Inversion from Audiovisual Speech to Articulatory Information by Exploiting Multimodal Data

A. Katsamanis[1], A. Roussos[1], P. Maragos[1], M. Aron[2], M.-O. Berger[2]

[1] National Technical University of Athens [2]LORIA/INRIA Nancy Grand-Est

E-mail: {nkatsam,troussos,maragos}@cs.ntua.gr {aron,berger}@loria.fr

Abstract

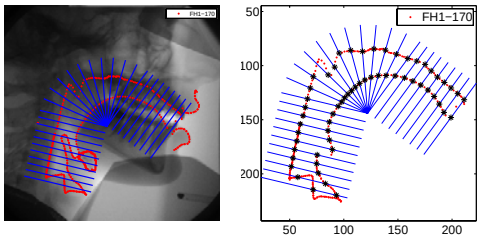
We present an inversion framework to identify speech production properties from audiovisual information. Our system is built on a multimodal articulatory dataset comprising ultrasound, X-ray, magnetic resonance images, electromagnetic articulography data as well as audio and stereovisual recordings of the speaker. Visual information is captured via stereovision while the vocal tract state is represented by a properly trained articulatory model. The audiovisual-to-articulation relationship is approximated by an adaptive piecewise linear mapping. The presented system can recover the hidden vocal tract shapes and may serve as a basis for a more widely applicable inversion setup.

1 Introduction

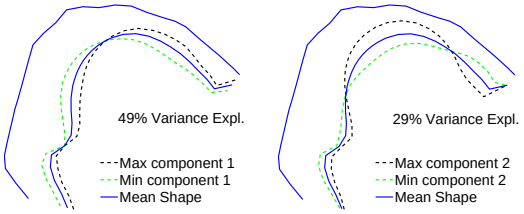
From an engineering viewpoint, inversion from speech to articulatory information refers to the problem of identification of the underlying speech production system given the observed output. For the development of an inversion setup, three primary design issues have to be addressed, namely the choice of the speech representation, the description of the production system and the adoption of a proper computational framework. Related decisions are typically motivated by speech production theory, the nature and amount of the available data and the specific application goals which the inversion scheme will have to serve. In this context, we build on available multimodal articulatory data and present an inversion framework to recover articulation from audiovisual speech information. Speech is represented by both audio spectral features and visual cues from the speaker's face. Articulation is described by means of an appropriate articulatory model and the speech-to-articulation mapping is approximated in an adaptive piecewise linear manner.

Speech inversion has been traditionally regarded as the determination of the vocal tract (VT) shape from the audio speech signal only [3, 8, 7]. Formant values, Line Spectral Frequencies or Mel-Frequency Cepstral Coefficients (MFCCs) are alternative acoustic parameterizations that have been used. Introduction of the visual modality in the speech inversion process can significantly improve inversion accuracy [5, 4]. Independent component analysis of the region around the lips [5] or active appearance modeling of the face [4] provide the practical means to achieve this.

Regarding the representation of the vocal tract, several options have been proposed, each satisfying certain requirements. For example, the tubelet model in [9] allows inversion from formants based on the linear speech production theory. It is quite restrictive though and does not easily account for sounds other than vowels. The representation via the coordinates of points on significant articulators as used in [3] is more realistic but spatially sparse and not so informative for the entire VT state. However, such data can be relatively easily acquired via Electromagnetic Articulography (EMA) and have allowed the application of data-driven techniques for inversion. A much more informative representation is the one achieved via an articulatory model [8, 7] that can describe the geometry or the cross-sectional area of the vocal tract and is controlled by a limited number of parameters. Such models have been built from real articulatory data, i.e., X-ray [6] or Magnetic Resonance Images (MRIs) [7, 2] of the vocal tract. The amount of corresponding data is however limited and thus such models are not easily usable in



(a) X-ray, contours & grid (b) Intersection Points



(c) 1st Component (d) 2nd Component

Figure 1: Building the articulatory model from the X-ray data; positioning the grid and finding the intersection points with the vocal tract boundary. The first model components are shown after PCA.

a data-driven inversion scenario.

As far as the inversion method is concerned, both model-based [9, 8] and data-driven approaches [3] have been reported. In [8] efficient methods are described for the use of codebooks relating formants and articulatory model parameters. To also exploit dynamic information, a piecewise-linear approximation of the audio-articulatory relation is presented in [3]. Each phoneme is modeled by a context-dependent Hidden Markov Model (HMM) and a separate linear regression mapping is trained at each HMM state between the observed MFCCs and the corresponding articulatory parameters.

Given this setting, in the current paper we propose an audiovisual speech inversion framework built on multimodal articulatory data. Visual information is incorporated in the form of 3D coordinates of markers painted on the speaker's face and recovered via stereovision. The VT shape is represented by an articulatory model that is constructed from manually annotated X-ray data. Inversion is achieved by an HMM-based framework similar to the one presented in [4]. The system has been properly adapted to recover the hidden articulatory model parameters. Experiments were performed on a recently acquired articulatory dataset comprising concurrent audio recordings, stereo-videos of the speaker's face, EMA data and Ultrasound (US) videos of his tongue. To extract proper articulatory parameters from this dataset we fitted the articulatory model to the visible tongue contour part in every US frame. Registration of the X-ray reference system in the US images is achieved by properly exploiting available head MRIs of the speaker and the stereovisual data. Recovered articulatory shapes after inversion closely follow the original ones and demonstrate the potential of the approach.

2 Framework Description

The articulatory dataset on which the framework has been built is described in detail in [1]. It includes audio (44kHz), stereo-videos (120Hz) of the face, US images of the tongue (65Hz) and recordings of EM sensors (40Hz) on the US-probe, the speaker's tongue and head. In the performed experiments, approximately 6 minutes of recordings have been exploited. There is only one French speaker and the uttered corpus includes a wide variety of isolated phonetic sequences (vowel-consonant-vowel or vowel-vowel) and a set of phonetically balanced French sentences. Additionally, 3D MRI data of the speaker's head are included (for 3 sustained vowels). There also exist approximately 700 VT shapes, corresponding to roughly 30 seconds of manually annotated x-ray images of the speaker's vocal tract (25Hz). These contours have been used to train a proper articulatory model.

Articulatory Model The role of the articulatory model is critical in our setup. It describes the midsagittal VT shape and is constructed as in [6]. A semi-polar grid (to which we refer as VT grid) is properly positioned on the midsagittal plane, Fig.1(a), and the coordinates of the intersection points with the vocal tract boundary are found, Fig.1(b). Principal Component Analysis (PCA) of the derived vectors determines the components of a linear model that can describe almost 96% of the shape variance using only 6 parameters. The first two components are shown in Figs.1(c), 1(d) for the maximum and minimum values of the corresponding parameters. The goal is then to fit this model to the US data of the tongue in order to derive an efficient description of the VT shape for the whole dataset. For this purpose, registration of the grid on

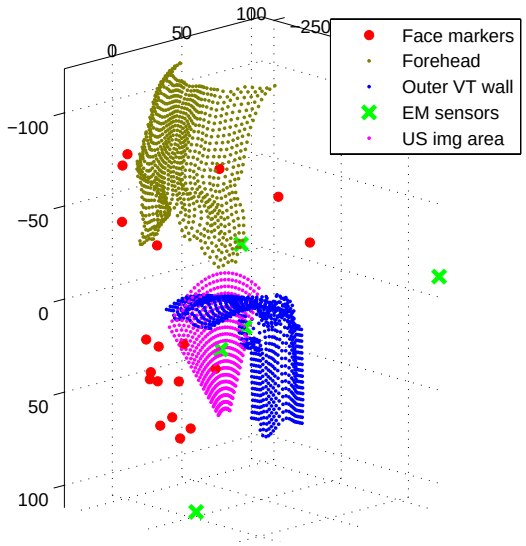


Figure 2: Registration of the multimodal articulatory data for a particular time instance. The HCS is used.

the US images is necessary.

Registration of the Articulatory Data As a preliminary step, preprocessing is applied separately on the different image modalities. The speaker's outer VT wall and forehead surfaces are reconstructed from 3D MRI of his head (see Fig.2). The 3D positions of the painted face markers at every time instance are tracked automatically using the image sequences of the stereo camera pair (see Fig.3). The US image sequences are filtered using the preprocessing method described in [1], which emphasizes the tongue contour (see Fig.4).

The basic steps to combine and exploit the different articulatory data modalities are the following:

Conversion to the EM coordinate system. The 3D positions of three EM sensors (behind the ears and on the probe) at the stereo coordinate system are approximated using the stereo images, at a number of time instances. In this way, we manage to have the same positions expressed in both the stereo and EM sensor coordinate system. So, the coordinate transformation from the one system to the other can be computed by registration.

Compensation of the head movement. A coordinate system whose position and orientation are fixed in time w.r.t. the speaker's head (which we refer to as Head Coordinate System - HCS) is used. The coordinate transformation from the EM system to the HCS is computed by registering the positions of the upper head part markers at every time to the positions of the same markers at a reference time instance. The 3D trajectories of the EM sensors and the rest of the face markers are expressed at the HCS (see Fig.2). Among these, the trajectories of the lip markers are afterwards used for the audiovisual inversion (see Sec.3).

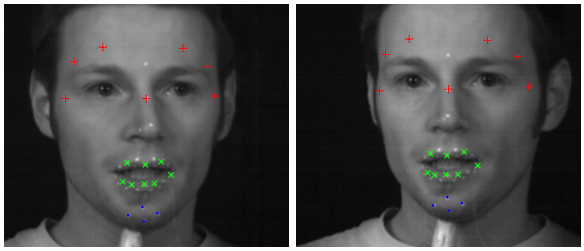


Figure 3: A stereovision image pair, with the tracked face markers. The markers on the upper head part ("+") are used at the registration, whereas the markers on the lips ("×") are used for inversion.

Registration of the VT grid on the US images. The VT semipolar grid is originally configured on the X-ray images and then is positioned in the MRI midsagittal slice by registering the contours of the outer VT wall at the two modalities. This grid is extended to 3D by considering that it is fixed to all MRI slices. Afterwards, it is expressed in the HCS by registering the forehead surface that is extracted from the MRI with the corresponding face markers. Further, the 3D position and orientation (in HCS) of the moving US image plane are recovered using the 6 degrees-of-freedom EM sensor on the US probe, Fig.2. In the end, the intersection of the US image plane with the 3D VT grid is computed for every time instance.

Model fitting to tongue points in US data Using the filtered US images, each grid line is considered to intersect the visible part of the tongue contour only if the maximum image intensity on that line is above a global threshold. If this is the case, the points of the current grid line whose intensities pass a line-specific threshold are kept and finally the closest point to the outer vocal tract boundary is marked as tongue point (Fig.4-Left).

At the last stage of the articulatory parameter extraction, the articulatory model is fitted to best match the coordinates of the tongue points on the US plane. Essentially we minimize the squared distance of the reconstructed shape from these specific points. The missing articulatory parameters are considered to be normally distributed with statistical properties de-

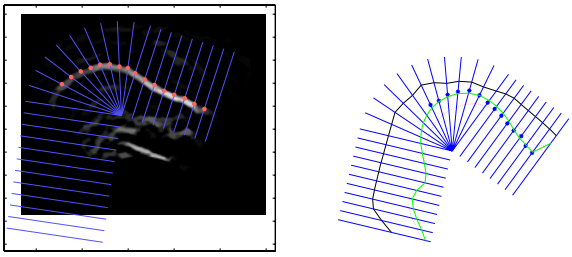


Figure 4: *Left:* Extraction of tongue points (*red dots*) on the VT semipolar grid (*blue lines*), for the same time instance as in Fig. 2. The corresponding preprocesed US frame is used. *Right:* Fitting the articulatory model to tongue points "$*$" determined from an ultrasound image. The green line corresponds to the fitted model while the red line is the mean shape.

termined from the model training data. Considering these distributions as priors, the problem is then solved by Bayesian inference. A result of the applied model fitting process is shown in Fig. 4-Right.

3 Inversion, Results and Discussion

Having extracted articulatory parameters for the whole dataset, as a result of the model fitting process, we then train an HMM-based audiovisual-articulatory mapping similar to the one described in [4]. Acoustic information is represented by 16 MFCCs while visual information is given in the form of the 3D coordinates of 8 markers on the speaker's lips. Phoneme-based multistream HMMs are trained and for each state a linear mapping between the audiovisual and the articulatory parameters is determined. For inversion, the optimal HMM state is first found via Viterbi decoding from the sequence of the audiovisual observations. Then the underlying sequence of articulatory parameters, i.e., VT shapes, is computed as a result of Maximum A Posteriori (MAP) estimation [4, 3].

Recovered VT shapes along with the reference ones are shown in Fig.5, for the phonemes /i/ and /s/ respectively. We are currently working on the introduction of an articulatory synthesizer in the proposed framework in order to further improve the inversion results in an analysis-by-synthesis manner. Detailed evaluation is under way. Though safe conclusions for the quality of the results cannot yet be drawn, it is clear that the main benefit of the proposed framework is that it exploits a rich variety of

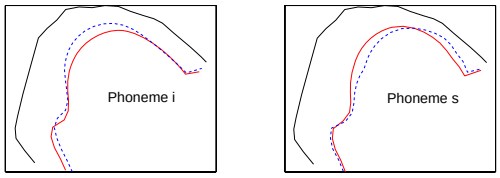


Figure 5: Recovered vocal tract shapes for the phonemes /i/ and /s/. The reference shapes are also given in dashed lines.

multimodal articulatory data, allows improved flexibility and can be more widely applicable.

Acknowledgements The authors would like to thank Y. Laprie and E. Kerrien at Loria for their help in making the articulatory data available and all the ASPI participants for very fruitful discussions. This work was supported by European Community FP6 FET ASPI (contract no. 021324) and partially by grant ΠΕΝΕΔ-2003-ΕΔ866 {co-financed by E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%)}.

References

[1] M. Aron, A. Roussos, M.-O. Berger, E. Kerrien, and P. Maragos. Multimodality acquisition of articulatory data and processing. In *EUSIPCO*, 2008.

[2] P. Badin and A. Serrurier. Three-dimensional linear modeling of tongue: Articulatory data and models. In *Proc. of ISSP*, 2006.

[3] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. on Sp. and Au. Proc.*, 12(2):175–185, March 2004.

[4] A. Katsamanis, G. Papandreou, and P. Maragos. Face active appearance modeling and speech acoustic information to recover articulation. To appear in IEEE Trans. on Au., Sp. and Lang. Proc., 2009.

[5] H. Kjellstrom, O. Engwall. Audiovisual-to-articulatory inversion. to appear in Sp. Comm., 2008.

[6] S. Maeda. *Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model*, chapter in Speech Production and Speech Modeling, pages 131–149 , 1990.

[7] P. Mokhtari, T. Kitamura, H. Takemoto, and K. Honda. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients. *Journ. of Phonetics*, 35(1):20–39, Jan. 2007.

[8] S. Ouni and Y. Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journ. of the Ac. Soc. of America*, 118(1):444–460, 2005.

[9] J. Schoentgen and S. Ciocea. Kinematic formant-to-area mapping. *Sp. Comm.*, 21(4):227–244, 1997.