Resynthesis of 3D tongue movements from facial data

Olov Engwall, Jonas Beskow

Centre for Speech Technology, KTH, SE-100 44 Stockholm, Sweden
{olov, beskow}@speech.kth.se

Abstract

Simultaneous measurements of tongue and facial motion, using a combination of electromagnetic articulography (EMA) and optical motion tracking, are analysed to investigate the possibility to resynthesize the subject's tongue movements with a parametrically controlled 3D model using the facial data only. The recorded material consists of 63 VCV words spoken by one Swedish subject. The tongue movements are resynthesized using a combination of a linear estimation to predict the tongue data from the face and an inversion procedure to determine the articulatory parameters of the model.

1. Introduction

In our work on 3D talking heads, we envisage to create an automatic articulation tutor, who can practise place and manner of articulation with its users, e.g. hearing-impaired children, second language learners or speech therapy patients. The goal is to provide relevant visual feedback allowing to correct deviant articulations. To do so the tutor should be able to contrast the user's own articulation with a correct one, which involves several subtasks: scaling of the articulatory model to correspond to the user, (audiovisual) speech recognition, acoustic-to-articulatory inversion and optimal mode of display.

This paper deals with one aspect of the articulatory inversion, i.e. to what extent facial data can contribute to the recovery of the tongue shape. The relation between the facial and tongue movements has been investigated in earlier studies [1, 2], using a combination of electromagnetic articulography (EMA) and facial motion tracking with Optotrack [1] or Qualisys [2]. Both studies however focused on the overall correlations between the two data sets (non-simultaneous recordings of two English and six Japanese sentences in [1] and simultaneous recordings of 17 CV syllables in [2]) rather than the actual tongue contour. The effects on the tongue contour were indeed studied in [3], using an articulatory model based on non-simultaneous cineoradiographic data of the vocal tract and video recordings of 168 coloured beads on the face. Eventhough an important correlation between the two datasets was found, the information was insufficient to recover the vocal tract constriction.

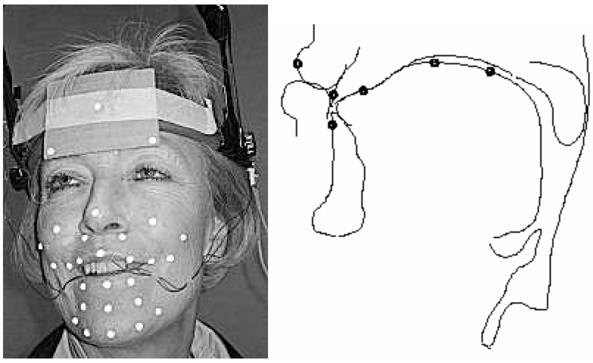
2. Data acquisition

In a real future application the facial data would have to be captured in stereo by one or several video or web cameras, which introduces additional difficulties in extracting the facial data, but to investigate an upper limit of the contribution of facial data, it is here captured using an optical tracking system.

2.1. Experimental setup

The setup consisted of the optical motion tracking system Qualisys [4], the electromagnetic articulograph Movetrack [5] and audio and video recorders.

The Qualisys system uses four infrared cameras to track in stereo 28 small (4 mm diameter, cf. Fig. 1) reflectors at a rate of 60 frames per second. The reflectors were glued to the Movetrack headmount (reference for head movements) and the subject's jaw, cheeks, lips and nose. Six EMA receiver coils were placed on the upper and lower incisor, the upper lip and on the tongue, approximatively 8 (T1), 20 (T2) and 52 mm (T3) from the tip (cf. Fig. 1). The EMA coils on the upper lip and the jaw were equipped with a Qualisys reflector (the latter during a special alignment recording) to allow for spatial alignment between the two data sets.



(a) Qualisys (b) Movetrack

Figure 1: The measurement setup; placement of Qualisys markers and EMA coils.

2.2. Subject and Corpora

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 VCV words with the consonants [p, t, k, b, d, g, f, s, ɕ, ɧ, m, n, ŋ, l, r, ɳ, ʈ, ɖ, v, j, h] in symmetric vowel context with V=[a, ɪ, ʊ].

2.3. Pre-processing

The Qualisys data, consisting of 3D coordinates for all 28 points, was first normalized with respect to global movement using the points on the Movetrack frame as reference. The Movetrack data was down-sampled to the frame rate of the Qualisys data, 60 Hz, and inserted into the subject's midsagittal plane, where it was roto-translated to align the lip and jaw coils with the corresponding Qualisys markers, forming a coherent data set of extra- and intraoral movement data. The silent pauses between the VCVs were removed before the analysis.

3. Predicting tongue shape from the face

Predicting the tongue shape from the face involves three parts. Firstly, an articulatory tongue model is needed (section 3.1). Secondly, an inversion procedure to recover the articulatory parameters of the model from EMA coil positions has to be defined (section 3.2). Thirdly, a training procedure is employed to be able to predict the articulatory parameters from the facial data (section 3.3). Using these estimated articulatory parameters, the tongue movements of the subject can be resynthesized.

3.1. Tongue model

The 3D tongue model, shown in Fig. 2, is based on a three-dimensional MRI database of one reference subject of Swedish [6]. The corpus included 13 Swedish vowels in isolation and 10 consonants in three symmetric VCV contexts with V=[a, ɪ, ʊ]. The tongue parameters consist of jaw height (JH), dorsum raise (TD), body raise (TB), tip raise (TT) and tip advance (TA).

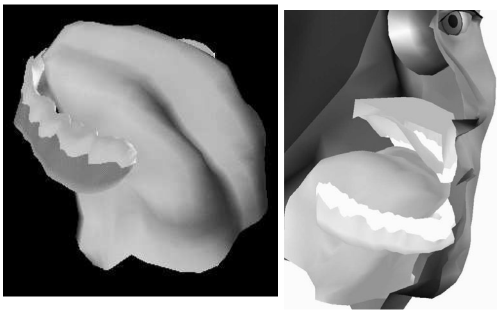


Figure 2: The tongue and jaw model alone and in the frame of a half face.

3.2. Inversion procedure

The articulatory parameters of the tongue model, and hence the tongue shape, are recovered from the positions of the EMA coils using an inversion procedure. The method is similar to the one in [7], where the midsagittal contour of the tongue was reconstructed from three tongue fleshpoints and the jaw height.

The parameter JH, controlling the jaw height, is estimated directly from the lower incisor coil data as

$$JH(t) = \frac{y_{Jaw}(t) - min(y_{Jaw})}{max(y_{Jaw}) - min(y_{Jaw})} \qquad (1)$$

where y_{Jaw} is the vertical position of the lower incisor coil. The parameters TB, TD, TT and TA were determined simultaneously using the Matlab® function fgoalattain, that solves the multiobjective optimization goal attainment problem. The object of the optimisation was 1) to minimise the Euclidean distances between the three EMA coils and the midsagittal tongue contour given by a combination of the articulatory parameters (three goals) and 2) to minimise the difference between the value of TA from the optimisation and that estimated by the position of the first EMA coil, x_{T1}. TA was estimated from x_{T1} as

$$\widetilde{TA}(t) = \frac{x_{T1}(t) - ref_{TA}}{max(x_{T1}) - ref_{TA}} \qquad (2)$$

where ref_{TA} was determined so as to centre $\widetilde{TA}$ on the MRI based model, i.e. assuming that the mean values $\overline{\widetilde{TA}}$ and $\overline{TA}_{MRI}$ are equal, and inverting Eq. 2:

$$ref_{TA} = \frac{\overline{x_{T1}} - max(x_{T1}) \cdot \overline{TA}_{MRI}}{1 - \overline{TA}_{MRI}} \qquad (3)$$

3.3. Linear estimation

The next step is to predict the tongue data, either the EMA coil positions or the articulatory parameters, from the facial data. Following [1] and [2], it is assumed that the the tongue data set can be determined from the face using linear estimators. The face data was arranged in a N-by-75 matrix $\boldsymbol{Y}$, where N is the number of frames in the corpus and each row is a time frame with the x-, y- and z-coordinates of the 25 points on the face, excluding the reference points on the Movetrack headmount. For the EMA data, a similar N-by-8 matrix $\boldsymbol{X}$ was constructed, containing x- and y- coordinates of the 4 Movetrack coils on the tongue and the jaw. Finally, the N-by-5 matrix $\boldsymbol{Z}$ holds the values of the articulatory parameters.

Applying linear regression, unknown EMA data can be estimated as

$$\widetilde{\boldsymbol{X}} = \boldsymbol{T}_{XY} \cdot \boldsymbol{Y}' \qquad (4)$$

where $\boldsymbol{Y}'$ is $\boldsymbol{Y}$ augmented with a column of ones, to allow direct prediction of non-zero-mean vectors, and the

Table 1: Correlation coefficients for the EMA coils and articulatory parameters.

a) Correlation coefficients for the four EMA coils.

	All vowel contexts					[a] context				[ɪ] context				[ʊ] context			
	Jaw	T1	T2	T3	Mean	Jaw	T1	T2	T3	Jaw	T1	T2	T3	Jaw	T1	T2	T3
x	0.98	0.83	0.72	0.69	0.81	1.00	0.78	0.60	0.47	1.00	0.55	0.49	0.31	1.00	0.78	0.74	0.46
y	0.99	0.58	0.35	0.80	0.65	1.00	0.60	0.55	0.36	0.97	0.54	0.51	0.62	0.99	0.63	0.26	0.66

b) Correlation coefficients for the prediction of the articulatory parameters, grouped by place of articulation.

	All VCVs	bilabial [p,b,m]	alveolar [t,d,s,n,l,r]	post-alveolar [ɕ,ɳ,ʈ,ɖ]	velar [k,g,ɧ,ŋ]	other [f,v,j,h]
JH	1.00	1.00	1.00	1.00	1.00	1.00
TB	0.54	0.62	0.45	0.21	0.44	0.60
TD	0.40	0.16	0.39	0.32	0.45	0.62
TT	0.75	0.86	0.79	0.62	0.76	0.79
TA	0.26	0.43	0.20	0.21	0.15	0.15

estimator matrix $\boldsymbol{T}_{XY}$ is given by

$$\boldsymbol{T}_{XY} = \boldsymbol{X} \cdot \boldsymbol{Y}'^{T} \cdot (\boldsymbol{Y}' \cdot \boldsymbol{Y}'^{T})^{-1} \tag{5}$$

A jackknife training procedure was applied, splitting the data into ten parts of which nine determine the estimator $\boldsymbol{T}_{XY}$ (Eq. 5), used to predict the tenth part from Eq. (4). This was repeated so that all parts were used for training and prediction, resulting in an estimated EMA data set ($\widetilde{\boldsymbol{X}}$) that could be compared with the measured ($\boldsymbol{X}$). The same jackknife training procedure was used to predict the articulatory parameters from the facial data.

4. Results

A global measure of the correspondence between the original and predicted EMA data is the correlation coefficients, given in Table 1a). The overall mean correlation is 0.74 for the four coils and 0.66 for the three tongue coils, which is lower than in, but comparable to, earlier studies. Table 1a) shows that the recovery of the positions of the three tongue coils are on average similar to results in [1, 2], with the exception of y_{T2}, which is poorly predicted. The table further indicates that the recovery is dependent on the vowel context, with e.g. substantially higher correlation coefficients for $T1_x$ in [a,ʊ] than in [ɪ] context and for $T3_y$ in [ɪ,ʊ] than in [a] context.

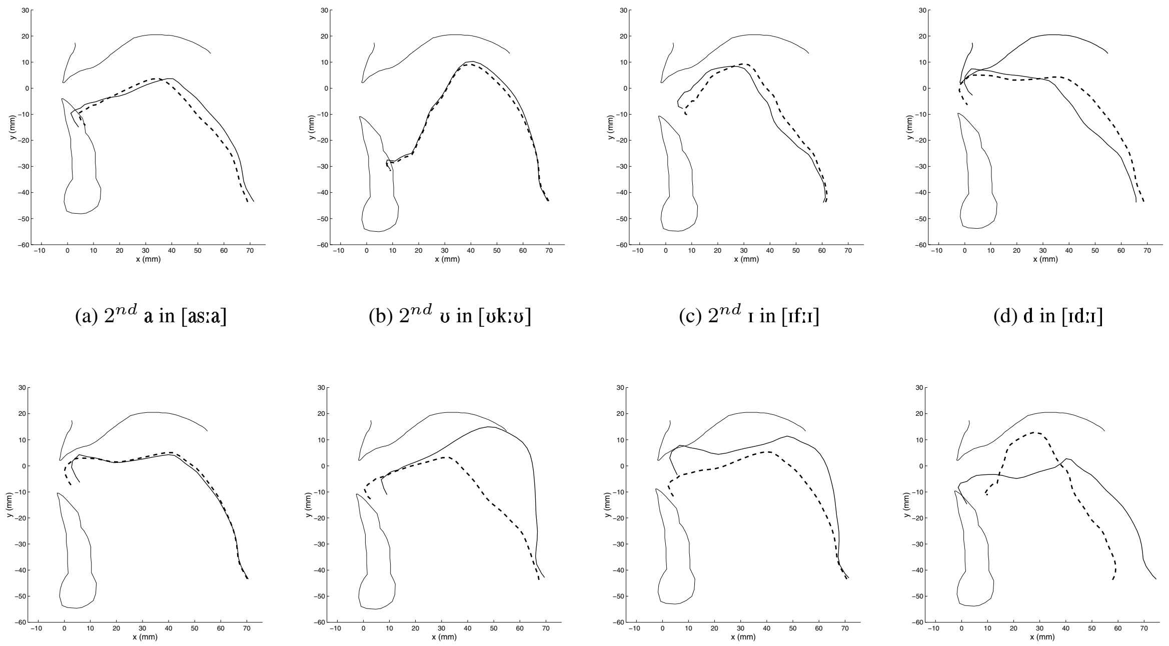
The combination of articulatory inversion followed by linear estimation gives the correlation coefficients in Table 1b). The recovery is not as good as that in [3], with lower correlation coefficients for TB, TD and TA. JH and TT are the best predicted, while TD and TA cannot be recovered from the face. Neither can TB for the current corpus. Table 1b) also gives the correlation coefficients for different VCV groups, based on the place of articulation. Note the large differences in the prediction of different parameters for different articulation groups. Generally, the more neutral the parameter is in the subcorpus, the higher is the correlation coefficient.

The correlation coefficients do however not indicate whether the articulation of the consonant was acheived. The resulting tongue shape and midsagittal contour were hence studied, and the most representative results are summarized in Table 2. The face clearly does not give sufficient information to recover the tongue contour for velars and bilabials. However, contrary to [3], the alveolar stops were successfully recovered, and while the tongue tip raising in [l] was missed, due to the combination of an open jaw and a high tongue tip, which is in conflict with the main correlation between a open jaw and a low tongue tip, the corresponding tongue tip raising was actually predicted for the retroflexes. Concerning the vowels, [a] was generally well predicted, cf. Fig. 3(a), whereas [ɪ,ʊ], Figs. 3(e)-3(c), were sometimes recovered with too low tongue body and dorsum, respectively.

Examples of the 3D resynthesis animations, resulting from EMA measurements as well as from prediction from the facial data are available on
`http://www.speech.kth.se/multimodal/qsmt`

Table 2: Summarized resynthesis results.

Group	Observation	Example
Successful recoveries		
alveolar stops [t, d, n] fricative [s] [r], retroflexes [ʈ ɳ]	Good tongue tip control	Fig. 3(d) Fig. 3(e)
Unsuccessful recoveries		
velar stops [k, g]	No velar closure	Fig. 3(f)
lateral [l]	no tongue tip raising	Fig. 3(g)
bilabials [p, m, b]	no tongue body control	Fig. 3(h)
palatal & velar fricatives [ɕ, ɧ]	No, or too frontward, constriction	



(a) 2^{nd} a in [asːa]

(b) 2^{nd} ʊ in [ʊkːʊ]

(c) 2^{nd} ɪ in [ɪfːɪ]

(d) d in [ɪdːɪ]

(e) s in [ʊsːʊ]

(f) k in [akːa]

(g) l in [ɪlːɪ]

(h) b in [abːa]

Figure 3: Examples of midsagittal tongue contours, reconstructed from EMA (solid lines) and facial data (dashed).

5. Conclusions and future work

The results show that the facial data provides relevant information on the jaw opening and the tongue tip, but that the data is insufficient to accurately predict a non-alveolar vocal tract constriction. Note however that the resynthesis by inversion in this paper was based on the facial data only, whereas it would be based on facial data *and* acoustic-to-articulatory inversion from speech recognition in a future articulation tutor application. Due to the many-to-one mapping of articulation to acoustics, the acoustics cannot be used as the only source of information for the inversion. As concluded above, nor can the facial data. The facial data *does* however reduce the number of candidate articulations in the acoustic-to-articulatory inversion, increasing the possibilities to recover the tongue shape from acoustics and facial data.

6. Acknowledgements

This research is carried out within the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. Olov Engwall is funded by the Wenner-Gren foundation and Hans Werthén-fonden. The authors would like to thank Peter Branderud and Bertil Lyberg for giving access to the Movetrack and Qualisys systems and the enduring subject, Anne-Marie Öster.

7. References

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

[2] J. Jiang, J. Alwan, L. Bernstein, P. Keating, and E. Auer, "On the correlation between facial movements, tongue movements and speech acoustics," in *Proc of the* 6^{th} *ICSLP*, vol. 1, 2000, pp. 42–45.

[3] G. Bailly and P. Badin, "Seeing the tongue from outside," in *Proc of the* 6^{th} *ICSLP*, 2002, pp. 1913–1916.

[4] *http://www.qualisys.se/*.

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

[6] O. Engwall, "Tongue talking – studies in intraoral visual speech synthesis," Ph.D. dissertation, KTH, Stockholm, Sweden, 2002.

[7] P. Badin, E. Baricchi, and A. Vilain, "Determining tongue articulation: from discrete fleshpoints to continuous shadow," in *Proc of Eurospeech97*, vol. 1, 1997, pp. 47–50.