Articulatory features for speech-driven head motion synthesis

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proc. Interspeech

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Articulatory features for speech-driven head motion synthesis

Atef Ben-Youssef 1, Hiroshi Shimodaira1, David A. Braude1

1Centre for Speech Technology Research, University of Edinburgh, United Kingdom
abenyou@inf.ed.ac.uk, h.shimodaira@ed.ac.uk, d.a.braude@sms.ed.ac.uk

Abstract

This study investigates the use of articulatory features for speech-driven head motion synthesis as opposed to prosody features such as F0 and energy that have been mainly used in the literature. In the proposed approach, multi-stream HMMs are trained jointly on the synchronous streams of speech and head motion data. Articulatory features can be regarded as an intermediate parametrisation of speech that are expected to have a close link with head movement. Measured head and articulatory movements acquired by EMA were synchronously recorded with speech. Measured articulatory data was compared to those predicted from speech using an HMM-based inversion mapping system trained in a semi-supervised fashion. Canonical correlation analysis (CCA) on a data set of free speech of 12 people shows that the articulatory features are more correlated with head rotation than prosodic and/or cepstral speech features. It is also shown that the synthesised head motion using articulatory features gave higher correlations with the original head motion than when only prosodic features are used.

Index Terms: head motion synthesis, articulatory features, canonical correlation analysis, acoustic-to-articulatory mapping

1. Introduction

Speech sound may be complemented with visual information (e.g. movements of the mouth; lips, jaw and tongue, and also eyebrows, eyelids, and head movements). Such complementary information increases speech intelligibility. Munhall et al. [1] found that the display of head motion improves speech perception. Research on speech-driven talking faces began with work on synthesis of lip and mouth motions that are synchronised with speech, i.e. lip sync. [2]. In contrast to the lip sync on which a significant number of studies has been done, automatic synthesis of head motion from speech has not been studied that extensively, especially in terms of the use of machine learning techniques.

Graf et al. [3] showed a link between the prosody expressed by the voice and that given by the head. Yehia et al. [4] proposed a frame-wise mapping based on a linear-regression model to estimate head rotation angles (Euler angles) from F0. They found that the linear model had to be separately trained on each utterance sample otherwise the correlation between F0 and head motion almost disappeared. A GMM-based simple frame-wise articulatory features can be regarded as an intermediate parametrisation of speech that are expected to have a close link with head movement. Measured head and articulatory movements acquired by EMA were synchronously recorded with speech. Measured articulatory data was compared to those predicted from speech using an HMM-based inversion mapping system trained in a semi-supervised fashion. Canonical correlation analysis (CCA) on a data set of free speech of 12 people shows that the articulatory features are more correlated with head rotation than prosodic and/or cepstral speech features. It is also shown that the synthesised head motion using articulatory features gave higher correlations with the original head motion than when only prosodic features are used.

Speech sound may be complemented with visual information (e.g. movements of the mouth; lips, jaw and tongue, and also eyebrows, eyelids, and head movements). Such complementary information increases speech intelligibility. Munhall et al. [1] found that the display of head motion improves speech perception. Research on speech-driven talking faces began with work on synthesis of lip and mouth motions that are synchronised with speech, i.e. lip sync. [2]. In contrast to the lip sync on which a significant number of studies has been done, automatic synthesis of head motion from speech has not been studied that extensively, especially in terms of the use of machine learning techniques.

Graf et al. [3] showed a link between the prosody expressed by the voice and that given by the head. Yehia et al. [4] proposed a frame-wise mapping based on a linear-regression model to estimate head rotation angles (Euler angles) from F0. They found that the linear model had to be separately trained on each utterance sample otherwise the correlation between F0 and head motion almost disappeared. A GMM-based simple frame-wise mapping has also been employed for a talking head [5], longer temporal information was used in [6] and [7]. In the former, HMMs were employed to map F0 and energy to a frame-wise VQ code of head rotation angles, whereas in the latter a discrete HMM was used to decode a sequence of animation cluster codes from the pitch and intensity features at every input syllable. Sar- gin et al. [8] developed a fully HMM-based approach for mapping the trajectory of F0 and intensity to the one of head rotation angles, in which parallel HMMs were used to cluster trajectory-
2.2. Speech driven head motion synthesis

Similar to the articulatory prediction form speech, we estimate the head motion from the predicted articulation. In training stage, streams of head motion and articulatory feature vectors are used to train multi-stream HMMs, whose model units are determined by the HMM-based clustering technique described in Section 3.3.1. For each stream, the emission probability density function of each state is modelled by a multivariate Gaussian distribution with a diagonal covariance matrix.

In the mapping stage, i.e. head motion synthesis stage, the sequence of head motion feature vectors (i.e. rotations of the head) $\hat{Z}$ is estimated from the intermediate articulatory feature vectors $\hat{Y}$ predicted from speech feature vectors $X$ (as shown in Eq. 1). The mapping form acoustic speech to head motion is performed as follows:

$$\hat{Z} = \arg \max_{\hat{Z}} \{ p(Z|\hat{\lambda}^{y,x}, \hat{Q}^{y,x}) P(\hat{\lambda}^{y,x}, \hat{Q}^{y,x}|\hat{Y}) \}$$  \hspace{1cm} (2)$$

where $\lambda^{y,x}$ is the parameters set of the articulatory-head motion HMM, $Q^{y,x}$ is the head-motion cluster HMM state sequence decoded from the predicted articulatory features $\hat{Y}$, $\hat{\lambda}^{y,x}$ is the parameters set of the acoustic-articulatory HMM and $\hat{Q}^{y,x}$ is the phone-size HMM state sequence decoded from acoustic speech. $\hat{Z}$ is obtained by maximizing all conditional probabilities. After predicting the articulatory features $\hat{Y}$, we decode the head-motion cluster HMM state sequence by maximizing $\arg \max_{\hat{Z}} \{ p(\hat{Z}|\hat{\lambda}^{y,x}, \hat{Q}^{y,x}) \}$ using the Viterbi algorithm. Second, we synthesise the head motion by estimating $\hat{Z} = \arg \max_{\hat{Z}} \{ p(\hat{Z}|\hat{\lambda}^{y,x}, \hat{Q}^{y,x}) \}$, using the MLPG algorithm [18].

3. Experiments

3.1. Data sets

In the present study, we used 12 speakers (denoted by $R00\text{xx}c\_e.S.\_X$) of the Edinburgh Speech Production Facility (ESPF) corpus [19]. This corpus contains speech movements over time synchronously recorded with audio and electropalatography. Using two Carstens AG500 electromagnetic articulometers, the speech movements of English speakers in dialogue were recorded.

3.1.1. Articulatory data

The articulatory data have been recorded by means of an Electromagnetic Articulograph (EMA) that tracks motion of flesh points of the articulators thanks to small electromagnetic receiver coils glued on the organs. Six coils are used: a jaw coil is attached to the lower incisors, whereas three coils are attached to the tongue tip, the tongue middle, and the tongue back; a coil is attached to the upper lip coil and another one to the lower lip coil in the midsagittal plane. The data was down-sampled to 100 Hz and their first derivatives was added.

3.1.2. Head motion data

Head motion is represented by the head correction of the articulatory data. Extra four coils attached to the upper incisor, to the nose and to the left and right ears served as references to extract the head correction. Head translations and rotations were calculated in order to remove the contribution of head movement.
from the articulatory data. In this study, head motion are represented by head rotations ($\theta_x, \theta_y, \theta_z$) about the z, y and x axes, respectively. The data was down-sampled to 100 Hz and their first derivatives was added.

3.1.3. Acoustic features extraction

Audio-speech signal was recorded, synchronously with EMA data, at a sampling frequency of 22,050 Hz and down-sampled to 16 kHz. Pitch denotes the combined features of the fundamental frequency (F0) that was extracted via an autocorrelation and cepstrum based method, log-energy, loudness contours, voicing probability, and voice quality. All these features (i.e. Pitch) were extracted with openSMILE [20], and then smoothed with a moving average filter with a window length of 10 frames. The first 12 MFCCs and 12 LPCs were extracted using SPTK. Two other LPC representation was tested: Log Area Ratios (LAR) represented by 12 LPC reflection coefficients extracted using HTK and Line Spectral Pairs (LSP) coefficients calculated from the 12 LPC. Pitch, MFCCs, LPC, LAR and LSP were computed from the audio signal over 25 ms windows at a frame rate of 10 ms to match the frame rate of the articulatory and head motion data. Their first time derivatives (i.e. delta parameters) were also added.

3.2. Head motion and speech correlation

Canonical correlation analysis (CCA) is employed in the present study to measure the linear relationship between two streams of vectors. The original CCA between two column vectors of random variables, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ is defined as the maximisation problem of the correlation between the linear combinations $A^TX$ and $B^TY$, with respect to the set of canonical coefficients $A \in \mathbb{R}^p$ and $B \in \mathbb{R}^q$. It is possible to find $d$ sets of canonical coefficients, where $d = \min(p, q)$.

We define global CCA as the average of $d$ canonical correlations over the whole data streams such that

$$r_G = \frac{1}{d} \sum_{i=1}^d \max_{A,B} \text{corr} \left( A_i^TX_{[1:T]}, B_i^TY_{[1:T]} \right)$$

The resulting matrices $U^{[i]}_{[1:T]} = A_i^TX_{[1:T]}$ and $V^{[i]}_{[1:T]} = B_i^TY_{[1:T]}$ are the $i^{th}$ canonical variables that maximise the Pearson’s correlation $\text{corr}()$. In practice, it is important to use a sufficiently large set of samples to avoid the trap of spurious correlation. Since the correlations between the speech and head motion streams are believed to change over time, it is more useful for us to define a local CCA for a time window of $n$ frames that starts at $t^{th}$ frame such that

$$r_L = \frac{1}{d} \sum_{i=1}^d \text{corr} \left( A_i^{[t:T]}X_{[t:t+n-1]}, B_i^{[t:T]}Y_{[t:t+n-1]} \right)$$

Figure 2: Global CCA, $r_G$, between speech and head motion features.

Figure 3: Average local CCA, $r_L$, between speech and head motion features.
initialise an HMM. The HMMs parameters were re-estimated using the EM algorithm and then new cluster labels were decoded using the Viterbi algorithm. This process was repeated until convergence was reached.

In order to find the optimal number of clusters and the optimal HMM topology that match best with the task of head motion synthesis, we varied the number of clusters, $K$. To define the best HMM configuration, we synthesise the head motion trajectories from the recognised sequence of clusters and the trained HMMs. Then, we evaluate it by a comparison with the original head motion trajectories. Preliminary experiment shows that the optimal number of clusters varies between 11 and 15, although it varies across the speakers.

A similar experiment was done for the number of states per HMM to confirm that there is no clear strategy for deciding the optimal number of states when clustering is concerned. Thus the number was fixed to 5 for the following experiments.

### 3.3.2. Evaluation of head motion synthesis

We used 15 clusters to train speaker-dependent multi-stream HMMs. 5-state left-to-right no-skip context-independent HMMs were used to model speech and head motion streams. A 3-fold cross validation procedure was used to evaluate the performance of the predicted head motion. Figure 5 presents the average local CCA $r_L$, between original and estimated head motion from different speech features, for all speakers. By looking to the results over all speakers, head motion estimated form both measured and predicted articulatory features are more correlated with original head motion than those predicted for prosodic or cepstral features. Medium correlation that we found may explained by the other factors that are involved on head motion such as speech stance and speaker culture and style semantic meaning. Figure 6 presents the average local CCA $r_L$, between speech and head motion estimated from different input speech features and averaged over all speakers. The estimated head motion from measured and predicted articulatory features are more correlated with speech than those estimated from prosodic and cepstral features. We conclude that the articulatory features give better results than prosodic or cepstral input speech features.

3.3.2. Evaluation of head motion synthesis

We used 15 clusters to train speaker-dependent multi-stream HMMs. 5-state left-to-right no-skip context-independent HMMs were used to model speech and head motion streams. A 3-fold cross validation procedure was used to evaluate the performance of the predicted head motion. Figure 5 presents the average local CCA $r_L$, between original and estimated head motion from different speech features, for all speakers. By looking to the results over all speakers, head motion estimated form both measured and predicted articulatory features are more correlated with original head motion than those predicted for prosodic or cepstral features. Medium correlation that we found may explained by the other factors that are involved on head motion such as speech stance and speaker culture and style semantic meaning. Figure 6 presents the average local CCA $r_L$, between speech and head motion estimated from different input speech features and averaged over all speakers. The estimated head motion from measured and predicted articulatory features are more correlated with speech than those estimated from prosodic and cepstral features. We conclude that the articulatory features give better results than prosodic or cepstral input speech features.

Other experiments were done using a tree-based state-tying strategy to train context-dependent HMMs. 4 mixture component Gaussian distributions by state were also tested. We note that there is no significant improvement in the system’s performances when either of these approaches was applied. It may be because of the small amount of the available data per speaker or the strategy used to tie the states of different clusters.
6. References


