Head Motion Analysis and Synthesis over Different Tasks

Atef Ben Youssef, Hiroshi Shimodaira, and David A. Braude

Centre for Speech Technology Research, University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
abenyou@inf.ed.ac.uk, h.shimodaira@ed.ac.uk, D.A.Braude@sms.ed.ac.uk

Abstract. It is known that subjects vary in their head movements. This paper presents an analysis of this variety over different tasks and speakers and their impact on head motion synthesis. Measured head and articulatory movements acquired by an ElectroMagnetic Articulograph (EMA) synchronously recorded with audio was used. Data set of speech of 12 people recorded on different tasks confirms that the head motion variate over tasks and speakers. Experimental results confirmed that the proposed models were capable of learning and synthesising task-dependent head motions from speech. Subjective evaluation of synthesised head motion using task models shows that trained models on the matched task is better than mismatched one and free speech data provide models that predict preferred motion by the participants compared to read speech data.

Keywords: head motion variety, head motion synthesis.

1 Introduction

Head movement demonstrates a wide variety of meaning. For example, nodding can be used not only for agreement, but also for emphasis, indicating attention and to indicate thinking during dis-fluencies [10,12]. Such motion can be different for the same speaker in other tasks or for others speakers doing the same task.

In recent years, the problem of driving head motion from speech has become a popular topic for research. Head motion may considered as complementary information for speech or other visual information (e.g. movements of mouth; lips, jaw and tongue, and also eyebrows, eyelids movements). This information increases speech intelligibility. Munhall *et al.* [12] found that the display of head motion also improves speech perception.

Research on speech-driven talking faces began with work on synthesis of lip and mouth motions that are synchronised with speech, (lip sync.) [11]. 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 extensively, especially in terms of the use of machine learning techniques. However, existing speech-driven head motion system often ignore the variance of head movements over different situation and speakers.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 285–294, 2013.

Graf *et al.* [4] showed a link between the prosody expressed by the voice and that given by the head. Yehia *et al.* [17] 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 [7], longer temporal information was used in [3,2] and [9]. 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.* [13] 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.* [6,5] 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. Lee *et al.* [8] evaluate 3 different machine learning techniques in head nods and eyebrow movements prediction. They found that the behaviors generated by the different models affect the human perception of the agent.

In this paper, we used an ElectroMagnetic Articulograph (EMA) corpus that contain articulatory and head movements recorded synchronously with audio. 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 [18]. The goal of the work described in this paper is to analysis the head motion variety over different tasks and speakers and their impact on head motion synthesis.

2 Data Set

In the present study, we used 12 English native speakers (4 males and 8 females denoted by $R00XX\_csX$) of the Edinburgh Speech Production Facility (ESPF) corpus [15]. This corpus contains articulatory and head movements over time synchronously recorded with audio and electropalatography. Using two Carstens AG500 electromagnetic articulometers positioned 8.5m apart to avoid electromagnetic inter-machine interference, the articulatory and head movements of English speakers in *dialogue* was recorded using 3D positions of sensors glued on the lips, tongue, jaw, and head. Communication among participants and experimenters is regulated via a talkback system.

Each speaker recorded different tasks. The recorded tasks were:

- Script reading: the speaker reads the script "Comma gets a cure"
- Map Giver: Map task, the speaker is the instruction giver.
- Map follower: Map task, the speaker is the instruction follower.
- Spot diff.: 3 Spot the difference picture tasks were recorded (Street, Diapix, Farm), the speakers was collaborating to find the differences.
- Repet. Teller: 2 Repetition task (Dance story, Loch story), the speaker is the reader.
- Repet. Shadower: 2 Repetition task, the speaker repeats what the teller reads out.
- Story Teller: Shadowing task, the speaker tell a story of his choice
- Story Shadower: Shadowing task, the speaker follow the the partner's story.

Depending on the speaker, the duration of the speech is between 11 and 38 minutes and the number of the available task is between 8 and 10.

2.1 Head Motion Data

Head motion is represented by the head correction of the articulatory trajectories in the data set. Four coils attached to the upper incisor, to the nose and to the left and right ears served as references to extract the head movements. Head translations and rotations were calculated in order to remove the contribution of head movements from the articulatory data. In this study, head motion are represented by head rotations (R_z, R_y, R_x) about the z, y and x axes, respectively. In order to use a common frame shift of 10 ms, the data was down-sampled to 100 Hz and their first derivatives was added.

2.2 Speech Data

Articulatory Data. Articulatory movements correspond to the horizontal and vertical midsagittal (x, y) coordinates of six coils attached to the speech organs. A jaw coil is attached to the lower incisors, three coils are attached to the tongue (tongue tip, tongue middle and tongue back), a coil is attached to the upper lip and another to the lower lip. The articulatory data (denoted by EMA and represented by 12 parameters) was down-sampled to 100 Hz to match the head motion data and their first derivatives were added. Note that audio-speech signal was recorded synchronously with EMA data (not used in this study).

3 Head Motion Variation

It is well-known that subjects vary greatly in their head movements. Although head movements is associated with many factors that can be explained by the settings of more/less speaking, the dialogue partner, the seating arrangement and also speaker's personality, social stance, physiological state and visual focus

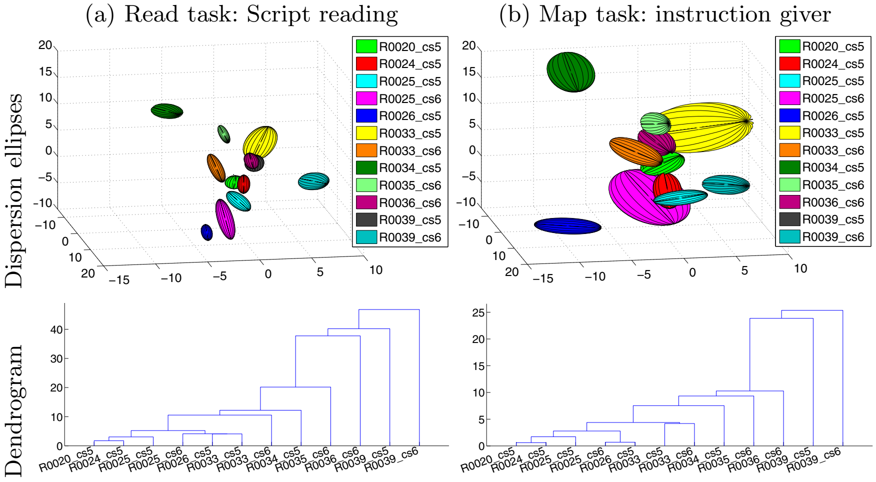

Fig. 1. Head motion variation of the same tasks over different speaker

of attention. Modelling the impacts of these factors and their dependency is a challenging problem.

This section discuss how the head motion varies between tasks and speakers. The dispersion ellipses of the head motion is represented in the $3D$ head space (i.e. x, y and z axes) by the mean and the full covariance matrix over the sets of the task instance (cf. Fig. 1) and speakers (cf. Fig. 2).

For a clear representation, we display all the task for two speakers in Fig. 2 and all speakers for two tasks in Fig. 1. Fig. 1 presents the distribution of two different tasks: (a) Read task (i.e. Script reading) when the speaker reads a script and (b) Map task (i.e. instruction giver) when the speaker is given the map instruction to the follower. This illustrates the very low variability of the head motion for the read task, as expected since the speaker's head is focusing on the script. The high variability of the head motion observed on the map task, can be explained by the free speech given by the speaker when he is collaborating with the follower to find the way. The variation of the head motion changes from speaker to another specially on the free speech (i.e. map task). Confusion trees have been built for speakers, based on the matrix of Mahalanobis distances of the head motion between each pair of speakers We confirm that the head motion is speaker dependent.

Fig. 2 displays the dispersion ellipses of head motion over all tasks for two speaker (i.e. male speaker $R0020\_cs5$ and female speaker $R0039\_cs6$), as well as the confusion trees that have been built for tasks, based on the matrix of Mahalanobis distances of the head motion between each pair of task. Using hierarchical clustering to generate dendrograms, we find different distance between tasks depending on the speaker. We can confirm that head motion is not only speaker dependent but also task dependent.

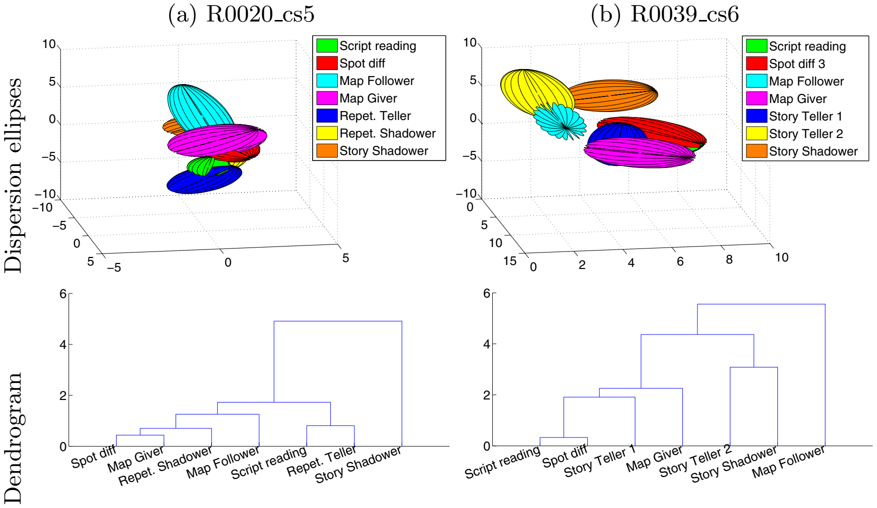

Fig. 2. Head motion variation of the same speaker over different tasks

Hierarchical clustering was performed based on Mahalanobis distances. The results are viewed in a dendogram, which displays the nodes arranged into their hierarchy and also shows how far apart the items were. Dendrograms of Fig. 1 and Fig. 2 that display the Mahalanobis distances between tasks and speakers, respectively, show that the distance between tasks (up to 6) is lower than the distance between speakers. This illustrate that head motion is depends more on speaker variation rather than tasks variation.

4 Speech Driven Head Motion Synthesis

The key idea of this paper is to model the head motion variation and its impact on head motion synthesis from speech.

We recall the experiments published previously in [1]. We found that canonical correlation analysis (CCA) on a free speech data shows that the articulatory features are more correlated with head rotation than prosodic and/or cepstral speech features. Therefore, we used measured articulatory features as input feature for speech driven head motion synthesis.

4.1 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. Furthermore, as head motion is concerned not only one segmentation will necessarily be right.

In our experiments, the training data of head motions were automatically labelled using an HMM-based clustering technique that may be able to provide both short and long segments that provide a statistical description of a particular motion. Algorithm 1 explain the instructions used to label automatically the data. We used GMM clustering to initialise the HMMs. Over the task data of a speaker, GMM with K distributions was trained using EM algorithm. Then, the data was clustered using the trained GMM into K clusters. Each cluster was used to initialise an HMM. The HMMs parameters were re-estimated using EM algorithm and then new cluster labels were decoded using Viterbi algorithm. This process was repeated until convergence was reached.

Algorithm 1. CLUSTERING of head motion data

```
Input: Head motion data of a task of one speaker
Output: Head motion cluster labels with their durations
Train GMM with $K$ distributions using EM algorithm.
Cluster the data into $K$ cluster labels using the trained GMM based on the
maximum likelihood.
Initialise $K$ HMMs with $K$ clusters
repeat
    Re-estimate the HMMs parameters using EM algorithm
    Decode new cluster labels using the re-estimated HMMs and Viterbi
    algorithm
until convergence is reached
return cluster labels
```

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. Busso *et al.* [3] found that 16 clusters achieves the best result of generating head motion sequences from prosodic features.

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.

4.2 Head Motion Synthesis

An overview of the multi-stream HMM-based speech-driven head motion system is presented in Fig. 3.

In this experiment, we used 15 clusters to train task-dependent multi-stream HMMs. 5-state left-to-right no-skip context-independent HMMs were used to model speech and head motion streams of the task. The proposed technique is based on the joint modelling of articulatory and head motion features, for each cluster.

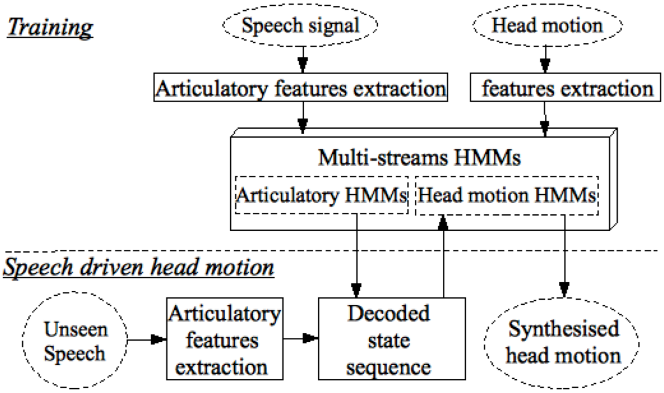

Fig. 3. Overview of the speech-driven head motion system

In training stage, streams of head motion and articulatory feature vectors are joined to train multi-stream HMMs, whose model units are determined by the HMM-based clustering technique [1]. For each stream, the emission probability density function of each state is modelled by a multivariate Gaussian distribution with a diagonal covariance matrix.

The speech driven head motion synthesis is achieved in 2 steps: 1) finding the most likely HMM state sequence from the articulatory observations; 2) inferring the head motion from the decoded state sequence. For a given speaker's articulatory feature vectors X, we predict the head motion features Y such as

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

where $\lambda^{y,x}$ is the parameters set of the head-motion cluster-size HMM and $Q^{y,x}$ the HMM state sequence. $\hat{Y}$ is obtained by maximizing separately the two conditional probability terms of Eq. 1. First, we decode the HMM state sequence by maximising $\left\{ \left(\hat{\lambda}^{y,x}, \hat{Q}^{y,x}\right) = \arg\max_{\lambda,Q} \left\{ P\left(\lambda^{y,x}, Q^{y,x}|X\right) \right\} \right\}$ using the Viterbi algorithm. Second, we synthesise the head motion 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 [14].

4.3 Evaluation

To evaluate the impact of the head motion variation over tasks, we used data of two tasks recorded by the same speaker (i.e. $R0020\_cs5$): map as instruction giver and script reading. The data of each task was split in two partition:

1. Training partition: two-third of the task was used to train the models (mapHMMs trained from the map task training data and readHMMs trained from script reading training data).

Table 1. Pearson's correlation between the original head motion and the synthesised one using matched models and mismatched models

Task of speech input \ used models	mapHMMs	readHMMs
Map task	0.47	−0.34
Read task	−0.38	0.48

2. Test partition: the remaining third was used for test. In order to evaluate the impact of the task on the head motion, cross-task speech-driven head motion synthesis was tested.

The articulatory speech data of the test task was the input for the two trained models to form matched and mismatched models on synthesis stage.

Objective Evaluation. To evaluate the impact of the task difference, Pearson's correlation between the original head motion and the synthesised one using the matched models and mismatched models was calculated. As can be observed in Table 1, the correlation on the matched condition is higher than the mismatched one. This result suggests that the synthesised data follow the motion of the task used for training rather than following the speech input specially for the mismatched condition.

Mismatched condition gives high, even though negative correlation. This means that as one rotation angle increases in value, the synthesised one decreases in value (i.e. when the head moves from up to down, the estimated movements was from down to up). This confirm that the proposed models was capable of learning and synthesising task-dependent head motions from speech.

Subjective Evaluation. We performed a subjective A/B comparison test to measure the opinions on the naturalness of the synthesised head motion. The participant are asked to chose between two head motion on a scale of A better than B, no preference and B better than A. 6 side-by-side comparison pairs was used: 3 pairs for each task. The 3 comparison was between the measured head motion (*org*) and two synthesised ones from matched HMMs and from mismatched one. Each comparison pairs is 50 seconds video length. The subjective tests were performed by 11 participants. The average preference are shown in Fig. 4.

The original data is typically perceived as much more realistic, except for the Read task, for which the mapHMM appears better. By looking to the map task results, we found that matched HMMs are more preferred than mismatched ons. However for read task results, contrary to our expectation, synthesised head motions using the mismatched models that was trained on free speech (i.e. mapHMMs) are preferred by the participant rather than the matched models. This can be explained that the free speech data that have more variation compared to read speech (see Fig. 1) may provide more expressive and preferred motion compared to read speech.

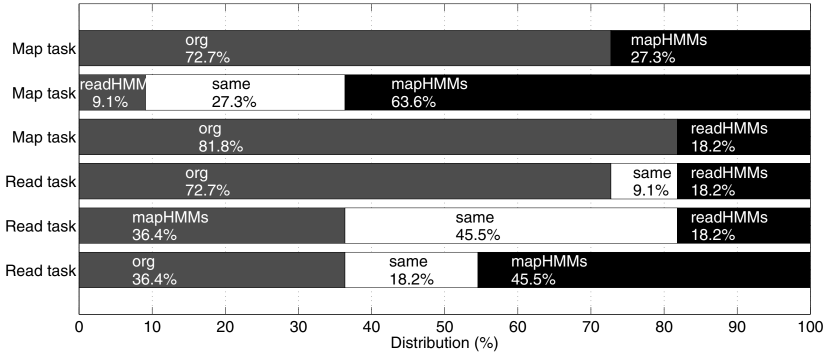

Fig. 4. Subjective A-B tests results over 11 participants

5 Conclusion

We have presented an analysis of head motion variation and the impact of this variety on synthesis. Over 12 speakers, we confirm that the head motion varies depending not only on the speaker but also on task. Articulatory features that have more correlation with head motion than acoustic features were used to drive head motion [1]. Experimental results confirmed that the proposed model was capable of learning and synthesising task-dependent head motions from speech. Synthesised head motion trajectories are more correlated with original motion when it was synthesised using a models trained on matched tasks than using mismatched ones. The subjective evaluation tests indicates that the free speech data may provide more expressive motion compared to read speech. A better head movement models could be trained using free speech data collected over similar tasks to the ones the avatar is supposed to perform.

This work could be extended in several ways. The advantage of HMM-based head motion synthesis is that a possible emotional and personalised motion can be achieved using adaptation techniques [16]. Further studies will include an extension to speaker-independent models with speaker adaptation.

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, acoustic-to-articulatory mapping system may be used to predict articulatory features from an acoustic signal. Another motivation of using acoustic-to-articulatory mapping is to use the predicted articulatory features for lip sync.

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.

References

1. Ben Youssef, A., Shimodaira, H., Braude, D.A.: Articulatory features for speech-driven head motion synthesis. In: Proceedings of Interspeech, Lyon, France (2013)
2. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1075–1086 (2007)
3. Busso, C., Deng, Z., Neumann, U., Narayanan, S.: Natural head motion synthesis driven by acoustic prosodic features. Computer Animation and Virtual Worlds 16(3-4), 283–290 (2005)
4. Graf, H., Casatto, E., Strom, V., Huang, F.J.: Visual Prosody: Facial Movements Accompanying Speech. In: Proc. 5th International Conf. on Automatic Face and Gesture Recognition, pp. 381–386 (2002)
5. Hofer, G.: Speech-driven Animation Using Multi-modal Hidden Markov Models. PhD thesis, Uni. of Edinburgh (2009)
6. Hofer, G., Shimodaira, H.: Automatic head motion prediction from speech data. In: Proc. Interspeech 2007 (2007)
7. Le, B., Ma, X., Deng, Z.: Live speech driven head-and-eye motion generators. IEEE Transactions on Visualization and Computer Graphics 18(11), 1902–1914 (2012)
8. Lee, J., Marsella, S.: Modeling speaker behavior: A comparison of two approaches. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 161–174. Springer, Heidelberg (2012)
9. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. In: SIGGRAPH Asia 2009 (2009)
10. McClave, E.Z.: Linguistic Functions of Head Movements in the Context of Speech. Journal of Pragmatics 32(7), 855–878 (2000)
11. Morishima, S., Aizawa, K., Harashima, H.: An intelligent facial image coding driven by speech and phoneme. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1989, vol. 3, pp. 1795–1798 (1989)
12. Munhall, K., Jones, J., Callan, D., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychological Science 15(2), 133–137 (2004)
13. Sargin, E., Yemez, Y., Erzin, E., Tekalp, A.M.: Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Trans. Patt. Anal. and Mach. Intel. 30(8), 1330–1345 (2008)
14. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1315–1318 (2000)
15. Turk, A., Scobbie, J.M., Geng, C., Macmartin, C., Bard, E.G., Campbell, B., Diab, B., Dickie, C., Dubourg, E., Hardcastle, B., Hoole, P., Kainada, E., King, S., Lickley, R., Nakai, S., Pouplier, M., Renals, S., Richmond, K., Schaefer, S., Wiegand, R., White, K., Wrench, A.: An edinburgh speech production facility
16. Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., Nakano, Y.: Model adaptation approach to speech synthesis with diverse voices and styles. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV–1233–IV–1236 (2007)
17. Yehia, H., Kuratate, T., Vatikiotis-Bateson, E.: Linking Facial Animation, Head Motion, and Speech Acoustics. Journal of Phonetics 30, 555–568 (2002)
18. Zafar, H., Nordh, E., Eriksson, P.O.: Temporal coordination between mandibular and headneck movements during jaw opening closing tasks in man. Archives of Oral Biology 45(8), 675–682 (2000)