# SPEECH DRIVEN TALKING HEAD FROM ESTIMATED ARTICULATORY FEATURES

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

Centre for Speech Technology Research, University of Edinburgh, United Kingdom
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom

## ABSTRACT

In this paper, we present a talking head in which the lips and head motion are controlled using articulatory movements estimated from speech. A phone-size HMM-based inversion mapping is employed and trained in a semi-supervised fashion. The advantage of the use of articulatory features is that they can drive the lips motions and they have a close link with head movements. Speech inversion normally requires the training data recorded with electromagnetic articulograph (EMA), which restricts the naturalness of head movements. The present study considers a more realistic recording condition where the training data for the target speaker are recorded with a usual motion capture system rather than EMA. Different temporal clustering techniques are investigated for HMM-based mapping as well as a GMM-based frame-wise mapping as a baseline system. Objective and subjective experiments show that the synthesised motions are more natural using an HMM system than a GMM one, and estimated EMA features outperform prosodic features.

***Index Terms***— inversion mapping, clustering, head motion synthesis

## 1. INTRODUCTION

In an embodied virtual agent or animated character speech is supplemented with visual information. Some examples of possible extra information is the motion of the mouth, lips, and other articulators, eyebrows, eyelids and other facial features, and the movement of the head and body. Such supplementary information has been shown to increase intelligibility [1]. For example, nodding can be used not only for agreement, but also for emphasis, indicating attention and to indicate thinking during disfluencies [2, 3].

The context of the speech has been shown to be an important factor. It was shown that free speech had more expressive head motion than read speech [4]. In dialogues the interlocutor also has an impact on head motion and there is a bias towards nodding [5].

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. This work was supported in part by the Japan Science and Technology Agency (JST) CREST (uDialogue).

The head also moves according to the speech production [6]. Such motions, which the present study is interested in, are indispensable for making animated agents appear natural. Prior work in the analysis of the relationship between head motion and speech has mostly found on prosodic features (F0, energy, etc.). While it has been shown that there is a link between head motion and prosody [7], there is no clear linear mapping between acoustic features like F0 and head motion [8]. Our recent research has shown that articulatory features (movements of the lips, jaw and tongue) are more correlated with head motion than prosodic and cepstral features, even when they are estimated from the speech [9].

In addition to the speech features, motion unit and type of model are the important factors to achieve realistic and natural speech-driven head motion synthesis.

With regards to the motion units, there is a range from frame based approaches where each frame is estimated individually [8, 10], to using longer temporal information [11, 12]. When using longer segments as the basis of motion one could either work with meaningful units such as [13] or semi-or-unsupervised learning for example [14]. Regarding the modelling, a range of techniques have been used. Focusing on machine learning approaches, Gaussian mixture model (GMM) [10, 14] and hidden Markov model (HMM) [9, 11, 12, 13, 15], have been attempted.

In our previous work [9], electromagnetic articulograph (EMA) was employed to record articulatory motions and head motions, and an HMM-based clustering method was used to find motion units. The problem with an EMA is that head motions were constrained because of the nature of the recording method and thus they were not as natural as those when recorded without EMA. To mitigate the problem, the present study employs a motion capture system based on optical markers to record head motions. The present study investigates also different clustering techniques to find suitable head motion units for speech-driven head motion synthesis that uses estimated articulatory features. To be specific, HMM-based temporal clustering and Aligned Cluster Analysis (ACA) are considered for HMM-based mapping as well as GMM-based frame-wise mapping as a baseline system. In addition to objective evaluations, a subjective experiment is carried out to evaluation the naturalness of the synthesised animations.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

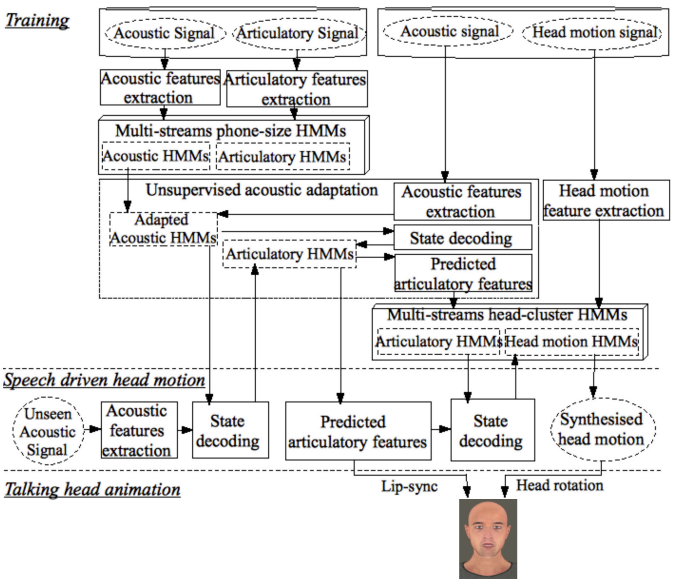


**Fig. 1**. Overview of the speech driven talking head animation.

## 2. SPEECH DRIVEN TALKING HEAD

Our overall approach is shown in Figure 1. There are three key parts, the articulatory features prediction, the head motion synthesis, and the talking head animation which is driven by the predicted articulatory features.

### 2.1. Articulatory features prediction

The articulatory features are estimated versions of the movements estimated by an Electro-magnetic Articulograph (EMA).

Previously we have shown that an HMM-based acoustic-to-articulatory system have created reasonable head motion [9]. For this paper we use the same phone-sized HMMs. The models were initialised on labelled data from different speakers then adapted to the target speakers' voices using unsupervised methods. We then iteratively create phonetic transcriptions and then used the new transcription to retrain the models until convergence of the difference in accuracy of the phonetic transcriptions [16].

The predicted articulatory features are $(x, y)$-coordinates of 6 active EMA coils (i.e. 2 coils attached to the upper and lower lip, 1 to the jaw and 3 to the tongue). We will thus refer to them as EMA features.

### 2.2. Head motion synthesis

For the many-to-many mapping of synthesising head motion from speech, we estimate the head motion from the intermediate articulation predicted from speech.

#### *2.2.1. HMM-based head motion synthesis*

In training stage, streams of head motion and speech feature vectors are fed to train multi-stream HMMs, whose model units are determined by the clustering technique described in Section 2.2.2. For each stream, the emission probability density function of each state is modelled by a multivariate Gaussian distribution with a diagonal covariance matrix.

In 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 $X$.

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

where $\lambda^{y,x}$ is the parameters set of the acoustic-articulatory HMM, $Q^{y,x}$ is the articulatory HMM state sequence decoded form speech $X$, $\lambda^{z,y}$ is the parameters set of the articulatory-head motion HMM and $Q^{z,y}$ is the head motion HMM state sequence decoded from the predicted articulatory features $\hat{Y}$. Articulatory features $\hat{Y}$ is obtained by maximizing separately the two conditional probability terms of: state sequence decoding $\left\{\left(\hat{\lambda}^{y,x}, \hat{Q}^{y,x}\right) = \arg\max_{\lambda,Q}\left\{P\left(\lambda^{y,x}, Q^{y,x}|X\right)\right\}\right\}$ and articulatory prediction $\left\{\hat{Y} = \arg\max_{Y}\{p(Y|\hat{\lambda}^{y,x}, \hat{Q}^{y,x})\}\right\}$. Then, head motion $\hat{Z}$ is obtained by maximizing all conditional probabilities. After predicting the articulatory features $\hat{Y}$, we decode the head motion HMM state sequence by maximising $\left\{\left(\hat{\lambda}^{z,y}, \hat{Q}^{z,y}\right) = \arg\max_{\lambda^{z,y}, Q^{z,y}}\left\{P\left(\lambda^{z,y}, Q^{z,y}|\hat{Y}\right)\right\}\right\}$ using the Viterbi algorithm. Then, 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 [17].

#### *2.2.2. Temporal clustering of head motion*

Manual annotation is often time-consuming and expensive. In our previous work [9], we used HMM-based clustering to annotate head motion data. Unfortunately, it is difficult to evaluate the clustering because there is no manual annotation. To resolve partially this problem, we compare different clustering techniques and analyse their impact on head motion synthesis. In this paper, we introduce another clustering procedure based on the Aligned Cluster Analysis (ACA), initially proposed for human motion segmentation by Zhou *et al.* [18]. ACA combines kernel $k$-means with a dynamic time alignment kernel to cluster time series.

To segment a sequence $X \in \mathcal{R}^{D,T}$ of data with $D$-dimension and $T$ frames into $M$ segments, segmentation matrix $S$ and cluster matrix $G$ are used to assign each segment to one cluster. The segmentation matrix $S$ contain the start and the end frames of each segment and the cluster matrix $G$ indicate if a segment belongs the cluster $c$. If $g_{c_t} = 1$ than $X_{S_t}$ belongs to class $c$ otherwise $g_{c_t} = 0$ where $S_t$ denote the segment begin at frame $t$ and end at frame $t + n - 1$. Note that $n$ represent the length of the segment and variate form $n_{min}$ to $n_{max}$. The segmentation $(G, S)$ were found using the equations detailed in [18].

#### *2.2.3. GMM-based head motion synthesis*

GMM-based approach was also applied to estimate head motion from speech, in order to compare our approach to prior

work such as [10, 11, 12] and use it as a baseline system to evaluate the effectiveness of our approach. We used frame-wise GMM-mapping that estimates head motion using Maximum Likelihood Estimation (MLE) from the input speech feature. This method was originally used for articulatory-to-acoustic mapping and inversion mapping detailed in [19].

### 2.3. Talking head animation

A $3D$ virtual talking head can be controlled by several parameters of different models. The head motion is described using three degree of freedom (3DOF) rotation. Lip motion can be described using three controls: The distance between the upper and lower lip (closure), horizontal lip coordinates (protrusion), and the vertical motion of the lips [20]. The lip motion is calculated from EMA data in [21]. Due to the high correlation of lip motion to the estimated EMA data subjective testing on the lip motion should give an indication of the quality of the predicted articulatory features.

## 3. EXPERIMENTS

### 3.1. Data sets

Data of four participants of 2 males (*m1* and *m2*) and 2 females (*f1* and *f2*) were used in the following experiments [22]. The available free speech are about 16, 25, 17, and 24 minutes for *m1*, *m2*, *f1* and *f2*, respectively.

#### *3.1.1. Head motion data*

The recordings were performed at 100 Hz with the Natural-Point OptiTrack[1] motion capture system which is expected to provide more natural head movement expression than ElectroMagnetic Articulograph (EMA), used in [9], where speakers wear EMA sensors and their head positioned inside a plexiglass cube.

Rotation matrices for the head and body were estimated from maker data using singular value decomposition, and then the relative head motions to the body were estimated by removing the effect of body motion. The obtained relative head motions were converted into extrinsic Euler angles using trigonometric identities. Finally, delta features were added.

#### *3.1.2. Acoustic features extraction*

Audio-speech signal was recorded synchronously with head and body motions, and down-sampled to 16 kHz. To compare the performance of the predicted articulatory features, we used prosodic features that was usually used in the literature. Pitch denotes the combined prosodic features of the fundamental frequency (F0) that was extracted via an auto-correlation and cepstrum based method, log-energy, loudness contours. All these features (i.e. Pitch) were extracted with openSMILE [23], and then smoothed with a moving average filter with a window length of 10 frames. Pitch features were computed from the audio signal over 25 ms windows at a frame rate of 10 ms to match the frame rate of the head motion data. Their first time derivatives were also added.

[1] http://www.naturalpoint.com/optitrack/

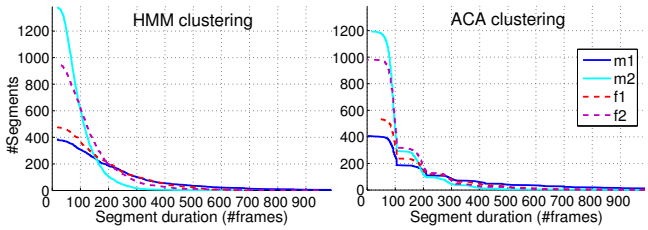


**Fig. 2**. Segment distribution sorted by duration.

### 3.2. Experiments results

Similar to [16], we employed the $mngu0$ corpus to train the reference phone-size acoustic-articulatory streams HMMs. The first 12 MFCCs and log-energy with their first derivatives was used as acoustic input features and $(x, y)$-coordinates of the 6 actives EMA coils with their first derivatives was used as articulatory output features. Using 5-fold cross validation, we find an RMSE of 1.13 mm and Pearson's correlation of 0.87. Subjective evaluation is used to evaluate the predicted articulation on term of lip-sync.

A 5-fold cross validation procedure is used to evaluate the performance of the predicted head motion. Preliminary experiment shows that the optimal number of clusters variate between 11 and 15, although it varies across the speakers. Busso *et al.* [11] found that 16 clusters achieves the best result of generating head motion sequences from prosodic features. In this experiments, we used 15 clusters to train speaker-dependent multi-stream HMMs and 16 mixture components to train GMM.

Figure 2 shows the segments distribution by ACA and HMM clustering techniques over all speakers. Using ACA clustering, we can see that there are fast motion (less than 1 $second$ = 100 $frames$), medium motion (between 1 and 2 $seconds$) and slow motion (up to 19 $seconds$). This division has previously been seen in recorded data [6]. This classification is much clearer with ACA clustering than HMM-clustering. Note that the shortest segments variate from 150 ms (i.e. phoneme size) to 390 ms (i.e. syllable size) across speakers. Total number of segments is slightly different between ACA clustering technique and HMM one.

#### *3.2.1. Objective evaluation*

Canonical correlation analysis (CCA) is employed to evaluate objectively the performance of the head motion synthesiser. To measure the correlation between the original head motions $X \in \mathcal{R}^p$ and the synthesised ones $Y \in \mathcal{R}^q$, we define *local CCA* $r_t$ for a time window of $n$ frames that starts at $t^{th}$ frame as

$$r_t = \frac{1}{d}\sum_{i=1}^{d} \text{corr}\left(A^{[i]T}X_{[t:t+n-1]}, B^{[i]T}Y_{[t:t+n-1]}\right) \quad (2)$$

where $d = \min(p, q)$ and $A^{[i]}, B^{[i]}$ are the canonical coefficients obtained by maximising the Pearson's correlation corr() of $A^{[i]T}X_{[1:T]}$ and $B^{[i]T}Y_{[1:T]}$ over the whole data streams such that

**Table 1**. Average local CCA and symmetric KLD ($r_L/KLD$) between original and synthesised head motion. The $*$ $+$ $\times$ and $\triangleleft$ denote significant differences ($p < 0.05$) found from $\wedge$ $\#$ $\circ$ and $\dagger$, respectively.

| *Speaker* / *Approach+Cluster-Feature* | *m1* | *m2* | *f1* | *f2* |
|---|---|---|---|---|
| GMM+∅-Pitch | *0.35/322.9 | *0.35/452.0 | *0.35/241.9 | *0.38/182.3 |
| GMM+∅-EMA | *0.36/19.0 | *0.36/35.1 | *0.35/19.9 | *0.39/26.4 |
| HMM+HMM-Pitch | $\wedge^{+}$0.47/2.0 | $\wedge^{+}$0.42/2.1 | $\wedge$0.48/2.0 | $\wedge^{+}$0.45/3.2 |
| HMM+HMM-EMA | $\wedge$0.49/1.7 | $\wedge^{\#}$0.46/0.6 | $\wedge$0.49/1.5 | $\underset{\times}{\wedge^{\#}}$0.48/2.4 |
| HMM+ACA-Pitch | $\wedge$0.48/2.1 | $\wedge^{\#}$0.47/1.0 | $\wedge^{\triangleleft}$0.48/1.6 | $\underset{\circ\triangleleft}{\wedge}$0.46/2.8 |
| HMM+ACA-EMA | $\wedge^{\#}$**0.51/1.6** | $\wedge^{\#}$**0.47/0.5** | $\wedge^{\dagger}$**0.51/1.2** | $\underset{\circ\dagger}{\wedge^{\#}}$**0.50/2.3** |

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

The *average local CCA* $r_L$ is defined such that

$$r_L = F^{-1}\left(\frac{n}{T}\sum_{t=1;t=t+n;}^{T-n+1} F(r_t)\right) \quad (4)$$

where $T$ is the total number of frames, $F(r)$ 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.

We consider also the similarity of the original and the synthesized motions by using the symmetric Kullback-Leibler divergence (KLD), where a smaller divergence value indicates that the motions are more consistent.

Table 1 presents average local CCA $r_L$ and symmetric KLD for all speakers. The approaches shown are GMM that does not require any segmentation (denoted by GMM+∅), HMM using HMM-clustering (denoted by HMM+HMM) and HMM using temporal clustering (denoted by HMM+ACA). We found that predicted articulatory features suits speech driven head motion better than prosodic features. We also found that the HMM-based method is significantly better than the GMM-based one and ACA temporal clustering gives better results than HMM-clustering for this task.

### *3.2.2. Subjective evaluation*

In this study, we used Poser Pro software[2] to design a female and a male virtual avatar. The $3D$ virtual avatar's head can be controlled by several parameters. We decided to use lip protrusion and lip aperture, as defined in [21], in order to simplify the control of the movements of the mouth. From the predicted articulatory features, we estimated the mouth opening from the predicted vertical lips coordinates and mouth pucker from the predicted horizontal lips coordinates. We use the estimated head motion (i.e. 3 rotation angles) to animate the avatar's head.

We performed a subjective evaluation through an online web application presenting the talking head animations with

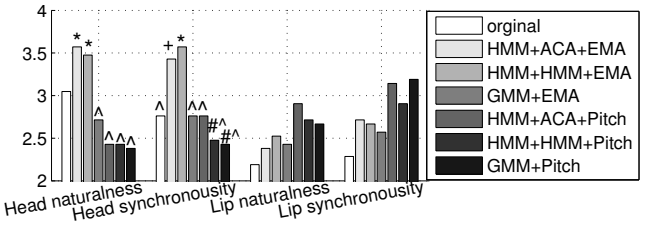


**Fig. 3**. Subjective evaluation results over 13 participants. The $*$ and $+$ denote significant differences ($p < 0.05$) found from $\wedge$ and $\#$, respectively.

question to answer. The animations[3] were represented as video clips showing lips and head motions. Note that there was no movements of eyebrow and body. Lip movements were the same in all the videos and was estimated form the predicted articulatory features (EMA). There are 7 video clips of 30 sec length for each speaker. After viewing the avatar's talking head animation of a selected speaker, the participants are asked to choose from 5 point score scale (i.e. from very bad (1) to very good (5)) related to the $naturalness$ of lip and head motions of the talking head animation as well as the $synchronousity$ between the speech and the animation.

The subjective tests were performed by 13 participants aged between 25 and 58. The average preferences over speakers' animations are shown in Figure. 3. Synthesised head motion using HMM-based approach was tend to be more realistic and natural than the synthesised motion using GMM-based one. We also found that articulatory input features provide head motion significantly better than those predicted from prosodic features, and ACA clustering is slightly better than HMM clustering. Lip motion was the same in all the animations. However, human perception of the lips vary according to the head motion. Participants gave higher score for lip-sync when the variation of head motion is smaller. This could be explained by the fact that participants focus on the lips when there is very small head movement. One factor that maybe involved in this medium results is that we used only two parameters to control Poser lips models from EMA lips coils coordinates.

## 4. CONCLUSION

This study shows that the articulatory features estimated from speech were more effective than prosodic features for the task of speech-driven head motion synthesis. HMM-based speech driven head motion approach is better than GMM-based one. ACA clustering is slightly better than HMM clustering. Both objective and subjective evaluation show that our proposed models to synthesise head motion give reasonable quality of the virtual talking head animation.

Further studies will include an extension to speaker-independent models with speaker adaptation. It will be interesting to also take into account the other factors involved in head motion such as the emotional state of the speaker.

[2]http://poser.smithmicro.com/

[3]http://homepages.inf.ed.ac.uk/abenyou/icassp2014.html

## 5. REFERENCES


[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] K.G. Munhall, J.A. Jones, D.E. 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.

[3] E. Z. McClave, "Linguistic Functions of Head Movements in the Context of Speech," *Journal of Pragmatics*, vol. 32, no. 7, pp. 855 – 878, 2000.

[4] A. Ben Youssef, H. Shimodaira, and D. A Braude, "Head motion analysis and synthesis over different tasks," in *Intelligent Virtual Agents*. Springer, 2013, pp. 285–294.

[5] C. T. Ishi, H. Ishiguro, and N. Hagita, "Analysis of relationship between head motion events and speech in dialogue conversations," *Speech Communication*, 2013.

[6] U. Hadar, T.J. Steiner, E.C. Grant, and F.Clifford Rose, "Kinematics of Head Movements Accompanying Speech During Conversation," *Human Movement Science*, vol. 2, no. 1-2, pp. 35–46, 1983.

[7] H.P. 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.

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

[9] A. Ben Youssef, H. Shimodaira, and D. A. Braude, "Articulatory features for speech-driven head motion synthesis," in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 2758–2762.

[10] B.H. 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.

[11] 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.

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

[13] G. Hofer, *Speech-driven Animation Using Multi-modal Hidden Markov Models*, Ph.D. thesis, Uni. of Edinburgh, 2009.

[14] D.A. Braude, H. Shimodaira, and A. Ben Youssef, "Template-Warping Based Speech Driven Head Motion Synthesis," in *Interspeech*, 2013, pp. 2763 – 2767.

[15] 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.

[16] 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.

[17] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *In Processing of IEEE ICASSP*, 2000, vol. 3, pp. 1315–1318.

[18] F. Zhou, F. De la Torre Frade, and J. K. Hodgins, "Aligned cluster analysis for temporal segmentation of human motion," in *8th IEEE Conference on Automatic Face and Gestures Recognition*, 2008.

[19] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215 – 227, 2008.

[20] D. Beautemps, P. Badin, and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2165–2180, 2001.

[21] P.K. Ghosh and S.S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *In Processing of IEEE ICASSP*, 2011, pp. 4624–4627.

[22] D. A Braude, H. Shimodaira, and A. Ben Youssef, "The university of edinburgh head-motion and audio storytelling (UoE-HaS) dataset," in *Intelligent Virtual Agents*. Springer, 2013, pp. 466–467.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, Eds. 2010, pp. 1459–1462, ACM.