Introducing visual cues in acoustic-to-articulatory inversion

Olov Engwall
Centre for Speech Technology, KTH, Stockholm, Sweden
ológ@speech.kth.se

Abstract

The contribution of facial measures in a statistical acoustic-to-articulatory inversion has been investigated. The tongue contour was estimated using a linear estimation from either acoustics or acoustics and facial measures. Measures of the lateral movement of lip corners and the vertical movement of the upper and lower lip and the jaw gave a substantial improvement over the audio-only case. It was further found that adding the corresponding articulatory measures that could be extracted from a profile view of the face; i.e. the protrusion of the lips, lip corners and the jaw, did not give any additional improvement of the inversion result. The present study hence suggests that audiovisual-to-articulatory inversion can as well be performed using front view monovision of the face, rather than stereovision of both the front and profile view.

1. Introduction

Acoustic-to-articulatory inversion, i.e. recovering the shape of the vocal tract from an acoustic signal, can be considered as one of the big remaining challenges in speech technology research.

The first reason for this is that the potential applications are vast. A reliable and robust inversion method would provide a major breakthrough in computer assisted language learning, as automatic feedback could be given based on the student’s production [1]. It could moreover be beneficial in speech coding or acoustics and facial measures. Measures of the lateral and vertical movement of lip corners and the jaw, approximately 8, 20 and 52 mm from the tip are of importance for this paper. The EMA coils on the upper lip and lower incisor were used to align the Qualisys and Movetrack data sets and the coil on the jaw was not used in this study.

The second reason is however that important problems need to be solved first, such as the non-unicity of the inverse solutions, the need to eliminate phonetically irrelevant solutions and the impossibility to use standard spectral data as input.

The problem has mostly been attacked using inversion by synthesis, i.e. an articulatory model is used to create a feature description of all possible sounds [2, 3]. Inversion from acoustic data to an articulatory description is then performed by matching the acoustic data to the output of the model, through lookup in an articulatory codebook.

A different approach towards finding articulatory information from corresponding acoustic data is to statistically determine the quantitative association of acoustics and articulatory data of the tongue [4, 5]. Both [4] and [5] found rather high correlations between the acoustic, facial and vocal tract data.

Acoustic-to-articulatory inversion can be improved by visual information of the subject’s face, as data on jaw position, mouth opening and lip rounding may provide additional information to improve the recovery of the position of the tongue. To be useful in a real application, the added visual information must however be automatically retrievable in real-time from video images. This signifies that only a small number of articulatory measures can be extracted in each time frame and that they should be extractable using computer vision. The problem of tracking facial features, especially the lip contour, with automatic algorithms has been studied extensively (e.g., [6]). This paper does not cover the tracking, and instead focuses on the improvement in articulatory inversion that can be made assuming that the tracking is successful.

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 one repetition of 138 symmetric VCV words (henceforth referred to as VCV words), where $V = \{a, i, u\}$ and $C = \{p, t, k, b, d, g, f, s, z, s, s, m, n, l, t, j, t, c, v, j, b, z, zh, l, k, gl, fl, gr, br, kr, tr, k, n, l, tr, dr, fr, sl, sp, sk, sh, str, spr, skr, skl\}$ and 178 short everyday sentences. The sentences follow a rationale for audio-visual speech perception tests [7] and were adapted to Swedish by G. Öhngren.

2.2. Measurement setup

The audio signal, electromagnetic articulography (EMA) data of the tongue and facial data were recorded simultaneously by [8] in a sound-proofed room, using a DAT tape recorder, the electromagnetic articulograph Movetrack [9] for the tongue movements and the stereo-motion capture system MacReflex from Qualysis [10] for the face. Six EMA receiver coils were used in the acquisition, as shown in Fig. 1. The three placed on the tongue, approximately 8, 20 and 52 mm from the tip are of importance for this paper. The EMA coils on the upper lip and upper incisor were used to align the Qualysis and Movetrack data sets and the coil on the jaw was not used in this study.

Figure 1: Placement of EMA coils and Qualysis markers. Black Qualysis markers were used in the audiovisual-to-articulatory inversion; J=jaw, RC=right lip corner, LC=left lip corner, LL=lower lip edge, UL=upper lip edge. Arrows indicate if horizontal or vertical movement was used for the markers.
Fig. 1 further shows the placement of the 25 small reflectors in the Qualisys system that were used to capture the subject’s facial movements. An additional 3 markers were glued to the Movetrack headmount to be able to adjust for head movements. The Qualisys system tracks the reflectors using four infrared cameras and calculates their 3D-coordinates at a rate of 60 Hz.

3. Data processing

3.1. Speech acoustics

The speech signal was originally sampled at 16 kHz and a downsampling to the sample rate of the articulatory data was hence needed. The speech signal was divided into frames with a frame length of 24 ms and a frame shift of 16.67 ms to match the 60 Hz frame rate of the Qualisys system. Each acoustic frame was first pre-emphasized and multiplied by a Hamming window. A covariance-based LPC algorithm [11] was then applied to generate 16 line spectrum pairs (LSP). LSP coefficients were used, following [4] and [5], as they are closely related to the formant frequencies and the vocal tract shape. The speech signal was represented by the RMS amplitude and the 16 LSP coefficients in each frame.

3.2. Qualisys data

The Full Qualisys (FQ) data contained 3D positions of all markers, i.e. 75 measures, but as the objective of the study was to investigate the contribution of facial features that could be extracted easily from video images, two sub-sets were used: Monovision (MV) contained only data that could be recovered from a frontal image of the face: the vertical (z) movement of markers on the jaw (J), upper (UL) and lower lip (LL) and of the horizontal lateral (y) movement of the markers in the right (RC) and left (LC) lip corners, i.e. five measures. Refer to Fig. 1 for the placement of these markers.

Stereovision (SV) in addition to the measures in MV contained the horizontal protrusion (x) of the MV markers, in total 10 measures. This set was intended to mimic measures that could be found in the combination of both frontal and profile images of the face.

3.3. EMA data & Articulatory parameters

The EMA data was down-sampled to 60 Hz and aligned with the corresponding Qualisys lip and jaw markers to create a coherent 3D data set of face and tongue movements.

Parameter trajectories for a three-dimensional tongue model was then estimated from the EMA data. The 3D tongue model was derived from a statistical analysis of three-dimensional MR images of one Swedish subject producing a corpus of 13 vowels in isolation and 10 consonants in three symmetric VCV contexts [12].

The estimation of the parameter trajectories for the jaw height (JH), dorsum raise (TD), body raise (TB), tip advance (TA) and tongue width (TW) is described in detail in [8]. To summarize, the estimation is based on a fitting procedure to minimize the error function ε, defined as

$$\varepsilon(y) = \varepsilon_{fit}(y) + \varepsilon_{vol}(y) + \varepsilon_{range}(y)$$

where y is a combination of parameter values, \(\varepsilon_{fit}(y)\) is a measure of the goodness of fit between the model and data, \(\varepsilon_{vol}(y)\) is the difference between the reference tongue volume and that generated by the parameter values y, \(\varepsilon_{range}(y)\) is a penalty function for breaking the parameter range defined in the statistical model and \(w_{vol}\) and \(w_{range}\) are empirically derived weights.

The goodness of fit is calculated as the absolute difference between the positions of the real EMA coils and the corresponding virtual coils in the model. The volume conservation error \(\varepsilon_{vol}\) penalizes variations in the tongue volume, as it should remain constant over all articulations. Finally, the parameter range error \(\varepsilon_{range}(y)\) is used to introduce soft limits on the values that the articulatory parameters can take.

The benefit of trying to estimate tongue model parameters rather than EMA coil positions from the acoustic signal is that the parameters have an articulatory relevance and 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. [5] grouped different CV syllables based on the consonant place and manner of articulation, voicing and the vowel context and found some statistical differences between the groups, but did not investigate the significance on the articulatory level.

4. Statistically based articulatory inversion

4.1. Acoustic-to-articulatory inversion

Unseen tongue data \(X\) was estimated from the acoustic input \(Y\), using linear estimators ([4], [5]):

$$\hat{X} = T_{XY} \cdot Y'$$

where \(Y'\) is \(Y\) augmented with a column of ones, to allow for direct prediction of non-zero-mean vectors. The estimator \(T_{XY}\) is defined from training data as

$$T_{XY} = X \cdot Y'^T \cdot (Y' \cdot Y'^T)^{-1}$$

Linear regression estimation of the N-by-6 matrix \(X\) was performed for both EMA coil positions and the articulatory parameters. \(N\) is the number of time frames in the corpus and each row contains the data of that frame: either the x- and y-coordinates of the three EMA coils on the tongue or the values of the articulatory parameters.

The input data was a N-by-17 matrix \(Y\), where each row contained the 16 LSP coefficients and the RMS amplitude. A jackknife training procedure was applied, rotating the corpus in order to use all parts for training and prediction. Each utterance of a corpus (either one VCV word or one sentence) was estimated from Eq. (2), using an estimator \(T_{XY}\) of Eq. (3), determined from all the other utterances in the corpus.

4.2. Introducing cues from facial data

Several studies ([4], [5], [13]) have found strong correlations between facial data and tongue data. Facial data was therefore used to supplement the audio signal, in order to improve the inversion. Eqs. (2)–(3) were used, but the number of columns in \(Y\) was augmented to 92 (17+75) for the FQ data, 27 (17+10) for the SV data and 22 (17+5) for the MV data.

5. Results

5.1. Acoustic-only inversion

The correlation between the original tongue data set \(X\) and that estimated from the acoustic signal only, \(\hat{X}_{AO}\), is moderate, as shown by the white bars in Fig. 2. The correlation is substantially lower than that found by [5] for CV syllables (mean 0.78).
and by [4] for sentences of English (mean 0.61) and Japanese (mean 0.60), due to the fact that every utterance appear only once in this corpus, while both [5] and [4] used four (or five) repetitions. This meant that other productions of the utterance to estimate were included in the training set in their case. Statistically based acoustic-to-articulatory inversion of an utterance may hence achieve high correlation levels if other productions of the utterance are present in the training, but if it is completely unseen, the chance of a successful recovery is lower.

5.2. Acoustic inversion supported by articulatory measures

Significant improvements were achieved when the acoustic input data was supplemented with face data, as shown in Fig. 2. The improvement from audio-only to audio and Full Qualisys inversion is 50% for the VCV words and 70% for the sentences. However, even a much 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 for the VCV words, and a very slight improvement (8%) of the EMA coil estimation for the sentences.

The contribution of the MV measures was further evaluated, grouped based on the articulatory features jaw height (one measure of vertical position of marker J), mouth width (two measures, one each of the horizontal position of markers LC and RC) and mouth opening (two measures, one each of the vertical position of markers UL and LL).

An improvement was achieved with all the three cases of partial information from the articulatory measures (an increase of 14% to 25%), as shown in Fig. 3. Considering the importance that is traditionally attached to the jaw movements in speech production modelling, it may seem somewhat surprising that the jaw marker information gave the least increase in the VCV case, but note that the vertical position of LL provides some information on the jaw movement as well.

Note also the difference between the two corpora, as the upper and lower lips gave much more additional information in recovering the VCV articulations than the sentences.

5.3. Articulatory analysis

The articulatory analysis focused on what improvement the facial measures can contribute to 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 h, j, h, t/1, the facial data actually decreased the performance (-12%), mainly due to the fact that the combination of a lowered jaw and a raised tongue tip for h, t/ goes against the general tendency in the corpus that the movement of the jaw and tongue tip are positively correlated. In agreement with the importance of the lip corners in Fig. 3, the increase was the larger for VCVs in rounded /a/ or spread h/ than for /a/.

| 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). 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  | 0.34  | 0.62  |
| VCV AMV| [0.94]| 0.58  | 0.68  | [0.72]| 0.66  |
| VCV MVO| 0.93  | 0.47  | 0.62  | [0.67]| 0.52  |
| Sent AO| 0.45  | 0.31  | 0.32  | 0.33  | [0.50]|
| Sent AMV| [0.85]| 0.40  | 0.40  | 0.45  | 0.55  |
| Sent MVO| 0.80  | 0.24  | 0.30  | 0.38  | 0.37  |
Fig. 4: Midsagittal contours reconstructed from EMA measures (solid line), estimated from acoustic 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.

6. Discussion & Conclusions

This study has shown that important improvements can be achieved in acoustic-to-articulatory inversion by adding a small number of articulatory measures of the subject’s face. As these measures can potentially be extracted from a computer vision analysis of full 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. The analysis further indicates 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 – if the tracking of points is perfect. Stereovision may on the other hand provide a more robust tracking. We now plan to pursue the current research path by trying to extract the image features from video images.

7. Acknowledgements

This research is carried out within the ARTUR project, funded by the Swedish research council. The Centre for Speech Technology is supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The estimation of parameter trajectories from EMA data was performed by Jonas Beskow. The author would like to thank Peter Branderud for the loan of the Movetrack system, Bertil Lyberg for giving access to the Qualisys system and the enduring subject, Anne-Marie Öster.

8. References