Articulatory features for speech-driven head motion synthesis

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

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, Hiroshi Shimodaira, and David A. Braude.

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.

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.

Grad 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 trajectories of speech and head motion separately. Hofer et al. [9, 10] proposed the use of human-understandable head-motion units (e.g. nodding and shaking) as the model unit of HMMs. In their approach HMMs are trained with the combined streams of audio speech features (MFCC, F0, and energy) and head rotation angles. Despite the very low frame-wise correlations they found between the speech and head motion features, it was shown that head motion units were correctly recognised with an accuracy of approximately 70% on a free-speech data set, and reasonably natural head motions were synthesised.

Apart from linguistic features available from text, it is clear that the literature has focused on the speech features that are derived directly from acoustic signals. The present study, on the other hand, investigates the use of articulatory features for head motion prediction for the first time. The rationale for considering articulatory features is that there is some evidence that articulatory movements, e.g. opening the jaws, contribute to the movement of the head [11]. Articulatory features have been used successfully for automatic speech recognition [12] and emotional speech synthesis [13], but not yet for head motion synthesis.

The challenge of using articulatory features for head motion synthesis is that the training data of the target speaker normally do not come with articulatory data such as electromagnetic articulography (EMA), meaning supervised training of the model that maps speech features to articulatory features is not possible. To tackle this problem, semi-supervised learning using speaker adaptation is employed in this study.

It should be noted that, compared with the previous studies, which employed head-motion data of only one or two speakers, the present study employs the data of 12 people in order to improve the reliability of experiments in terms of speaker variety.

2. Speech driven head motion synthesis

The outline of the proposed approach is depicted in Figure 1.

2.1. Articulatory features prediction

To predict the articulatory features from speech, we used HMM-based acoustic-to-articulatory inverse mapping. In [14, 15], the author develop a multi-speaker inverse mapping system based on supervised adaptation, where each HMM of acoustic stream was adapted to the new speaker’s voice using the maximum likelihood linear regression (MLLR) technique and a small amount on labelled audio data.

In the present study, we consider a more realistic scenario where no labelled speech data are available for the target speaker. To address the problem, we developed a semi-supervised adaptation technique, where we train initial models on the labelled data of a different speaker, and we adapt the models to the target speaker in an unsupervised manner.

The mingul corpus [16] was employed to train the ini-
tional phone-size acoustic-articulatory streams HMMs using the
Minimum Generation Error (MGE) criterion [17]. To evaluate
the trained models, we used 5-fold cross validation to find an
RMSE of 1.13 mm and Pearson’s correlation of 0.87.

The unsupervised adaptation is done by running a recogni-
siser using the un-adapted models to obtain an initial phonetic
transcription of the data. Then, we used the recognised tran-
scription to re-adapt the models and re-decode a new phonetic
transcription from the adapted models, iteratively until conver-
gence. The convergence test was achieved by calculating the
accuracy between the new and old phonetic transcription for
each iteration. For all speakers, the accuracies were less then
23% (i.e. between −74% and 23%) on the first iteration. Af-
fer 8 iterations, the accuracies converge to more than 84% (be-
tween 84% and 97%). These differences of accuracies may be
ascribed to the sex difference, and the difference in size and
content of the corpus.

The predicted articulatory feature vectors (denoted by
\( P_{\text{red,EMA}} \)), for each speaker used in this study, are re-
presented by the trajectories of \((x, y)\)-coordinates of the 6 actives
EMA coils. The prediction of these trajectories was achieved in
two stages. First, phonetic state decoding was performed by the
Viterbi algorithm using the adapted acoustic HMMs. Second,
given the decoded HMM state sequence, the articulatory feature
vector sequence was inferred, using the articulatory HMMs. For
a given speaker’s acoustic feature vectors \( X \), we predict the
articulatory features \( Y \) such as

\[
\hat{Y} = \arg \max_{Y} \{ p(Y | \lambda^{y,x}, Q^{y,x}) P (\lambda^{y,x}, Q^{y,x} | X) \}
\]

where \( \lambda^{y,x} \) is the parameters set of the phone-size HMM
and \( Q^{y,x} \) the HMM state sequence. \( \hat{Y} \) is obtained by max-
imizing separately the two conditional probability terms of
Eq. 2. First, we decode the HMM state sequence by maximis-
ing \( \{ \lambda^{y,x}, \hat{Q}^{y,x} \} = \arg \max_{\lambda, \hat{Q}} \{ P (\lambda^{y,x}, \hat{Q}^{y,x} | X) \} \) using
the Viterbi algorithm. Second, we predict the articulatory
trajectories by estimating \( \hat{Y} = \arg \max_{Y} \{ p(Y | \lambda^{y,x}, \hat{Q}^{y,x}) \} \),
using the maximum-likelihood parameter generation algorithm
(MLPG) algorithm [18]. Pearson’s correlation between mea-
ured and predicted articulatory features decreases from 0.87
for the reference speaker to a correlation between 0.61 and 0.21
for the target speakers. This medium correlations can be ex-
plained by the difference of speakers vocal tract and also by the
difference of EMA coils position.

2.2. Speech driven head motion synthesis
Similar to the articulatory prediction form speech, we esti-
mate 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 den-
sity function of each state is modelled by a multivariate Gaus-
ian 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 features
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 such

\[
\hat{Z} = \arg \max_{\hat{Z}} \{ p(\hat{Z} | \lambda^{y,x}, Q^{y,x}) P (\lambda^{y,x}, Q^{y,x} | \hat{Y}) \}
\]

\[
p(\hat{Z} | \lambda^{y,x}, Q^{y,x}) P (\lambda^{y,x}, Q^{y,x} | \hat{Y})
\]

where \( \lambda^{y,x} \) is the parameters set of the articulatory-head mo-
tion HMM, \( Q^{y,x} \) the head-motion cluster HMM state se-
quence decoded from the predicted articulatory features \( \hat{Y} \), \( \lambda^{y,x} \)
the parameters set of the acoustic-articulatory HMM and \( Q^{y,x} \)
the phone-size HMM state sequence decoded from acou-
sic speech. \( \hat{Z} \) is obtained by maximizing all conditional prob-
abilities. After predicting the articulatory features \( Y \), we de-
code the head-motion cluster HMM state sequence by maximiz-
ing \( \{ \lambda^{y,x}, \hat{Q}^{y,x} \} = \arg \max_{\lambda, \hat{Q}} \{ P (\lambda^{y,x}, \hat{Q}^{y,x} | \hat{Y}) \} \)
using the Viterbi algorithm. Second, we synthesise the head
motion by estimating \( \hat{Z} = \arg \max_{\hat{Z}} \{ p(\hat{Z} | \lambda^{y,x}, \hat{Q}^{y,x}) \} \),
using the MLPG algorithm [18].

3. Experiments

3.1. Data sets
In the present study, we used 12 speakers (denoted by
\( R0002 \mathrm{c} \times \mathrm{c} \times X \)) of the Edinburgh Speech Production Facility
(ESPF) corpus [19]. This corpus contains speech movements over
time synchronously recorded with audio and elec-
tropalography. Using two Carstens AG500 electromagnetic
articulators, 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 Elec-
tromagnetic Articulograph (EMA) that tracks motion of flesh
points of the articulators thanks to small electromagnetic re-
ceiver coils glued on the organs. Six coils are used: a jaw coil
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 articu-
latory data. Extra four coils attached to the upper incisor, to
the nose and to the left and right ears served as references to ex-
tract the head correction. Head translations and rotations were cal-
culated in order to remove the contribution of head movement.

Figure 1: Overview of the speech driven head motion synthesis
system.
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 were also 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, voiced 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 of the correlation between the linear combinations \(A^T X\) and \(B^T Y\), 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

\[
r_G = \frac{1}{d} \sum_{i=1}^{d} \max_{A, B} \text{corr} \left( A^{[i]} X_{[1:T]}, B^{[i]} Y_{[1:T]} \right)
\]

The resulting matrices \(U^{[i]}_{[1:T]} = A^{[i]} X_{[1:T]}\) and \(V^{[i]}_{[1:T]} = B^{[i]} Y_{[1:T]}\) 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

\[
r_L = \frac{1}{n} \sum_{t=1}^{n} \text{corr} \left( A^{[i]} X_{[t:t+n-1]}, B^{[i]} Y_{[t:t+n-1]} \right)
\]

where \(A^{[i]}, B^{[i]}\) are the canonical coefficients obtained in the global CCA.

Average local CCA over time can be defined as

\[
r_{L} = F^{-1} \left( \frac{n}{T-n+1} \sum_{t=1}^{T-n} F(r_t) \right)
\]

where \(F(r)\) is the Fisher transformation defined as \(\frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)\), which is employed to make the values additive, and \(F^{-1}()\) is its inverse function. Note that if \(n = T\) then \(r_L = r_G\). We employ CCA to measure the correlation not only between speech and head motion features, but also between the original head motions and synthesised ones.

Figure 2 presents the global CCA, \(r_G\), between speech and head motion features. It is clearly shown that the measured articulatory features is more correlated with head motion that acoustic features.

Figure 3 presents the average local CCA, \(r_L\), between different speech features and head motion features. Approximately more than 70% of local CCA are greater than or equal to 0.2 when articulatory trajectories were used. This percent decreases to less than 30% for acoustic features.

3.3. Speech driven head motion synthesis

3.3.1. Clustering of head motion data

Data annotation is an essential step in the HMM training process. However, manual annotation is often time-consuming and expensive. In our experiments, the training data of head motions were automatically labelled using an HMM-based clustering technique.

We used GMM clustering to initialise the HMMs. Over the whole data of each speaker, GMM with \(K\) distributions was trained using the EM algorithm. Then, the data was decoded using the trained GMM into \(K\) clusters. Each cluster was used to

---

**Notes:**

2. http://btk.eng.cam.ac.uk

**References:**

[20] Pitch denotes the combined features of the fundamental frequency (F0) that was extracted via an autocorrelation and cepstrum based method, log-energy, loudness contours, voiced 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.

**Equations:**

\[
r_G = \frac{1}{d} \sum_{i=1}^{d} \max_{A, B} \text{corr} \left( A^{[i]} X_{[1:T]}, B^{[i]} Y_{[1:T]} \right)
\]

\[
r_L = \frac{1}{n} \sum_{t=1}^{n} \text{corr} \left( A^{[i]} X_{[t:t+n-1]}, B^{[i]} Y_{[t:t+n-1]} \right)
\]

\[
r_{L} = F^{-1} \left( \frac{n}{T-n+1} \sum_{t=1}^{T-n} F(r_t) \right)
\]
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 synthesised 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.

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.

![Figure 4: Distribution of local CCA, $r_L$, between speech and head motion features over all speakers.](image)

![Figure 5: Head motion synthesis from speech: average local CCA, $r_L$, between original and estimated head motion. Minimum, mean and maximum values over the speakers for each input speech features are shown as well.](image)

![Figure 6: Head motion synthesis from speech: mean over all speakers of the average local CCA, $r_L$, between input speech features and estimated head motion.](image)

### 4. Conclusion and perspectives

This paper presents the effectiveness of articulatory features for head motion synthesis from speech. In real-world head motion synthesis scenarios, it is not practical to assume the availability of articulatory measurements from a user. To address this challenge, HMM-based acoustic-to-articulatory mapping techniques have been proposed to predict articulatory features from an acoustic signal. This study confirmed that the articulatory features estimated from speech were more effective than prosodic and cepstral features for speech-driven head motion synthesis. Since those features are expected to be complementary, it would be interesting to investigate more sophisticated manners of feature integration. Further studies will include an extension to speaker-independent models with speaker adaptation and subjective evaluation of synthesised animation.

Another motivation of using HMM-based acoustic-to-articulatory mapping is to use the predicted articulatory features for lip sync. The lip motion may be modelled using two features, that is, the mouth opening (i.e. determined by the distance between $y$-coordinates of the coils attached to the upper and lower lip) and the mouth pucker (i.e. determined using $x$-coordinates of the coils attached to the upper and lower lip).

### 5. Acknowledgements

This work was supported by EU FP7 SSPNet (grant agreement no. 231287). The financial assistance of the National Research Foundation of South Africa (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.
6. References


