PAPER 7.

Beskow, J., Engwall, O. and Granström, B. (submitted3). Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. To appear in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain.

RESYNTHESIS OF FACIAL AND INTRAORAL ARTICULATION FROM SIMULTANEOUS MEASUREMENTS

Jonas Beskow, Olov Engwall and Björn Granström

Centre for Speech Technology, KTH, SE-100 44 Stockholm, Sweden

ABSTRACT

Simultaneous measurements of tongue and facial motion, using a combination of electromagnetic articulography (EMA) and optical motion tracking, are analysed to improve the articulation of an animated talking head and to investigate the correlation between facial and vocal tract movement. The recorded material consists of VCV and CVC words and 270 short everyday sentences spoken by one Swedish subject. The recorded articulatory movements are re-synthesised by a parametrically controlled 3D model of the face and tongue, using a procedure involving minimisation of the error between measurement and model. Using linear estimators, tongue data is predicted from the face and vice versa, and the correlation between measurement and prediction is computed.

1. INTRODUCTION

In our work on three-dimensional models for articulatory and visual speech synthesis at KTH, we have exploited several kinds of data sources. For the externally visible articulators and the facial surface, we have used optical motion tracking. For modelling of the tongue and internal vocal tract, three-dimensional data from magnetic resonance imaging (MRI) and kinematic data from electropalatography (EPG) and electromagnetic articulography (EMA) have been used [1]. While each of these methods in isolation can provide useful information, none yields complete 3D data with good temporal resolution, and they hence need to be combined. This paper reports on simultaneous measurements of vocal tract and facial motion using EMA and optical motion tracking. The data is used to improve and extend the articulation of an animated talking head.

2. PREVIOUS STUDIES

The study differs from previous related studies as it uses simultaneous recordings of a large set of sentences. Yehia et al. [2] used non-simultaneous recordings with Optotrack and EMA of two English and six Japanese sentences to derive quantitative association between the two data sets. Jiang et al. [3] collected the data simultaneously, using Qualisys and EMA, but for CV syllables only and using 17 Qualisys markers. Bailly & Badin [4] studied the correlation between facial and tongue movements using an articulatory model based on video and cineoradiographic recordings. All three studies concluded that information from the face supplies information on the articulation of the speech organs, but Bailly & Badin warned that the information is insufficient to recover the lingual constriction. The most important difference is that none of these studies aim directly at applying the results to articulatory speech synthesis of the face, jaw and the entire tongue.

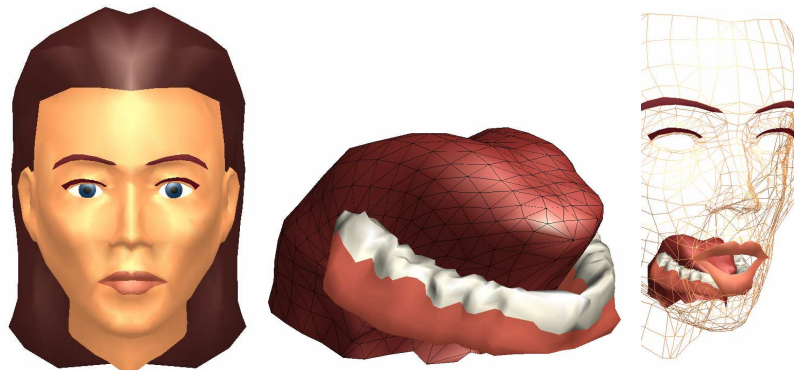

Figure 1. The face model (left), tongue and jaw model (middle) and the combined model (right)

3. MODELS AND DATA

3.1 Face and tongue models

The face and tongue models are based on concepts [5] first introduced by Parke [6] defining a set of parameters that deform a static 3D-wireframe mesh by applying weighted transformations to its vertices. The parameters for the face are *jaw opening*, *jaw shift*, *jaw thrust*, *lip rounding*, *upper lip raise*, *lower lip depression*, *upper lip retraction* and *lower lip retraction*. The 3D tongue model is based on a three-dimensional MRI database of one reference subject of Swedish [1]. The corpus consisted of 13 Swedish vowels and 10 consonants in three symmetric VCV contexts. As the acquisition time of 43 seconds required the subject to artificially sustain the articulations, EPG and EMA data has been used to adjust the articulations to normal and to obtain information on articulatory dynamics. The tongue parameters include *dorsum raise*, *body raise*, *tip raise* and *tip advance*.

3.2 Measurement setup

The EMA data is collected with the Movetrack system [7], using two transmitters on a light-weight head mount and six receiver coils (1.5x4 mm) positioned in the midsagittal plane as depicted in figure 2: three coils on the tongue (around 8 mm, 20 mm and 52 mm from the tip of the tongue) and two coils above and below the upper and lower incisors respectively. One coil was placed on the upper lip for co-registration with the optical system.

The optical motion tracking is done using a Qualisys system [8] with four cameras. The system tracks 28 small reflectors (4 mm diameter) glued to the subject's jaw, cheeks,

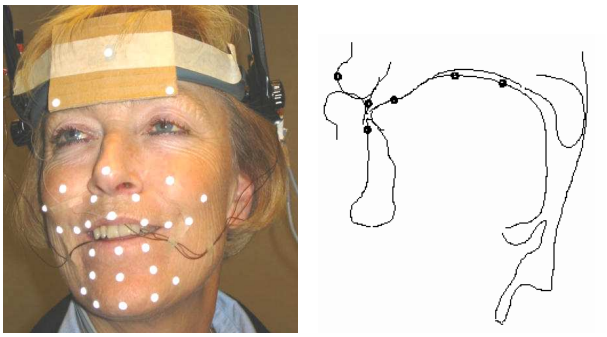

Figure 2 Placement of Qualisys markers (left) and Movetrack coils (right)

lips, nose and eyebrows and the Movetrack headmount (to serve as reference for head movements) and calculates their 3D-coordinates at a rate of 60 frames per second. The EMA coils on the upper lip and the jaw were equipped with a reflector (the latter during a special alignment recording) to allow for spatial alignment between the two data sets. The data was collected in sets of one minute each, with a break between sets. In the analysis below, silent pauses between the speech sequences were removed.

3.3 Subject and corpora

The subject was a female native speaker of Swedish, who has received high intelligibility ratings in audio-visual tests.

3.3.1 Sentence corpus

The 270 Swedish everyday sentences (listed in [9]) have been developed specially for audio-visual speech perception tests by G. Öhngren, based on [10]. The sentences are independent of each other and generally seven to nine syllables long (4-5 words), such as "*Katten lekte med ett nystan*" ("*The cat played with a ball of wool*"). The sentences were articulated clearly.

3.3.2 Nonsense VCV word corpus

The 138 VCV and VCC{C}V words consisted of the consonants [p, t, k, ʈ, b, d, g, ɖ, m, n, ɳ, ŋ, l, ɭ, f, s, ʃ, ç, j, r, v, h] and the consonant clusters [jk, rk, pl, bl, kl, gl, pr, br, kr, gr, kt, nt, tr, dr, st, sp, str, spr, sk, fl, fr, sl, skl, skr] in symmetric vowel context with V=[a, ɪ, ʊ].

3.3.3 Nonsense CVC word corpus

The corpus consisted of 41 asymmetric C_1VC_2 words, with firstly the long vowels V=[uː, oː, ɑː, iː, eː, ɛː, øː] in C_1=[k], C_2=[p] and C_1=[p], C_2=[k] context, secondly the short vowels V=[ʊ, ɔ, a, ɪ, e, ɛ, ʏ, ø] in C_1=[k], C_2=[pː] and C_1=[p], C_2=[kː] context. The [r] allophones V=[æː, œː, æ, œ] were collected with C_1=[k] and C_2=[r].

4. DATA PROCESSING AND MODEL FITTING

4.1 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 facial model was scaled and the Qualisys data was roto-translated in such a way that an optimal fit between the facial surface and the measured points was achieved. The EMA data was down-sampled to the frame rate of the Qualisys data, 60 Hz, and inserted into the midsagittal plane of the model, 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.

4.2 Inner lip point estimation

In order to produce correct re-synthesis of labial articulations, it is important to have information about the inner lip contour. However, labial data in the current set is limited to points along the outer lip contour. Our geometrically based face model is incapable of predicting closure based on these outer points only, since the lips change shape and thickness in a complex way for example during rounded and protruded articulations. Rather than trying to capture these effects in the facial model, two data points $\mathbf{i}_1$ and $\mathbf{i}_2$ (see figure

3) on the inner lip contour are predicted from the outer contour during a separate pre-processing stage. The predicted points are then added to the data set used in the model fitting described below.

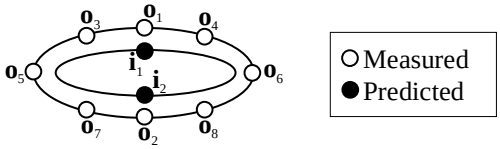

Figure 3. Prediction of inner lip contour from the outer.

Prediction is done in two steps. First, a training set is constructed based on phonetic information, then a linear estimator is trained on this data to do the actual prediction. The training set S consists of all frames in the corpus labelled as bilabial (S_{bilabial}), and the same number of frames[1] randomly selected from all non-bilabial segments ($S_{\text{non-bilabial}}$). For each of these partitions, the positions of $\mathbf{i}_1$ and $\mathbf{i}_2$ are estimated from the corresponding points on the outer contour, $\mathbf{o}_1$ and $\mathbf{o}_2$, as follows:

$$\mathbf{i}_{1i} = \mathbf{i}_{2i} = \frac{\mathbf{o}_{1i} + \mathbf{o}_{2i}}{2} \forall i \in S_{bilabial} \tag{1}$$

$$\begin{aligned} \mathbf{i}_{1i} &= \mathbf{o}_{1i} - \frac{\boldsymbol{\mu}_{o2o1}}{2} \\ \mathbf{i}_{2i} &= \mathbf{o}_{2i} + \frac{\boldsymbol{\mu}_{o2o1}}{2} \end{aligned} \quad \forall i \in S_{non-bilabial} \tag{2}$$

where $\boldsymbol{\mu}_{o2o1} = E[\mathbf{o}_2\text{-}\mathbf{o}_1]$ is the expected value of the vector from $\mathbf{o}_2$ to $\mathbf{o}_1$ taken over the bilabial set, thus representing the thickness of the lips.

The row vectors representing the inner lip points $\mathbf{i}_{1i}$ and $\mathbf{i}_{2i}$ for all i in S are arranged as a N-by-6 matrix $\mathbf{I}$ where N is the number of frames in S.

$$\mathbf{I} = \begin{bmatrix} \vdots & \vdots \\ \mathbf{i}_{1i} & \mathbf{i}_{2i} \\ \vdots & \vdots \end{bmatrix} \quad i \in S \tag{3}$$

A second matrix $\mathbf{O}$ is formed from all the points on the outer lip contour $\mathbf{o}_{ki}$, for k $\in\{1, 2 \ldots 8\}$, and all $i \in S$. The matrix is augmented with an additional column of ones to allow direct prediction of non-zero-mean vectors:

$$\mathbf{O} = \begin{bmatrix} \vdots & & \vdots & \vdots \\ \mathbf{o}_{1i} & \cdots & \mathbf{o}_{8i} & 1 \\ \vdots & & \vdots & \vdots \end{bmatrix} \quad i \in S \tag{4}$$

Using multiple linear regression, we calculate a linear estimator $\mathbf{T}_{\text{OI}}$ that can be used to make a prediction of inner lip points $\mathbf{I}$ from outer $\mathbf{O}$:

$$\tilde{\mathbf{I}} = \mathbf{T}_{\text{OI}} \cdot \mathbf{O} \tag{5}$$

where $\tilde{\mathbf{I}}$ is the least-square estimation of $\mathbf{I}$. The estimator matrix $\mathbf{T}_{\text{OI}}$ is given by

$$\mathbf{T}_{\text{OI}} = \mathbf{I} \cdot \mathbf{O}^{\text{T}} \cdot (\mathbf{O} \cdot \mathbf{O}^{\text{T}})^{-1} \tag{6}$$

[1] There are 2417 bilabial frames in the corpus

Using this estimator, inner lip points are calculated for all frames in the corpus, and appended to the original data set. The motivation for using a linear estimator to generate the new points, rather than predicting them directly using phonetic information (as was done for the training set) is that the latter strategy would result in loss of dynamical information, whereas the linear estimator is capable of retaining the dynamical properties in a natural way.

4.3 Resynthesis

Given the aligned and augmented data set, we want to estimate parameter trajectories to re-synthesize the data with our model of the face and tongue. There is no exact way of determining a single model parameter directly from measured points, since some areas of the face are influenced by more than one parameter. Thus, parameters must be jointly estimated using global optimisation. To carry out such an optimisation, we need a measure of goodness of fit between model and data.

4.3.1 Goodness of fit

Because the basic shape of the facial model does not exactly match the recorded speaker, we need a metric that does not penalize static shape differences. This is solved by defining a *virtual marker* $\mathbf{q}_i$ on the face and tongue of the model for each measured point $\mathbf{p}_i$. The virtual marker is defined in terms of *three* non co-linear model vertices, $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. The three vertices together with the surface normal vector form a coordinate system in which we can express the position of $\mathbf{q}$ in terms of the coordinates s, t and u:

$$\mathbf{q} = \mathbf{v}_1 + s\,(\mathbf{v}_2 - \mathbf{v}_1) + t\,(\mathbf{v}_3 - \mathbf{v}_1) + u\mathbf{e}_1 \tag{7}$$

where $\mathbf{e}_1$ is the normalized vertex normal of vertex $\mathbf{v}_1$. For each $\mathbf{q}_i$, the coordinates (s_i, t_i, u_i) are chosen so that $\mathbf{q}_i = \mathbf{p}_i$ for a given reference frame in the corpus. We used the middle of the [p]-segment in [apa] as the reference frame, since this allowed for a good definition of the bilabial closure and [apa] was included in the MRI data set of tongue shapes. The model was manually adjusted to match the production of this reference segment. The error function to measure goodness of fit between model and data for a the vector of model parameters $\mathbf{y}$ is given by

$$e_{fit}(\mathbf{y}) = \sum_{i=1}^{K} \left| \mathbf{q}_i(\mathbf{y}) - \mathbf{p}_i \right| \tag{8}$$

where K is the number of real and virtual markers. In order to make the parameter estimation more robust and less likely to produce un-physiological tongue shapes, we impose the constraints below on the fitting process.

4.3.2 Tongue volume preservation

To be physiologically correct, the volume of the tongue should remain constant during articulation. The volume enclosed by the polygonal tongue surface is calculated as

$$V = \frac{1}{6} \sum_{i \in P} (\mathbf{v}_{i3} - \mathbf{v}_{i1}) \times (\mathbf{v}_{i2} - \mathbf{v}_{i1}) \cdot \mathbf{v}_0 \tag{9}$$

where P is the set of all triangles that make up the tongue surface; $\mathbf{v}_{i1}$, $\mathbf{v}_{i2}$ and $\mathbf{v}_{i3}$ represent the three vertices of triangle i and $\mathbf{v}_0$ is a vertex at the root of the tongue. Using this formula we calculate the reference volume V_{ref} when the model is in its reference position. During fitting, deviations from the reference volume is penalised by the error term

$$e_{vol}(\mathbf{y}) = \left| V(\mathbf{y}) - V_{ref} \right| \tag{10}$$

As described in [1], the tongue model parameters are based on a linear component analysis of MRI-data. Each parameter y_i has a well defined interval $[a_i, b_i]$ for which it represents a good approximation of observed tongue movements, but values outside of this range can result in un-physiological tongue shapes. Since the subject used in this study differs from the one used for construction of the tongue model, there is a risk of violating the valid parameter intervals during the fitting process. To simply restrict the parameters to the valid intervals could however result in unnatural dynamics. Instead we impose soft limits by adding a penalty term for violation of valid intervals:

$$e_{range}(\mathbf{y}) = \sum_{i \in T} \left[\left(a_i - \min(y_i, a_i)\right)^2 + \left(\max(y_i, b_i) - b_i\right)^2 \right] \tag{11}$$

where T is the set of tongue parameters.

4.3.4 Resulting error function

The global error function used in the re-synthesis is given as a weighted sum of the errors for the fit, volume and parameter range as defined in eqs. (8), (10) & (11):

$$e(\mathbf{y}) = e_{fit}(\mathbf{y}) + w_{vol} \cdot e_{vol}(\mathbf{y}) + w_{range} \cdot e_{range}(\mathbf{y}) \tag{12}$$

Suitable values for the weights were empirically determined to w_{vol} = 0.001 and w_{range} =1.0. Minimisation was then carried out for all frames in the corpus.

Examples of the resulting resynthesis animations are available at the following URL:

`http://www.speech.kth.se/multimodal/qsmt/`

5. CORRELATION BETWEEN THE DATASETS

Using the analysis principle described in [2] and [3], we can investigate the interrelation between the face (Qualisys) and tongue (Movetrack) datasets. Using linear regression, one data set is predicted from the other, and the correlation between the original and the predicted can be calculated. Face data is arranged in a N-by-75 matrix $\mathbf{X}$, where each row represents a time frame (N = number of frames in the given corpus), and the columns hold the x-, y- and z-coordinates of the 25 points, excluding the reference points on the Movetrack headmount. A similar N-by-$2K$ matrix $\mathbf{Y}$ is constructed for the EMA data, containing x- and y- coordinates of K Movetrack coils. The analysis below is carried out with K=3 coils (tongue only), 4 coils (tongue and jaw) and 5 coils (tongue, jaw and upper lip).

To predict face data from the EMA, analogous to equations (5) and (6), we can write

$$\tilde{\mathbf{X}} = \mathbf{T}_{YX} \cdot \mathbf{Y}' \tag{13}$$

where $\mathbf{Y}'$ is $\mathbf{Y}$ augmented with a column of ones and the estimator matrix $\mathbf{T}_{YX}$ is given by

$$\mathbf{T}_{YX} = \mathbf{X} \cdot \mathbf{Y}^{T} \cdot (\mathbf{Y} \cdot \mathbf{Y}^{T})^{-1} \tag{14}$$

A jackknife training procedure is used: the data is split into ten parts where one part is used for prediction (13) and the remainder to train the estimator (14). This is repeated so that all parts are used for training and prediction. Correlation coefficients between the predicted and the original are calculated. The same procedure is used to predict tongue data from face data. The results are shown in table 1.

It can be noted that for 3 and 4 coils (tongue, tongue + jaw), prediction of EMA data from face is better than face from EMA, but with 5 coils (tongue + jaw + lip) face is better recovered from the EMA than the opposite. This is consistent with what has been reported by other investigators [2,3]. The correlations are generally higher for VCV than for CVC

and sentences. For all corpora, lip and jaw coils are predicted nearly perfectly, while the mid and back coils of the tongue have the lowest predictability.

Table 1. Average correlation coefficients between the predicted and measured coordinates.

	EMA from face			**Face from EMA**		
Corpus	**3 coils**	**4 coils**	**5 coils**	**3 coils**	**4 coils**	**5 coils**
VCV	0.658	0.738	0.763	0.539	0.684	0.815
CVC	0.511	0.624	0.693	0.298	0.484	0.820
Sentences	0.525	0.636	0.690	0.357	0.581	0.724
All	0.520	0.624	0.670	0.385	0.595	0.737

6. CONCLUSIONS

The combination of optical motion tracking and EMA measurements provide a good source of dynamical data to improve the accuracy of visual articulatory speech synthesis. The extent to which the two data sets could be linearly predicted from each other is slightly lower than in previous studies, but this can partly be explained by the fact that a more diverse corpus was used in this study.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] O. Engwall, *Tongue Talking – Studies in Intraoral Speech Synthesis*, PhD Thesis, KTH, Sweden, 2002.
[2] H. Yehia, P. Rubin and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour", *Speech Communication*, vol. 26, pp. 23-43, 1998.
[3] J. Jiang, A. Alwan, L. Bernstein, P. Keating and E. Auer, "On the correlation between facial movements, tongue movements and speech acoustics", *Proc of ICSLP2000*, vol.1, 42-45, 2000.
[4] G. Bailly and P., Badin, "Seeing tongue movements from outside", *Proc of ICSLP2002*, 1913-1916, 2002.
[5] J. Beskow, "Animation of Talking Agents", *Proc of AVSP'97*, pp. 149-152, 1997.
[6] F.I. Parke, "Parameterized models for facial animation", *IEEE Computer Graphics*.vol. 2(9), pp 61-68, 1982
[7] P. Branderud, "Movetrack - a movement tracking system", *Proc of the French-Swedish Symposium on Speech*, Grenoble, 113-122, 1985.
[8] `http://www.qualisys.se`
[9] T. Öhman, "An audio-visual speech database and automatic measurements of visual speech", *KTH TMH QPSR*, vol. 1-2, pp. 61-76, 1998.
[10] A. MacLeod and Q. Summerfield, "A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise. Rationale, evaluation and recommendations for use", British Journal of Audiology, vol. 24, pp. 29-43, 1990.