongue advance (TA), the [ɪ] context is distinct. The inversion results will therefore be influenced more by the estimation of the vowels than the consonants and it will hence be necessary to analyze not only

the global estimation success (correlation coefficients and RMS error), but also in terms of articulatory features for the reconstruction of the consonants.

3.3 3D Motion Capture Data

The full motion capture dataset (FQ) contained 3D position of all 25 markers (Figure 1b), i.e. 75 measures. A frame of full motion capture data is denoted $\mathbf{q}_k$. To simulate the feature information that could be gained from markerless computer vision methods, two reduced sets were generated from FQ:

The monovision dataset (MV) contains only data that could readily be recovered from a frontal image of the face: the vertical movement of markers on the jaw (J), upper lip (UL) and lower lip (LL) and the horizontal movement of markers in the right (RC) and left (LC) lip corners, i.e. five measures (Figure 1b). The sparsity of the feature points simulates the effect of removing the markers and relying on markerless feature extraction. A frame of monovision data is denoted $\mathbf{q}_k^M$.

To study how important depth cues are for the inversion, the stereovision dataset (SV) was created. Apart from the five measures of the MV dataset, it also contains the horizontal protrusion of the MV markers, in total 10 measures. It simulates features that could be recovered from a 3D articulated tracking system such as the ones presented in (Ahlberg, 2001; Torresani and Hertzmann, 2004). Let $\mathbf{q}_k^S$ be a frame of stereovision data.

3.4 Video Data

The subject was wearing white markers for the 3D motion capture system, but they were not employed in the lip parameter extraction. This is further discussed in Section 3.4.3 below.

3.4.1 Stabilization of the Mouth

The subject's mouth was first stabilized in the image by rigid tracking of the upper part of the face, which usually displays less deformation than the mouth area (similar to previous work by Shdaifat and Grigat (2005)). Face detection and tracking is a well understood problem in the computer vision field (Yang et al., 2002). Thus, when employing the inversion method in a realistic system, the face can be robustly found in the image and tracked over time using an existing method in the literature.

Here, since our data was captured under controlled lighting conditions with the subject facing the camera, a template based 2D method was used (Figure 3):

The face pose $\mathbf{y}_k$ in frame k is a vector consisting of 5 parameters; position, size and orientation of a rectangle over the upper part of the face (above the mouth) in the image. Let $\mathbf{f}_k$ be the pattern within the rectangle in frame k.

The standard deviation σ_f of face patterns has been learned from a number of short sequences of face patterns. It reflects the variation in different parts of the face; for example, the variation in the eye region is higher than on the cheeks, due to eye blinks.

A template face pattern $\mathbf{f}_0$ is first manually extracted from a template frame by clicking on the eyes and mouth. A probability density function over pose $\mathbf{y}_k$ is then estimated in each frame k of a sequence by iteratively minimizing $||\frac{\mathbf{f}_k - \mathbf{f}_0}{\sigma_f}||$ using a particle filter (Isard and Blake, 1998; Doucet et al., 2001). In the first frame of the sequence, the pose value is assumed to be normally distributed over the state-space, which corresponds to initially searching over a wide range of possible poses.

The face was tracked in all sequences using the above method. After spatial down-sampling, a 33×23 pixel video of the mouth was obtained. We use $\mathbf{m}_k$ to denote the mouth frame at timestep k.

3.4.2 Low-Dimensional Representation of the Mouth

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 4a) 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) (Bregler and Konig, 1994; Dupont and Luettin, 2000; Matthews et al., 2002) and independent component analysis (ICA) (Hyvärinen et al., 2001). 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 4).

Figures 4b-z show the first 25 independent components learned from $N =$

472 difference images of dimensionality $d = 33 \times 23 \times 3$. The independent component separation was implemented in Matlab using the FastICA package (Hyvärinen, 2005) with a radial basis kernel of variance 5.

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 sequece of vectors $\mathbf{v}_k$ which were approximate representations of the mouth images $\mathbf{m}_k$ (Figure 5).

3.4.3 The White Motion Capture Markers

The subject was wearing reflective markers for the 3D motion capture system. These markers were not used in the visual parameter extraction, neither in the stabilization nor in the ICA learning. However, the markers clearly affected the component base (Figures 4b-z).

The question is do the markers improve the results? The reconstructions in Figures 5b and 5d indicate that the markers do *not* help in reconstruction; only four of the markers in Figure 5a and none of the markers in Figure 5c were reconstructed properly in Figures 5b and 5d respectively. 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 would even be possible that the markers cause the ICA method to fail to represent some 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.

4 Inversion

For inversion, we want to learn functions f that map acoustic data ($\mathbf{a}_k$) and/or visual data ($\mathbf{v}_k$, $\mathbf{q}_k$) to estimated articulatory parameters ($\hat{\boldsymbol{\tau}}_k$). The set of training triples ($\mathbf{a}_k, \mathbf{v}_k, \boldsymbol{\tau}_k$) and ($\mathbf{a}_k, \mathbf{q}_k, \boldsymbol{\tau}_k$) can be used to learn these functions. In all experiments, the training data was employed in a jackknife fashion: The training set was divided on an utterance level into 10 equally large parts. One part in turn was removed from the training data and used as the test set, while the mapping functions were learned from the 9 others. This gave estimates $\hat{\boldsymbol{\tau}}_k$ for all training frames, with no overlap between training and test sets.

All our inversion experiments handle separate frames, which means that context information and continuity of articulatory trajectories are not used.

We first test linear regression in the experiments in Section 5.1 and then compare the results to a non-linear regression method in Section 5.2.

4.1 Linear Regression

A straight-forward method to learn a mapping function from a state-space $\mathbf{y}$ (e.g., speech parameters $\mathbf{a}$) to another state-space $\mathbf{x}$ (e.g., tongue positions $\mathbf{t}$) is linear regression (Yehia et al., 1998; Jiang et al., 2002). In a linear regressor,

$$\hat{\mathbf{x}} = f^{Lin}(\mathbf{y}) = T_{XY}\mathbf{y} \tag{2}$$

where T_{XY} is an estimator matrix, learned from K training examples as

$$T_{XY} = \mathbf{X}\,\mathbf{Y}^T(\mathbf{Y}\,\mathbf{Y}^T)^{-1}\ . \tag{3}$$

$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]$ is an m by K matrix where each column is an m-dimensional out-parameter example, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_K]$ is an n by K matrix where each column contains the corresponding n-dimensional in-parameter example.

4.2 Non-Linear Regression

The mapping from audiovisual to articulatory parameters can be expected to be non-linear, which means that a linear regressor fails to capture some aspects of the mapping. On the other hand, a non-linear regressor has more degrees of freedom and requires more training data to perform well. The choice between non-linear and linear regressor depends thus on the expected non-linearity of the mapping, and the size of the training set. In our experiments we have tested support vector machines (SVM) (Cristianini and Shawe-Taylor, 2000), relevance vector machines (RVM) (Tipping, 2001) and artificial neural networks (ANN) (Petersen, 1997) for non-linear regression. The results for the different types of non-linear regression methods are qualitatively similar, and we therefore concentrate on one non-linear method, the ANN.

A neural regressor with a quadratic cost function from the DTU Toolbox (Petersen, 1997) was used. This regressor is a two-layer feed-forward neural network with hyperbolic tangent sigmoidal functions for the hidden layer (five hidden units were used, as this yielded the best results in initial preliminary tests) and a linear output layer. The weights were optimized with a maximum

a posteriori approach and the network was trained using a quasi-Newton code for unconstrained optimization. Hereafter, the ANN regressor is denoted

$$\hat{\mathbf{x}} = f^{ANN}(\mathbf{y}) \ . \tag{4}$$

Each parameter in the vector $\hat{\mathbf{x}}$ was estimated with a separate regressor.

4.3 Fusion of Audio and Video

There are two approaches to fusing two modalities $\mathbf{y}$ and $\mathbf{z}$ (e.g. speech $\mathbf{a}_k$ and video $\mathbf{v}_k$) in an inversion function; early and late fusion.

An early fusion approach is to simply concatenate the training vectors as

$$\hat{\mathbf{x}}^{Early} = f^{YZ}([\alpha\mathbf{y}^T, \mathbf{z}^T]^T) \tag{5}$$

where $\alpha = \frac{\bar{\sigma}^Y}{\bar{\sigma}^Z}$ is a normalizing scale factor, $\bar{\sigma}^Y$ and $\bar{\sigma}^V$ being the mean standard deviations in the $\mathbf{y}$ and $\mathbf{z}$ datasets. This approach makes no assumptions about the statistical (in)dependence of $\mathbf{y}$ and $\mathbf{z}$, but has the disadvantage that the dimensionality of the state-space is high.

Late fusion instead performs regression separately for the two modalities, combining the results as

$$\hat{\mathbf{x}}^{Late} = \Gamma^Y f^Y(\mathbf{y}) + \Gamma^Z f^Z(\mathbf{z}) \tag{6}$$

where Γ^Y and Γ^Z are diagonal matrices whose respective elements are $\frac{(\rho_i^Y)^2}{(\rho_i^Y)^2+(\rho_i^Z)^2}$ and $\frac{(\rho_i^Z)^2}{(\rho_i^Y)^2+(\rho_i^Z)^2}$, $\boldsymbol{\rho}^Y$ and $\boldsymbol{\rho}^Z$ being the correlations between true out-data $\mathbf{x}$ and the reconstructions $\hat{\mathbf{x}}^Y = f^Y(\mathbf{y})$ and $\hat{\mathbf{x}}^Z = f^Z(\mathbf{z})$ respectively. The late fusion assumes independence between the $\mathbf{y}$ and $\mathbf{z}$ channel given the data, an assumption that is not true in our case of speech and video data. However, the late fusion deals with lower-dimensional state-spaces than the early fusion.

Both these approaches were used in our study. The results are presented in Section 5.3.

5 Results

The quality of the reconstruction is evaluated by comparing the estimates of the parameter values, $\hat{\boldsymbol{\tau}}_k$, with the true articulatory data $\boldsymbol{\tau}_k$ for the same

frame. Our evaluation is mainly based on correlation coefficients, but articulatory analyses of the resulting tongue shapes in the articulatory model are also performed.

5.1 Introducing Visual Cues in the Inversion

These experiments were carried out in order to study what could be gained from introducing visual cues in the inversion. The three motion capture corpora described in Section 3.3 were used and four different linear inversion functions were learned,

$$\hat{\boldsymbol{\tau}}_k^{AO} = f^{Lin,AO}(\mathbf{a}_k) \tag{7}$$
$$\hat{\boldsymbol{\tau}}_k^{AMV} = f^{Lin,Early,AMV}(\mathbf{a}_k, \mathbf{q}_k^M) \tag{8}$$
$$\hat{\boldsymbol{\tau}}_k^{ASV} = f^{Lin,Early,ASV}(\mathbf{a}_k, \mathbf{q}_k^S) \tag{9}$$
$$\hat{\boldsymbol{\tau}}_k^{AFQ} = f^{Lin,Early,AFQ}(\mathbf{a}_k, \mathbf{q}_k) \tag{10}$$

where *Lin* = linear regression, *Early* = early fusion, *AO* = audio cues only, *AMV* = audio + monovision motion capture data, *ASV* = audio + stereovision motion capture data, and *AFQ* = audio + full 3D motion capture data.

5.1.1 Acoustic Only Inversion

The correlation between the original tongue data $\boldsymbol{\tau}_k$ and that estimated from the acoustic signal only, $\hat{\boldsymbol{\tau}}_k^{AO}$, is moderate, as shown by the white bars in Figure 6. The correlation is substantially lower than that found by Jiang et al. (2002) for CV syllables (mean 0.78) and by Yehia et al. (1998) for sentences of English (mean 0.61) and Japanese (mean 0.60). It should however be noted that both Jiang et al. (2002) and Yehia et al. (1998) used four (or five) repetitions of each utterance, while we use only one. In their case, alternative repetitons of the test utterances were included in the training set. In our case, the consonant part of each VCV word appear three times (once for each context), but the coarticulatory influence of the surrounding vowels makes the three occurences differ more in the articulation.

To corroborate that the correlation coefficients give an indication of how well the entire VCV corpus is reconstructed, rather than how good it is for the much more frequent vowels [a, ɪ, ʊ], a baseline correlation was calculated. For this baseline, it was assumed that the vowels were always reconstructed as the correct vowel prototype (each parameter in the vowel frames was assigned its mean value for that vowel), while the reconstruction of consonants was

random (each parameter in the consonant frames was assigned its mean value over all consonants in the corpus). The baseline correlation for the VCV words was 0.53.

Since the correlation of the acoustic-to-articulatory inversion is not higher than the baseline correlation, it can be argued that the acoustic only estimation will be unsuccessful, at least if the training and test data are less similar than in our case.

5.1.2 Acoustic Inversion Supported by Face Measures

Large improvements were achieved when the acoustic input data was supplemented with face data, as shown in Figure 6. The improvement from audio only inversion to audio and full 3D motion capture data (*AFQ*) inversion is 50% for the VCV words and 70% for the sentences. However, even a more limited support from the facial data, using the five articulatory measures in the *MV* set gives an important increase, of 35% for the VCV words and 38% for the sentences. Adding the horizontal protrusion of the five markers, as in the *SV* set, does not give any improvement over the *MV* set. This indicates that depth information is not essential for audiovisual-to-articulatory inversion, and that the relevant information can be readily retrieved from monocular frontal face video without 3D reconstruction. On the other hand, the substantially higher correlation for *AFQ* indicates that relevant information in the face is not limited to the mouth region.

5.1.3 Articulatory Analysis

The articulatory analysis focused on what improvement the facial measures can contribute in the articulatory inversion.

Table 1 shows the importance of the audio signal and facial data for different articulatory parameters. The facial measures (*MVO*) provided the most information to recover the movements of the jaw (JH) and of the tongue tip raising (TT), while the audio (*AO*) contributed the most to the horizontal position of the tongue tip (TA). The largest synergetic increase gained by combining the two sources was for the front-back movement of the tongue body (TB) and the velar arching of the tongue (TD).

When grouping the VCV depending on manner of consonant articulation, the largest increase when the facial data was added was for fricatives (64%) followed by nasals (53%). The increase for stops was significantly lower (29%) and for the approximant-tremulant group /l, j, h, r/, the facial data actually decreased the performance (-12%), mainly due to the fact that a linear estimation is unable to handle the combination of a lowered jaw and a raised

tongue tip for /r, l/, since this goes against the general tendency in the corpus that the movement of the jaw and tongue tip are positively correlated.

Figure 7 illustrates some of the general aspects of the *MV* supported linear inversion. The five facial measures improved the estimation of the tongue tip position significantly and permitted to find e.g., alveolar closures (Figure 7a). The estimation of *manner* of articulation was hence improved, even if the *place* of articulation was not always recovered for post-alveolars (Figure 7b). The facial measures were unable to contribute to a better inversion of the tongue tip for articulations for which it was positioned very independently of the jaw, as for /r, l/ (Figure 7c). It was further found that the facial data was a poor estimator of the dorsum part of the tongue (Figure 7d). It even occured that the facial information contributed to a better estimation of the tongue tip position, but at the same time made the recovery of the dorsum part worse (Figure 7e).

5.2 The Linearity of the Mapping Functions

The purpose of this experiment was to investigate whether the audiovisual-to-acoustic inversion could be improved by replacing the linear regressor with a non-linear estimation method. For the experiment, four different functions were learned,

$$\hat{\boldsymbol{\tau}_k}^{Lin,AO} = f^{Lin,AO}(\mathbf{a}_k) \tag{11}$$

$$\hat{\boldsymbol{\tau}_k}^{ANN,AO} = f^{ANN,AO}(\mathbf{a}_k) \tag{12}$$

$$\hat{\boldsymbol{\tau}_k}^{Lin,AV} = f^{Lin,Early,AMV}(\mathbf{a}_k, \mathbf{q}_k^M) \tag{13}$$

$$\hat{\boldsymbol{\tau}_k}^{ANN,AV} = f^{ANN,Early,AMV}(\mathbf{a}_k, \mathbf{q}_k^M) \tag{14}$$

where *Lin* or *ANN* = linear or ANN regression, *AO* = audio cue only, and *AMV* = audio + monovision.

The correlation between these estimates and the original articulatory parameters $\boldsymbol{\tau}_k$ is shown in Figure 8a. The ANN estimation outperforms the linear and the audiovisual inversion outperforms the audio only for both linear and non-linear regression. It can be noted, however, that the linear estimation benefits more from complementarity of the visual information for the VCVs, narrowing the gap to the ANN estimation almost completely, with a 42% increase in correlation for the linear estimation but only a 24% increase for the ANN. For the sentences, this is not the case, as the increase is similar, 59% for the linear and 61% for the ANN.

The RMS error (Figure 8b) gives a slightly different picture. For the sentences, the ANN estimation is slightly better than the linear and the audiovisual

slightly better than the audio only for both types of estimation, which corroborates the results of the correlation coefficients. For the VCV words, however, the RMS error *increases* when visual input is added. This is probably due to overfitting, caused by the increase in the state-space when visual data is added and the fact that the training set is too small for the higher dimensionality. It may hence be advisable to use late fusion of audio and visual data for the non-linear regression, if the training set is small. The problem is further discussed in Section 5.3. Two baselines were calculated for the RMS error. The first is that between the actual and the mean tongue shape over the entire corpus, which is 3.3 mm for the VCV words and 3.2 mm for the sentences. This baseline is a measure of what could be achieved by doing nothing in the regression and simply guessing at the mean shape in each frame. A second baseline was calculated for the VCV words, in a similar manner as described for the correlation coeficients in Section 5.1, by assigning each parameter its mean value for the correct vowel in the vowel frames, while setting each parameter in the consonant frames to its mean value over all consonants. This baseline indicates the error achieved by estimating the vowels correctly, without attempting to reconstruct the consonant articulations. The second baseline is 2.8 mm.

Apart from the difference for the sentence corpus, it would appear that the gain from using a non-linear estimation is only marginal for the audiovisual case, but an analysis of the reconstructed tongue shapes nevertheless gives a different view. The non-linear regression can be considered superior, since the tongue shapes are "closer" to the actual shape from an articulatory point of view, i.e., relevant features, such as place and manner of articulation are estimated better. This is in particular true for alveolars, retroflexes and [l], since the tongue tip raising is correctly reconstructed. The case of [l] is especially troublesome for the linear estimation, due to the non-linear relation between the jaw and tongue tip positions, as noted above. The only exceptions for which the non-linear method is not clearly better are [r], which is as problematic for both methods, and [k], where the ANN estimation successfully reconstructs a velar consonant, but without the stop closure, hence making the tongue shape resemble the fricative [ʃ].

5.3 Markerless Feature Extraction

Ultimately, one would of course like the visual input to the inversion to be markerless, since markers compromise the usability of a system employing the inversion method. In this experiment, we evaluated the markerless features described in Section 3.4 for inversion. All the experiments in this section were performed using the 25 Hz VCV dataset. For these experiments, 8 different inversion functions were learned,

$$\hat{\boldsymbol{\tau}_k}^{Lin,AO} = f^{Lin,AO}(\mathbf{a}_k) \quad (15)$$
$$\hat{\boldsymbol{\tau}_k}^{ANN,AO} = f^{ANN,AO}(\mathbf{a}_k) \quad (16)$$
$$\hat{\boldsymbol{\tau}_k}^{Lin,VO} = f^{Lin,VO}(\mathbf{v}_k) \quad (17)$$
$$\hat{\boldsymbol{\tau}_k}^{ANN,VO} = f^{ANN,VO}(\mathbf{v}_k) \quad (18)$$
$$\hat{\boldsymbol{\tau}_k}^{Lin,Early,AV} = f^{Lin,Early,AV}(\mathbf{a}_k, \mathbf{v}_k) \quad (19)$$
$$\hat{\boldsymbol{\tau}_k}^{ANN,Early,AV} = f^{ANN,Early,AV}(\mathbf{a}_k, \mathbf{v}_k) \quad (20)$$
$$\hat{\boldsymbol{\tau}_k}^{Lin,Late,AV} = f^{Lin,Late,AV}(\mathbf{a}_k, \mathbf{v}_k) \quad (21)$$
$$\hat{\boldsymbol{\tau}_k}^{ANN,Late,AV} = f^{ANN,Late,AV}(\mathbf{a}_k, \mathbf{v}_k) \quad (22)$$

where *Lin* or *ANN* = linear or ANN regression, *Early* or *Late* = early or late fusion, *AO* = audio cue only, *AO* = markerless video cues only and *AV* = audio + markerless video.

Figure 9 shows the correlation and RMS error between the true and reconstructed articulatory parameters.

The *AO* inversion performance is lower here than in the previous experiments with the VCV dataset, i.e., the correlation is lower and the RMS error is higher. The reason is most probably that the 25 Hz VCV dataset is a factor 2.4 smaller than the 60 Hz VCV; parameter combinations occuring only for a very short time have then been averaged out and are not present in the training data. This means that the performance of the audiovisual inversion can not be compared in absolute terms to the *AMV* performance in the previous sections. Instead, we compare relative improvements from the respective *AO* inversion results.

With this measure, it can be concluded that using linear regression, the improvement in correlation when adding the visual cue (44%) exceeds the improvement of *AMV* over *AO* (Figure 8, 40%) and is almost at par with the improvement when adding full motion capture data to the recognition (Figure 6, 50%). From this, we draw the conclusion that our proposed method for extracting information from face video is effective for articulatory inversion.

In the ANN case, the improvement in correlation when adding the visual cue (20%) is slightly lower than that of *AMV* over *AO* (Figure 8, 24%). Given the better performance of the linear estimators, this could have to do with the fact that the dataset is smaller, and at the same time, the dimensionality of the data is higher in the *AV* case (57 dimensions) than in the *AMV* case (23 dimensions), leading to a larger state-space. This causes the ANN to overfit to the specific training data to a higher degree. Heuristically, the training data simply fail to cover a state space of too high dimensionality.

Furthermore, the RMS error for *Early, AV* is higher in the ANN case than in the linear case. The correlation for *Early AV* was also lower than the corre-

lation for *Late AV* in the ANN case, while it is the other way around in the linear case. Again, this was probably caused by ANN overfitting.

A conclusion to be drawn from this is that more training data probably would lead to a significant improvement in *Early AV* performance in the ANN case.

5.3.1 Articulatory Analysis

Table 2 shows the importance of the audio and video signals for different articulatory parameters. The results in the linear case are consistent with those in the VCV case in Table 1. This supports the conclusion that the parameters extracted from the video contains, not only as much information as the 3D marker data, but *the same type* of data.

Interestingly, except for the tongue advance (TA), the video makes a larger contribution than the acoustic signal, even for tongue positions further back, which are often considered impossible to lipread.

In the ANN case, the *Early AV* correlation levels are lower, again probably due to overfitting. Most notably, the estimation of tongue advance (TA) actually is *worse* than in the *AO* case.

The parameter variations throughout three different VCV words shown in Figure 10 (linear estimation) and Figure 11 (ANN) indicate that the qualitative improvement when adding visual information is similar for the linear and non-linear estimations. The tongue tip position (TT) is better estimated both for the closure in [t] and the lowering in [k], which is in accordance with the estimation from motion capture data in Figure 7. The reconstruction of the velar closure in [k] is further improved through a better estimation of the parameters controlling the back part of the upper tongue (TB and TD).

Figure 11 also hints at a possible cause for the higher RMS error for the ANN estimation, since its output is less smooth with sudden changes, which may cause large errors in occasional frames. This is a result of the fact that the current estimation is made on separate frames without any continuity constraints on the parameter trajectories. Further improvement of the method should hence include criteria to take adjacent frames into account. This is also discussed in Section 6.

It should further be noted from the parameter trajectories that the surrounding vowels, which occur much more frequently in the corpus than the consonants, are not better reconstructed than the middle consonant. The differences between the actual and the estimated parameter trajectories are evenly distributed through the word.

6 Discussion & Conclusions

We have investigated how a discriminative inversion procedure benefits from the introduction of visual cues along with the acoustic information.

The study showed that improvements on the order of 35% can be achieved in acoustic-to-articulatory inversion by adding a small number of articulatory measures of the subject's face; achieved using 3D motion capture, but selected to mimick the information that could be extracted from markerless video. These results were confirmed in experiments with features actually extracted from video.

The analysis further indicated that the gain in audiovisual-to-articulatory inversion when going from monovision (full frontal video images) to stereovision (frontal and profile images) is very small, at least for the measures extracted in this study. Articulatory inversion supported by monovision would hence be as effective, even if stereovision may have additional benefits in terms of robustness.

We also compared linear and non-linear regression and found that a non-linear estimation using neural networks is superior to a linear estimation, but that the difference in reconstruction quality becomes much smaller if visual data is added. If a linear estimation is used, our results indicate that early and late fusion of the different modalities perform similarly. If, on the other hand, a non-linear estimation is used, the best results may be achieved by processing the audio and video streams separately and subsequently fuse the inversion estimations.

As with speech recognition, temporal dependencies in both the acoustic, visual and articulatory parameters can be employed to improve the estimation. Temporal smoothness constraints can be represented using, e.g., multi-stream Hidden Markov Models. Using this approach, Katsamanis et al. (2008) report a 60% correlation between true and reconstructed tongue shape using audio only, a 7% improvement from our results on the same data.

As the relevant visual features can be extracted without markers from frontal images from one videocamera, the findings are important for applications that need a simple audiovisual-to-articulatory inversion technique, e.g., articulatory phonetics training for second language learners or hearing-impaired persons. An illustrative example of this is the different Swedish fricatives [s, ʂ, ɕ, ʃj] which are difficult for both non-native and hearing-impaired speakers. Pronunciation errors of these fricatives may be difficult to detect automatically from the acoustic signal only, but they are more easily found from the tongue shapes reconstructed from audiovisual data, since the places of articulation differ more than the acoustic features. We have found in a previous study

(Engwall, 2006) that the estimation of the place of articulation for the fricatives improves drastically if the MV data is used compared to audio only: from 35% to 70% for [s], 28% to 56% for [ɕ] and from 10% to 57% for [ʃj] for the linear estimation and from 81% to 85% for [s], 33% to 56% for [ɕ] and from 23% to 70% for [ʃj] for the non-linear estimation.

Acknowledgements

This research is part of the two projects ARTUR, funded by the Swedish Research Council, and ASPI, Audiovisual SPeech Inversion. The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324

References

Ahlberg, J.: 2001, Candide-3 – an updated parameterized face, *Technical report*, Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linkping University, Sweden.

Bailly, G. and Badin, P.: 2002, Seeing the tongue from outside, *International Conference on Spoken Language Processing*, pp. 1913–1916.

Beskow, J., Engwall, O. and Granström, B.: 2003, Resynthesis of facial and intraoral motion from simultaneous measurements, *ICPhS*, pp. 431–434.

Branderud, P.: 1985, Movetrack – a movement tracking system, *Proc of the French-Swedish Symposium on Speech, Grenoble*, pp. 113–122.

Bregler, C. and Konig, Y.: 1994, "Eigenlips" for robust speech recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 669–672.

Cootes, T. F., Edwards, G. J. and Taylor, C. J.: 2001, Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(6), 681–685.

Cristianini, N. and Shawe-Taylor, J.: 2000, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK.

Doucet, A., de Freitas, N. and Gordon, N. (eds): 2001, *Sequential Monte Carlo Methods in Practice*, Springer Verlag, New York, NY, USA.

Dupont, S. and Luettin, J.: 2000, Audio-visual speech modeling for continuous speech recognition, *IEEE Transactions on Multimedia* **2**(3), 141–151.

Engwall, O., Bälter, O., Öster, A.-M. and Kjellström, H.: 2006, Designing the human-machine interface of the computer-based speech training system ARTUR based on early user tests, *Behaviour and Information Technology* **25**(4), 353–365.

Engwall, O. and Beskow, J.: 2003, Resynthesis of 3D tongue movements from facial data, *Eurospeech*, pp. 2261–2264.

Engwall, O.: 2003, Combining MRI, EMA & EPG in a three-dimensional tongue model, *Speech Communication* **41**(2–3), 303–329.

Engwall, O.: 2005, Introducing visual cues in acoustic-to-articulatory inversion, *Interspeech*, pp. 3205–3208.

Engwall, O.: 2006, Evaluation of speech inversion using an articulatory classifier, *International Seminar on Speech Production*, pp. 469–476.

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

Hyvärinen, A.: 2005, The FastICA package for Matlab, version 2.4, www.cis.hut.fi/projects/ica/fastica.

Isard, M. and Blake, A.: 1998, Condensation – conditional density propagation for visual tracking, *International Journal of Computer Vision* **29**(1), 5–28.

Jiang, J., Alwan, A., Keating, P. A., Auer, E. T. and Bernstein, L. E.: 2002, On the relationship between facial movements, tongue movements, and speech acoustics, *EURASIP Journal on Applied Signal Processing* **11**, 1174–1188.

Katsamanis, A., Papandreou, G. and Maragos, P.: 2008, Audiovisual-to-articulatory speech inversion using active appearance models for the face and hidden markov models for the dynamics, *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Kaucic, R. and Blake, A.: 1998, Accurate, real-time, unadorned lip tracking, *IEEE International Conference on Computer Vision*, pp. 370–375.

Kjellström, H., Engwall, O. and Bälter, O.: 2006, Reconstructing tongue movements from audio and video, *Interspeech*, pp. 2238–2241.

MacLeod, A. and Summerfield, Q.: 1990, A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise. Rationale, evaluation and recommendations for use, *British Journal of Audiology* **24**, 29–43.

Maeda, S. (ed.): 1994, *SpeechMaps, WP2 - From speech signal to vocal tract geometry*, Vol. III.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S. and Harvey, R.: 2002, Extraction of visual features for lipreading, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(2), 198–213.

Ouni, S. and Laprie, Y.: 2002, Introduction of constraints in an acoustic-to-articulatory inversion, *International Conference on Spoken Language Processing*, pp. 2301–2304.

Petersen, M.: 1997, *Optimization of Recurrent Neural Networks for Time Series Modeling*, PhD thesis, Technical University of Denmark, Denmark.

Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J. and Darrell, T.: 2005, Visual speech recognition with loosely synchronized feature streams, *IEEE International Conference on Computer Vision*, pp. 1424–1431.

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

Shdaifat, I. and Grigat, R.-R.: 2005, A system for audio-visual speech recognition, *Interspeech*, pp. 1221–1224.

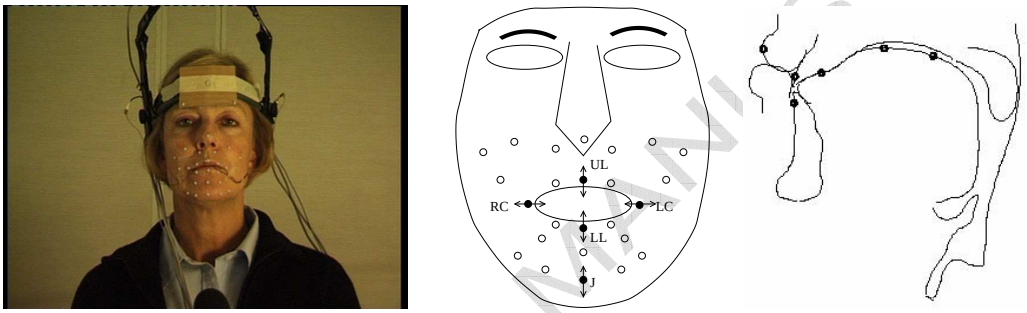
Sugamura, N. and Itakura, F.: 1986, Speech analysis and synthesis methods developed at ECL in NTT, *Speech Communication* **5**(2), 199–215.

Tipping, M. E.: 2001, Sparse Bayesian learning and the relevance vector machine, *Journal of Machine Learning Research* **2001**(1), 211–244.

Torresani, L. and Hertzmann, A.: 2004, Automatic non-rigid 3d modeling from video, *European Conference on Computer Vision*, pp. 299–312.

Yang, M.-H., Kriegman, D. J. and Ahuja, N.: 2002, Detecting faces in images: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(1), 34–58.

Yehia, H., Rubin, P. and Vatikiotis-Bateson, E.: 1998, Quantitative association of vocal-tract and facial behaviour, *Speech Communication* **26**, 23–43.



(a) Face video (b) Face marker positions (c) EMA coil positions

Fig. 1. Placement of EMA and motion capture markers for data collection. (a) Frontal face video used for markerless face feature extraction. (b) Position of markers used for 3D motion capture of points in the face. (c) Position of EMA coils used for estimation of the articulatory configuration.

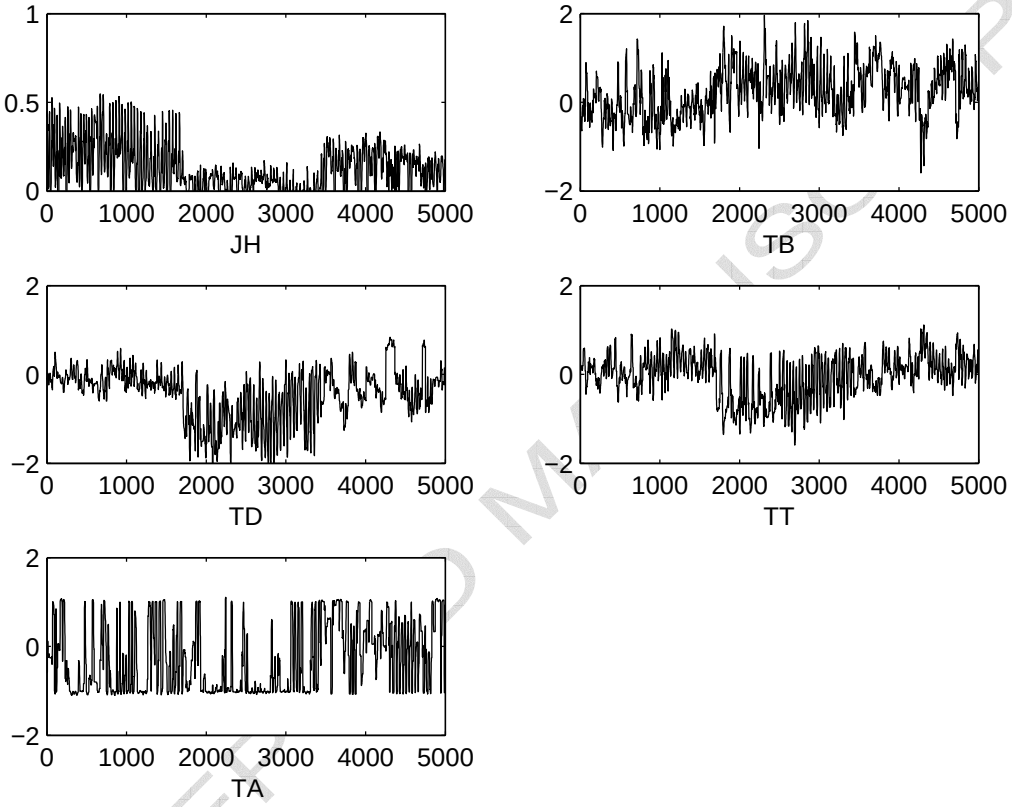


Fig. 2. Variation of parameter values over the frames of the VCV corpus, with JH=Jaw Height, TB= Tongue Body raise, TD=Tongue Dorsum raise, TT=Tongue Tip raise and TA=Tongue tip Advance. The vowel context is /a/ for the first third of the corpus, /ʊ/ for the second and /ɪ/ for the last.

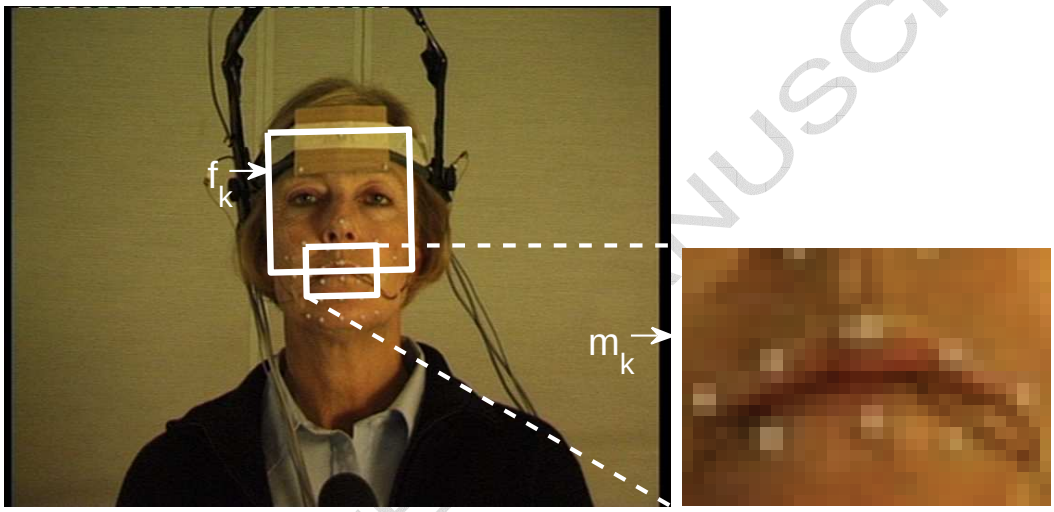
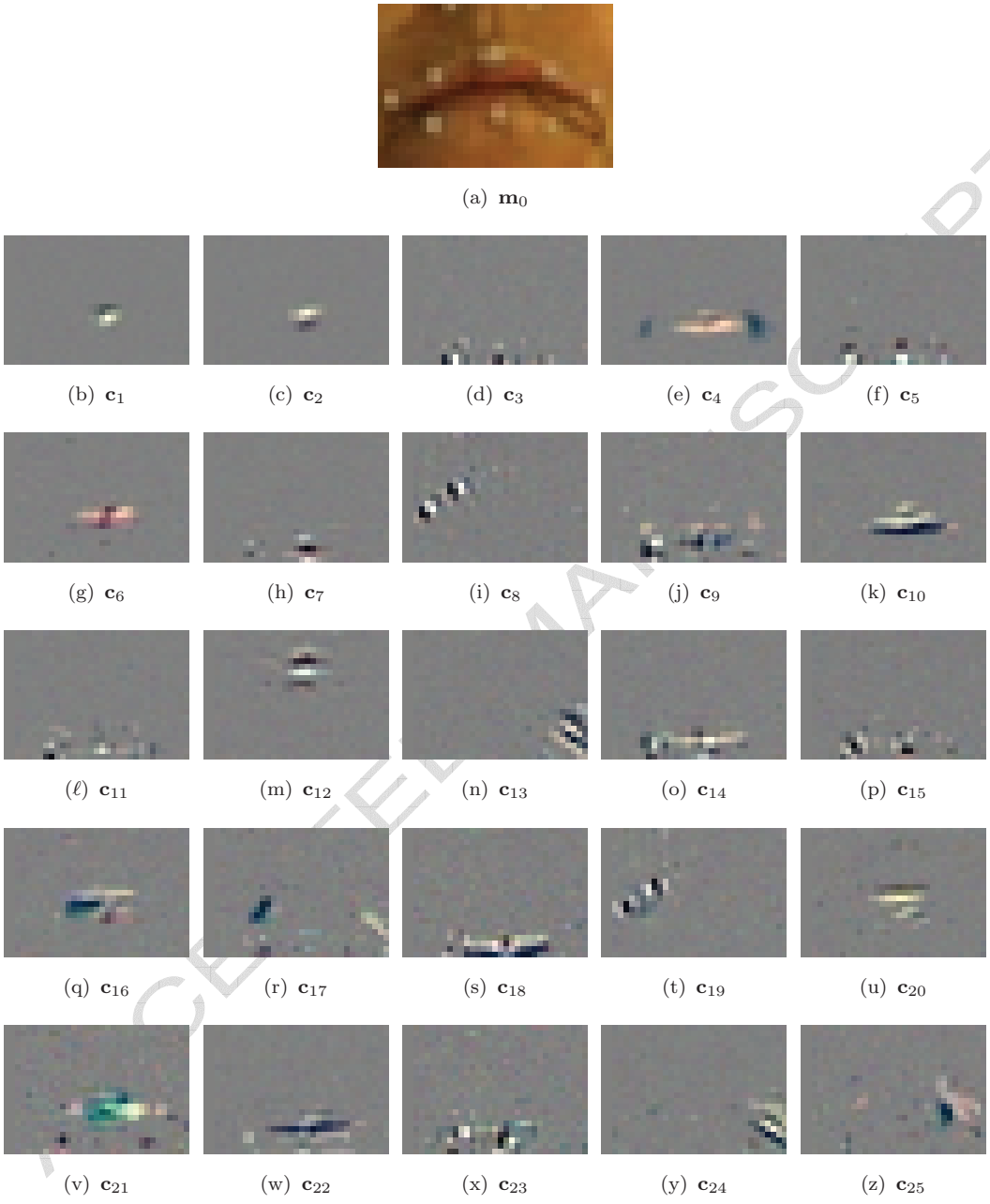


Fig. 3. By tracking the upper part of the face through a sequence, a stabilized sequence of the mouth region can be extracted. The upper part of the face undergoes relatively small appearance changes, the largest being eye blinks. This makes it more advantageous to track than the lip region itself.



(a) $\mathbf{m}_0$

(b) $\mathbf{c}_1$ (c) $\mathbf{c}_2$ (d) $\mathbf{c}_3$ (e) $\mathbf{c}_4$ (f) $\mathbf{c}_5$

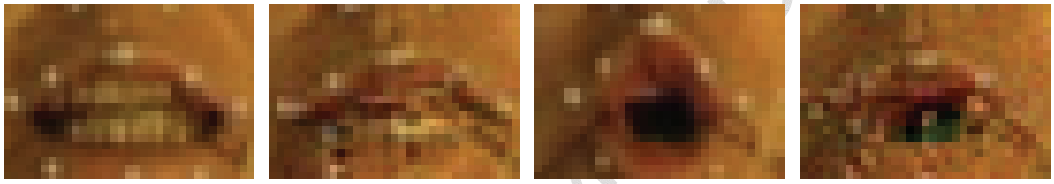
(g) $\mathbf{c}_6$ (h) $\mathbf{c}_7$ (i) $\mathbf{c}_8$ (j) $\mathbf{c}_9$ (k) $\mathbf{c}_{10}$

(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) $\mathbf{c}_{25}$

Fig. 4. 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 $\mathbf{x}_k = \mathbf{m}_k - \mathbf{m}_0$.



(a) Original /ɪ/ (b) ICA reconstr. of /ɪ/ (c) Original /ʉ̟ː/ (d) ICA reconstr. of /ʉ̟ː/

Fig. 5. 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 /ʉ̟ː/. (d) Reconstruction of the same frame $\mathbf{m}_0 + \Sigma_{i=1}^{n} v_{k,i}\mathbf{c}_i$, using $n = 50$.

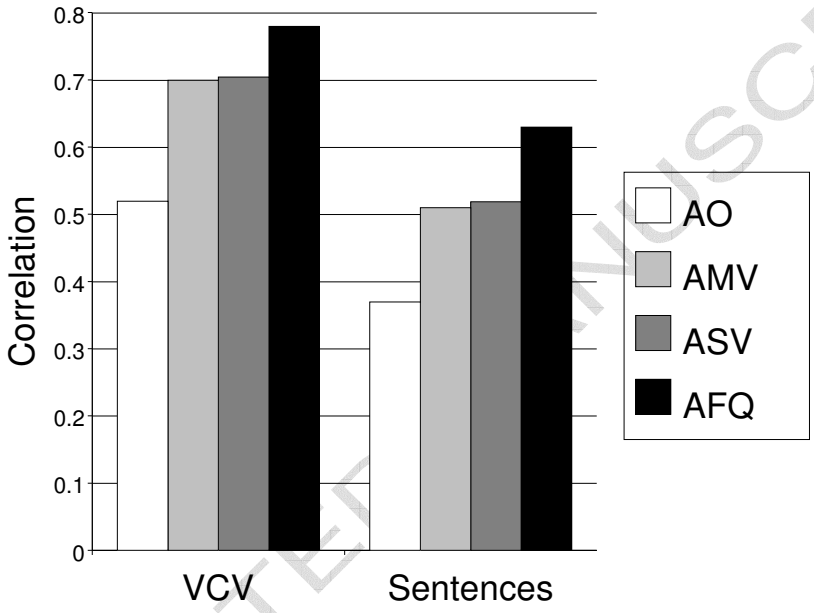
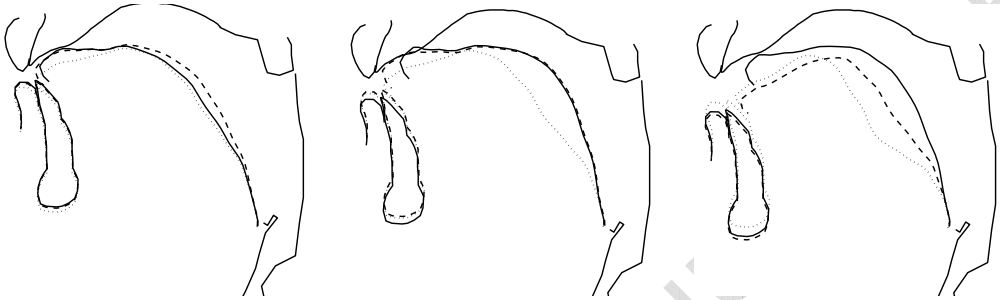
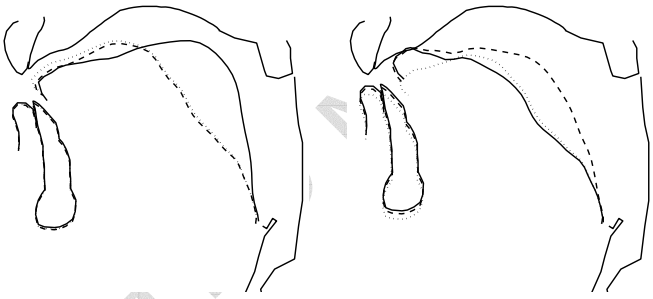


Fig. 6. Correlation coefficients for the articulatory parameters when the inversion was made from audio only (AO), from audio and five articulatory measures (AMV), audio and ten articulatory measures (ASV) and from audio and the entire 3D Qualisys data (AFQ).

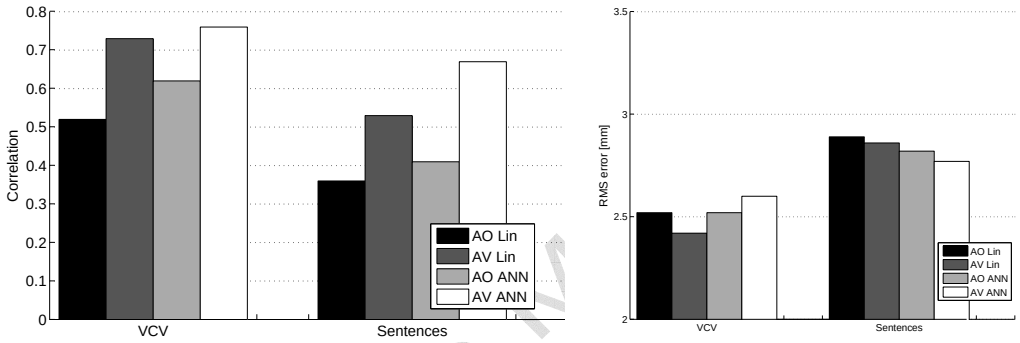


(a) /t/ in “atta” (b) /t̪/ in “irti” (c) /l/ in “alla”



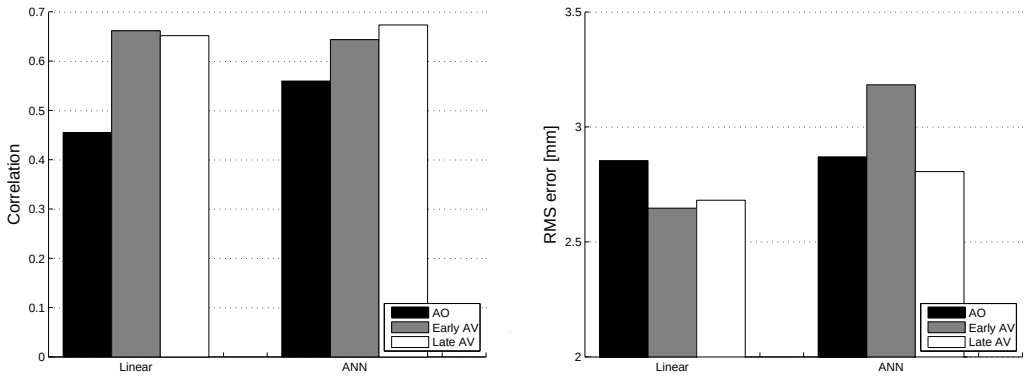
(d) j → k in “ijki” (e) /n/ in “anna”

Fig. 7. Midsagittal contours reconstructed from EMA measures (solid line), estimated from audio only (dotted line) and from the audiovisual AMV case (dashed line). Only the articulatory parameters of the tongue were applied, which means that the upper lip, palate, velum, pharyngeal wall and larynx are considered as fixed.



(a) Correlation (b) RMS error

Fig. 8. The correlation between estimated and actual parameter values and the RMS error for the estimated tongue contours compared to the original. Comparisons are made for audio only (*AO*) and early fusion of audio and sparse motion capture data (*Early AV*), using the linear (Lin) and neural networks (ANN) estimations.



(a) Correlation

(b) RMS error

Fig. 9. The correlation between estimated and actual parameter values and the RMS error for the estimated tongue contours compared to the original, for the 25 Hz VCV dataset. Comparisons are made for audio only (AO), early fusion of audio and video ($Early\ AV$), and late fusion of audio and video ($Late\ AV$), using the linear and neural networks (ANN) estimations.

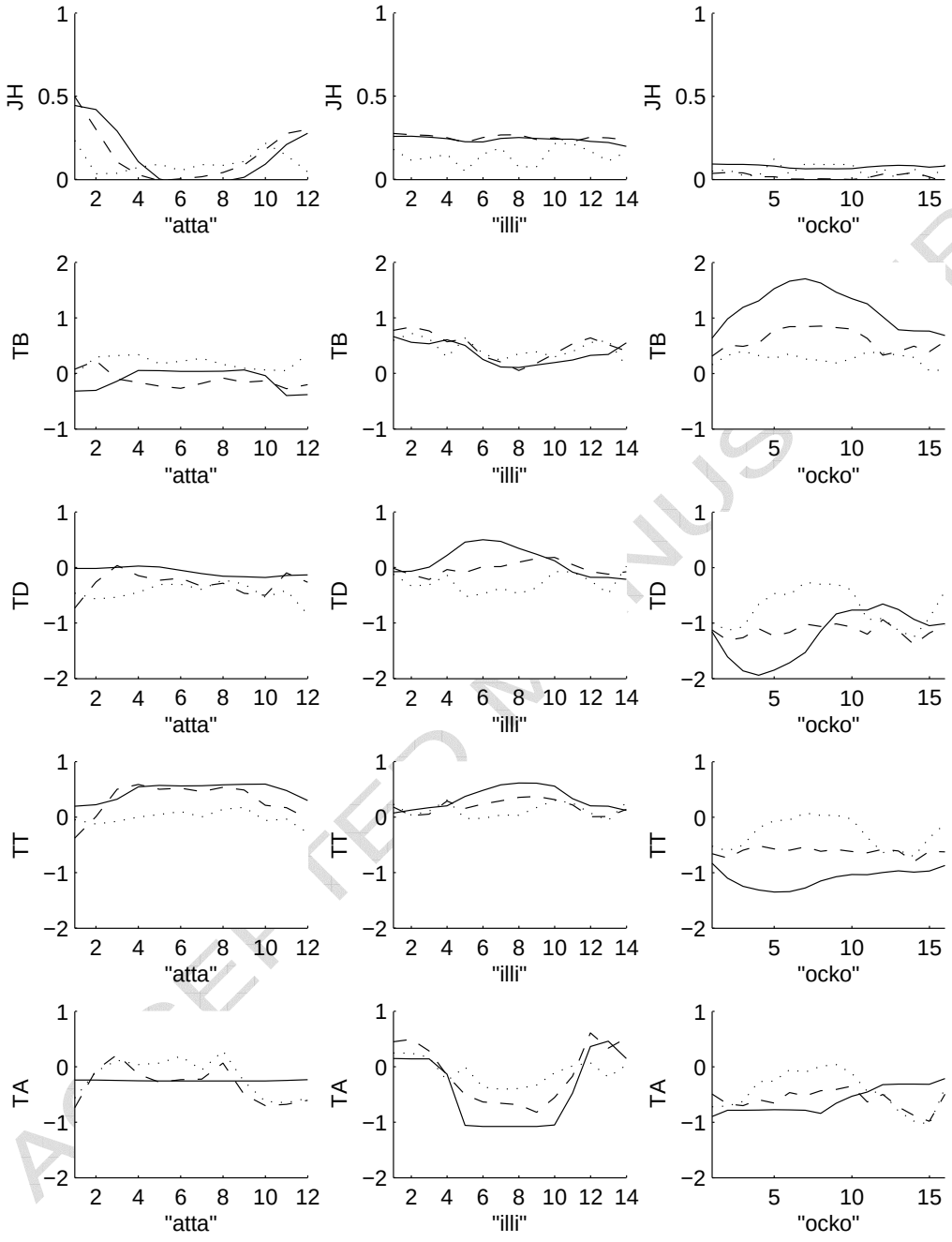


Fig. 10. Articulatory parameters for a number of consecutive time frames over the course of one VCV word, computed from EMA measures (solid line), estimated from audio only (dotted line) and from AV early fusion (dashed line), using the linear regressor.

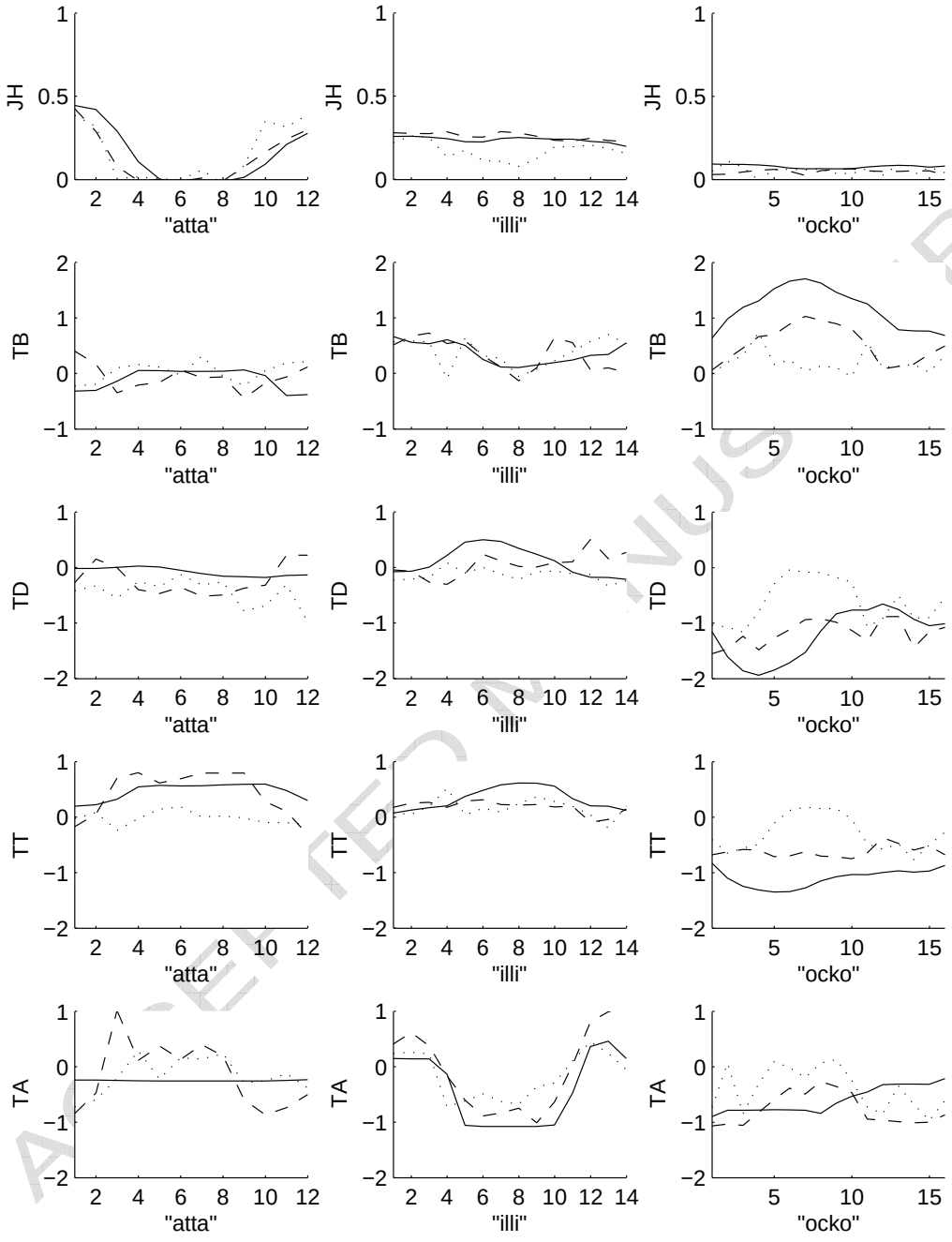


Fig. 11. Articulatory parameters for a number of consecutive time frames over the course of one VCV word, computed from EMA measures (solid line), estimated from audio only (dotted line) and from AV early fusion (dashed line), using the ANN regressor.

Table 1
Correlation coefficients for the articulatory parameters in the inversion using audio only (AO) and audio plus five face measures (AMV). The result is compared to the estimation from only the five measures (MVO). A linear estimator is used in all cases. Braces indicate that either the acoustic or the facial data had a dominant influence.

	JH	TB	TD	TT	TA
VCV AO	0.62	0.45	0.55	0.54	$\lceil 0.62 \rceil$
VCV AMV	$\lceil 0.94 \rceil$	0.58	0.68	$\lceil 0.72 \rceil$	$\lfloor 0.66 \rfloor$
VCV MVO	$\lfloor 0.93 \rfloor$	0.47	0.62	$\lfloor 0.67 \rfloor$	0.52
Sent AO	0.45	0.31	0.32	0.33	$\lceil 0.50 \rceil$
Sent AMV	$\lceil 0.85 \rceil$	0.40	0.40	0.45	$\lfloor 0.55 \rfloor$
Sent MVO	$\lfloor 0.80 \rfloor$	0.24	0.30	0.38	0.37

Table 2
Correlation coefficients for the articulatory parameters in the inversion using audio only (AO) and audio and video (*Early AV*). The result is compared to the estimation from only the video parameters (VO). The 25 Hz VCV dataset is used in all cases. Braces indicate that either the acoustic or the facial data had a dominant influence.

	JH	TB	TD	TT	TA
Linear, AO	0.44	0.31	0.47	0.47	⌈0.57⌉
Linear, *Early AV*	⌈0.92⌉	0.50	0.68	⌈0.72⌉	⌊0.61⌋
Linear VO	⌊0.92⌋	0.47	0.62	⌊0.68⌋	0.51
ANN, AO	0.73	0.36	0.53	0.61	0.61
ANN, *Early AV*	⌈0.92⌉	⌈0.46⌉	0.64	⌈0.71⌉	0.56
ANN, VO	⌊0.90⌋	⌊0.45⌋	0.59	⌊0.69⌋	0.54