Articulatory features for speech-driven head motion synthesis

Atef Ben-Youssef [1], *Hiroshi Shimodaira*[1], *David A. Braude*[1]

[1]Centre 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 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. Sargin *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 $mngu0$ corpus [16] was employed to train the ini-

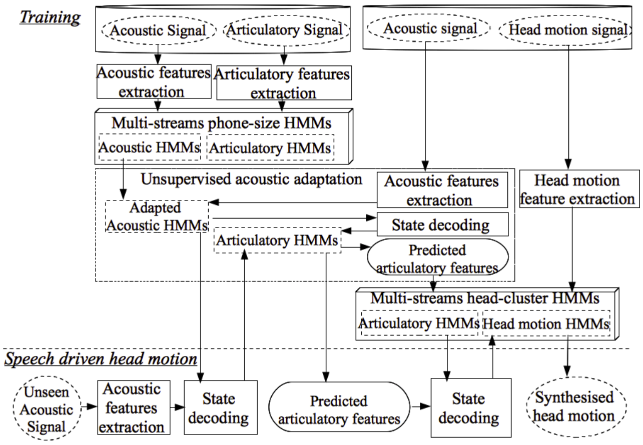
Training
Acoustic Signal
Articulatory Signal
Acoustic features extraction
Articulatory features extraction
Multi-streams phone-size HMMs
Acoustic HMMs
Articulatory HMMs
Acoustic signal
Head motion signal
Unsupervised acoustic adaptation
Acoustic features extraction
Head motion feature extraction
Adapted Acoustic HMMs
State decoding
Articulatory HMMs
Predicted articulatory features
Multi-streams head-cluster HMMs
Articulatory HMMs
Head motion HMMs
Speech driven head motion
Unseen Acoustic Signal
Acoustic features extraction
State decoding
Predicted articulatory features
State decoding
Synthesised head motion

Figure 1: Overview of the speech driven head motion synthesis system.

tial 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 recogniser using the un-adapted models to obtain an initial phonetic transcription of the data. Then, we used the recognised transcription to re-adapt the models and re-decode a new phonetic transcription from the adapted models, iteratively until convergence. 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. After 8 iterations, the accuracies converge to more than 84% (between 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 $Pred\_EMA$), for each speaker used in this study, are represented 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

$$\hat{Y} = \arg\max_{Y} \{p\left(Y|\lambda^{y,x}, Q^{y,x}\right) P\left(\lambda^{y,x}, Q^{y,x}|X\right)\} \qquad (1)$$

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 maximizing separately the two conditional probability terms of Eq. 2. First, we decode the HMM state sequence by maximising $\left\{\left(\hat{\lambda}^{y,x}, \hat{Q}^{y,x}\right) = \arg\max_{\lambda,Q} \{P\left(\lambda^{y,x}, Q^{y,x}|X\right)\}\right\}$ using the Viterbi algorithm. Second, we predict the articulatory trajectories by estimating $\left\{\hat{Y} = \arg\max_{Y}\{p(Y|\hat{\lambda}^{y,x}, \hat{Q}^{y,x})\}\right\}$, using the maximum-likelihood parameter generation algorithm (MLPG) algorithm [18]. Pearson's correlation between measured 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 explained 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 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 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 such as

$$\begin{aligned}\hat{Z} = \arg\max_{Z,Y} \{ & p\left(Z|\lambda^{z,y}, Q^{z,y}\right) P\left(\lambda^{z,y}, Q^{z,y}|\hat{Y}\right) \\ & p\left(\hat{Y}|\lambda^{y,x}, Q^{y,x}\right) P\left(\lambda^{y,x}, Q^{y,x}|X\right)\}\end{aligned} \qquad (2)$$

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

3. Experiments

3.1. Data sets

In the present study, we used 12 speakers (denoted by $R00xx\_csX$) 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

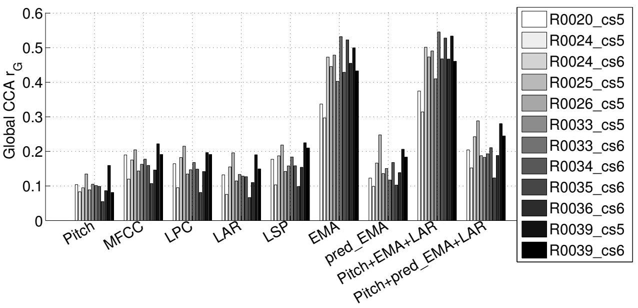

Figure 2: Global CCA, r_G, between speech and head motion features.

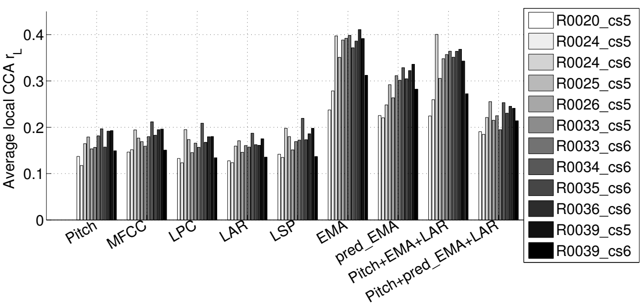

Figure 3: Average local CCA, r_L, between speech and head motion features.

from the articulatory data. In this study, head motion are represented by head rotations $(\theta_z, \theta_y, \theta_x)$ 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[1]. Two other LPC representation was tested: Log Area Ratios (LAR) represented by 12 LPC reflection coefficients extracted using HTK[2] 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, $X \in \mathcal{R}^p$ and $Y \in \mathcal{R}^q$ is defined as the maximisation problem of the correlation between the linear combinations $A^T X$ and $B^T Y$, with respect to the set of canonical coefficients $A \in \mathcal{R}^p$ and $B \in \mathcal{R}^q$. 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

$$r_G = \frac{1}{d} \sum_{i=1}^{d} \max_{A,B} \text{ corr} \left(A^{[i]T} X_{[1:T]}, B^{[i]T} Y_{[1:T]} \right) \quad (3)$$

The resulting matrices $U^{[i]}_{[1:T]} = A^{[i]T} X^{[i]}_{[1:T]}$ and $V^{[i]}_{[1:T]} = B^{[i]T} Y^{[i]}_{[1:T]}$ are the i^{th} canonical variables that maximise the Pearson's correlation 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

$$r_t = \frac{1}{d} \sum_{i=1}^{d} \text{corr} \left(A^{[i]T} X_{[t:t+n-1]}, B^{[i]T} Y_{[t:t+n-1]} \right) \quad (4)$$

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

Average local CCA over time can be defined as

$$r_L = F^{-1} \left(\frac{n}{T} \sum_{t=1;t=t+n;}^{T-n+1} F(r_t) \right) \quad (5)$$

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 different speech features 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 for a window length of 300 frames, i.e. 3 sec. For all speakers, the highest correlation was found for the measured articulatory trajectories (denoted as EMA), followed by the predicted ones. Note that a combined speech features of pitch, MFCC and articulatory trajectories doesn't increase the correlation. Note that the correlations shown here are much lower than those reported in other studies because a large amount of free-speech data was used in the present study.

Figure 4 displays the distribution of local CCA, r_t, between speech 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

[1] http://sp-tk.sourceforge.net

[2] http://htk.eng.cam.ac.uk

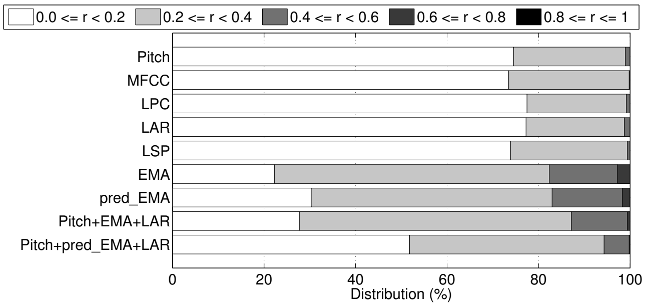

Figure 4: Distribution of local CCA, r_t, between speech and head motion features over all speakers.

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 variate 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.

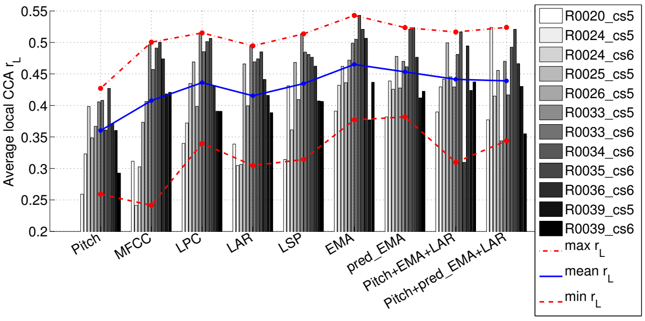

Figure 5: Head motion synthesis from speech: average local CCA, r_L, between original and estimated head motion. Minimum, mean and maximum values over the speakers for each input speech features are shown as well.

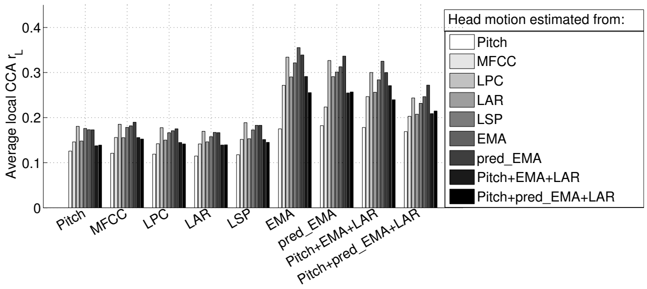

Figure 6: Head motion synthesis from speech: mean over all speakers of the average local CCA, r_L, between input speech features and estimated head motion.

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

[1] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson, “Visual prosody and speech intelligibility: head movement improves auditory speech perception,” *Psychological science*, vol. 15, no. 2, pp. 133–137, 2004.

[2] S. Morishima, K. Aizawa, and H. Harashima, “An intelligent facial image coding driven by speech and phoneme,” in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, no. 3, 1989, pp. 1795–1798.

[3] H. Graf, E. Casatto, V. Strom, and F. J. Huang, “Visual Prosody: Facial Movements Accompanying Speech,” *Proc. 5th International Conf. on Automatic Face and Gesture Recognition*, pp. 381–386, 2002.

[4] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, “Linking Facial Animation, Head Motion, and Speech Acoustics,” *Journal of Phonetics*, vol. 30, pp. 555 – 568, 2002.

[5] B. Le, X. Ma, and Z. Deng, “Live speech driven head-and-eye motion generators,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 11, pp. 1902–1914, November 2012.

[6] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, “Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–2007, March 2007.

[7] S. Levine, C. Theobalt, and V. Koltun, “Real-time prosody-driven synthesis of body language,” in *SIGGRAPH Asia 2009*, 2009.

[8] E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, “Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation,” *IEEE Trans. Patt. Anal. and Mach. Intel.*, vol. 30, no. 8, pp. 1330–1345, 2008.

[9] G. Hofer and H. Shimodaira, “Automatic head motion prediction from speech data,” in *Proc. Interspeech 2007*, 2007.

[10] G. Hofer, “Speech-driven Animation Using Multi-modal Hidden Markov Models,” Ph.D. dissertation, Uni. of Edinburgh, 2009.

[11] H. Zafar, E. Nordh, and P.-O. Eriksson, “Temporal coordination between mandibular and headneck movements during jaw openingclosing tasks in man,” *Archives of Oral Biology*, vol. 45, no. 8, pp. 675–682, 2000.

[12] P. Ghosh and S. Narayanan, “Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion,” *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–EL257, 2011.

[13] A. Black, H. Bunnell, Y. Dou, P. Kumar Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn, “Articulatory features for expressive speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, march 2012, pp. 4005–4008.

[14] A. Ben Youssef, T. Hueber, P. Badin, and G. Bailly, “Toward a multi-speaker visual articulatory feedback system,” in *Proceedings of Interspeech*, Florence, Italie, August 2011, pp. 589–592.

[15] A. Ben Youssef, “Control of talking heads by acoustic-to-articulatory inversion for language learning and rehabilitation,” THESIS, Grenoble University, 2011. [Online]. Available: http://tel.archives-ouvertes.fr/tel-00699008

[16] K. Richmond, P. Hoole, and S. King, “Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus,” in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[17] Y. Wu, W. Guo, and R. Wang, “Minimum generation error criterion for tree-based clustering of context dependent hmms,” in *Proceedings of Interspeech*. Pittsburgh, USA: ISCA, September 2006, pp. 2046–2049.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for hmm-based speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1315–1318.

[19] A. Turk, J. M. Scobbie, C. Geng, C. Macmartin, E. G. Bard, B. Campbell, B. Diab, C. Dickie, E. Dubourg, B. Hardcastle, P. Hoole, E. Kainada, S. King, R. Lickley, S. Nakai, M. Pouplier, S. Renals, K. Richmond, S. Schaefer, R. Wiegand, K. White, and A. Wrench, “An edinburgh speech production facility,.” [Online]. Available: http://www.lel.ed.ac.uk/projects/ema/home.html, 2010

[20] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *ACM Multimedia*, A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 1459–1462.