Head Emotion

Erika Chuang, Chris Bregler
Stanford CS Tech Report CSTR 2003-02

Abstract

When people engage in a conversation or story telling, they employ a rich set of nonverbal cues, such as head motion, eye gaze, and gesture, to emphasize their emotion. These nonverbal cues are very important for animating a realistic character. We propose a new data-driven technique that is able to generate realistic and idiosyncratic head motions. Modeling head motions is a very challenging task, since there are no obvious correlations between the speech and expression content and the head motions. For the same sentence and expression, many different head motions are possible. In this paper, we show how we use the audio pitch contour information to drive head motions. First, a database of example head motions is collected. Given new audio, a combination of non-parametric sampling and dynamic programming technique is used to generate the sequences of natural and complex head motion from the database.

Keywords: Animation from Motion/Video Data, Expressive Motion/Communication, Facial Animation, Learning for Animation.

1 Introduction

When people engage in a conversation or story telling, they engage head motion as part of the nonverbal cues to emphasize their intention and emotion. These nonverbal cues are very important for animating a realistic character. In fact, in traditional animation, animators often start out in creating the head motion first, and then fill in "less important" details, like lip-sync and other expressions. Many important aspects of facial speech are expressed or further emphasized by head motions.

Traditionally, head motion for animated faces are produced (1) manually, (2) randomly, or (3) by a set of rule derived from communication studies. These techniques are either tedious for a long animation sequence, or do not capture the final details of motion that constitute a realistic character.

We propose a method for generating such realistic and idiosyncratic head motions from motion capture data and statistical techniques. These types of approaches have been setting the recent trends in facial animation. For example a Facial Speech Animation system can be built by collecting a large amount of example video or motion capture data of a person speaking. Several techniques have been proposed on how to use such a database or statistical model for the generation of new facial speech animations. Using example data helps to capture person-specific idiosyncrasies and other subtleties. However, these techniques often neglect to generate realistic head motions. The generated animations have either no head motions, or the new lip-motions). Although the lip motions and other facial features are animated very realistically, the lack of the "right" head motions diminishes the quality of the facial animation.

Head motion modeling presents a different set of challenges to motion-capture driven facial speech animation. There is no deterministic relationship, or obvious correlation, between head motion and other communication cues, which is an important factor that many statistical techniques rely on. For the same sentence and expression many different head motions are possible. Nevertheless, this does not mean that there is no correlation at all, in which case we can just generate the motion randomly. A keen observer can usually tell that randomly generated head motion, although superior to having none, bears no relationship to the attitude conveyed in the animation, thus exhibits a bland personality.

We believe that among the non-verbal cues, intonation and tone of voice bears import cues to how head motion should be generated.

However, recent study in the phonetics and linguistic community suggests that there may be some anatomical evidence that leads to possible discovery of the coupling between head motion and pitch [Honda2000]. [Yehia2000] also noticed that the pitch contour in the audio signal is highly correlated to head motions, but also noted that there is a one to many mapping from pitch to head motions. This leads us to believe that with the right assumption, it is possible to use the audio pitch contour information to generate head motions.

We propose a new data-driven technique that is able to generate such realistic and idiosyncratic head motions. Given limited data, we use a non-parametric sampling technique to synthesize short segments of head motion. To generate a sequence of head motions, these segments need to be cascaded together. Due to the nature of one-to-many mapping, we keep multiple possible motion segments at each stage, and later integrate over a period of time and solve for the best matching sequence with a dynamic programming algorithm.

2 Related work

While most facial animation papers do not address the problem of head motion generation, Some of them use a so-called "background" sequence, that contains natural head-motions, and the new lip and other facial feature motion is aligned with the background sequence motion. Therefore the head-motion does not fit the new generated speech motion [Bregler97, Brand99, Cosatto2000, Ezzat2002]. Some systems produce generic head-motions or random head-motions that are not correlated to the facial speech [Perlin].

There are also systems that use procedural or rule-based techniques for the head motion generation. Head-motion procedures that model speaker turns and interaction with a user have been proposed by Cassell [Cassell94], Nagao and Takeuchi [Takeuchi93] and Poggi and Pelachaud [Poggi2000]. Most recently, DeCarlo et al [DeCarlo2002] proposed a new procedural system that enhances non-verbal aspects of speech with head motions. They also carefully studied non-verbal cues in TV newscast videos, and noticed that very often head-motions (brief nodding, side turns, etc) are used to highlight the intended interpretation of the utterance. Our goal is not to replace such context driven head motion. It is our intention to complement such approach by paying attention to fine motions and idiosyncrasies to create more realistic animation.

Another area of research that is very relevant to our approach, and inspired the use of audio pitch information, comes from subject studies in the audio-visual speech perception community. [Yehia2000] found that head motion accounts for up to 88% of the pitch variation in the audio signal. Using head-motion they could very accurately predict the pitch variance. Unfortunately, the reverse task of using pitch information to infer the head motion variance has been proven to be much more difficult. As we mentioned already, pitch to head-motion is a one-to-many mapping problem, and is harder to model. In the next section, we will show how we can circumvent this problem by proposing a solution that finds an optimal sequence from multiple plausible head motion matches.

This problem is, in fact, similar to concatenative speech synthesis

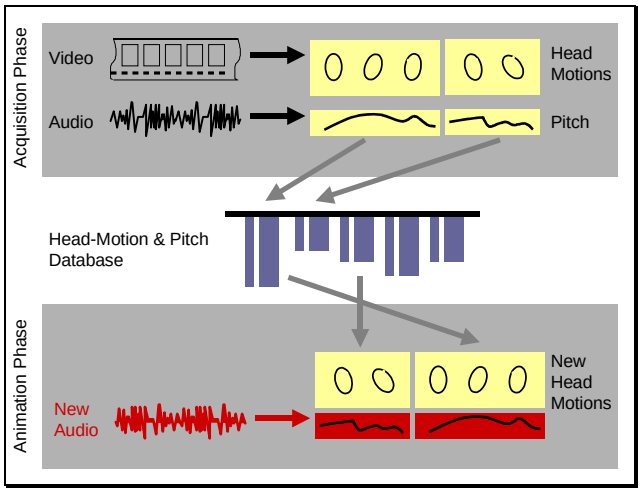

Figure 1: System overview. The system can be broken down into 2 stages. The first is the data acquisition stage, where …..

[Sorin94], facial speech synthesis [Bregler97], video synthesis [Schöedl2000], and motion capture synthesis [Kovar2002, Lee2002, Li2002, Arikan2002, Pullen2002].

3 Technical approach

Our system can be described as a two-phase approach: 1) Data Acquisition: This includes head-motion processing, audio processing and segmentation of the data into short segments. 2) Animation Mode: Finding a sequence of segments that matches the new audio signal, and post-processing of the new sequence. Figure 1 illustrates these two phases. More stuff…..

3.1 Data Acquisition

Head-motion acquisition: A marker-based motion-capture system from Vicon is used for motion acquisition. We employed 7 cameras with 120 Hz visual sampling frequency. 42 markers are placed at important facial position to obtain the facial expression. In addition, 8 markers are placed on a plastic hair-band, and 2 markers are on each of the ear lobes. In this study, we only use the position of the markers on the hair-band and the ear lobes to get accurate head motions, due to the rigidity of the points relative to each other. The Vicon software computes the 3D positions of the marker points. We computed 3D rotations and translations for the hair band and ear marker positions using a technique based on singular value decomposition (SVD) [Arun87].

Audio Pitch Analysis: Pitch is the fundamental frequency of the glottal oscillation (vibration of the vocal folds). A typical male voice has pitch range from 75 to 300 Hz, while a typical female speaker has pitch range from 100 to 600 Hz. We use [Praat], a program for doing phonetics, to extract pitch in the speech recorded simultaneously during the motion capture session. The algorithm used is based on an autocorrelation method [Boersma93]. The sampling rate is set to be 120 Hz to match the motion capture data. The silence and voicing thresholds are adjusted depending on the recording condition. The top part of figure (2) shows an example pitch contour obtained by the software.

Segmentation: Instead of treating each pitch and head motion sample independently, we partitioned the data into pieces of segments using the pitch contour information. First, the voiced region (where vocal folds vibration is present) detected by Praat has non-zero values. However, we only consider the pitch values that lie within the normal pitch range for longer than a threshold to be valid pitch values. Similarly, we only consider voiceless region longer than a threshold to be valid, because extremely short segments are hard to use in the matching process, as we will show in the next section. It is more important to observe the general pitch shape, or pitch accent. For region of zeros below the threshold length, we performed a simple interpolation from the neighboring pitch values. A data segment is defined as starting from the onset of one pitch segment and ending on the onset of the next pitch segment. Head motion continues to change regardless of whether it is a voiced and voiceless region. [However, our observation indicates that most of the large head motion happens right after the pitch onset, suggesting a good segment boundary at the beginning.] (((plot a graph of the velocity/acceleration of initial, mid, final part of sentence))))The bottom part of figure (2) shows the same pitch contours now partitioned and cleaned up after the processing.

In the rest of the paper, pitch segment refers to only the voiced region of the data segment.

3.2 Animation

3.2.1 Enlarging dataset: In order to increase the amount of data, we add to the head motion data by mirror imaging the head motion.
After we collect a database of pitch and head-motion segments, we are ready to generate new head-motions from a new audio input. This is done in four steps: 1) Partition the new input audio into a sequence of segments. 2) For each input segment, find N best matching database segments. 3) Find the best path through the sequence of matching segments. 4) Create the final head motion for the sequence.

The new input audio is segmented into pitch segments in the same way as described in the previous section. The rest is as follows.

3.2.2 Matching pitch: For a given sequence of new audio pitch segments, we need to find a matching sequence of segments from the database. We did this in a two-stage process:

In the **first stage**, we compared the pitch features for the entire phrase. This is important because the emotional, idiosyncratic content of the speech is often conveyed at a sentence level through pitch phrasing [Dellaert96]. These features include the statistics related to the speech rhythm, including the speaking rate, the average length of the voiced region and between voiced regions, and simple statistics on the pitch signal, including the minimum, maximum, mean, and standard deviation of pitch values. These statistics are normalized and used as a feature vector. Euclidian distance is used for calculating the top M sentences that best match with the test input. These M sentences are used as a subset of the database for the next stage.

In the **second stage**, we compared each individual pitch segment in the test sentence against the pitch segments in the subset from the first stage. Each pitch segment is re-sampled to match the length of average length of all pitch segments. Then we used root-mean-square difference to compute the distance between every other pitch segment.

$$Pdist = RMS(\mathbf{Pn}_{test} - \mathbf{Pn}_{template})$$

$\mathbf{Pn}_{test}$ and $\mathbf{Pn}_{template}$ are length normalized pitch contour. To avoid over stretching or shortening of the segments, a second criteria is used to penalize the cases where the length of the original segments are too different.

$$Ldist = \frac{\left|\text{length}(\mathbf{P}_{template}) - \text{length}(\mathbf{P}_{test})\right|}{\text{length}(\mathbf{P}_{test})}$$

The combined distance between 2 pitch segment is weighted by a constant c. The choice of c is chosen empirically by looking at the segments found.

$$Dtotal = c \cdot Pdist + (1 - c) \cdot Ldist \qquad (1)$$

The top N choices are kept for each test segment.

Finding the best path: Given K pitch segments in the test audio, and the top N matches for each test segment, we need to find a path through these segments that produce a good head motion trajectory. Here we take the head motion data accompanying each matching segment and compute the cost of transitioning from one segment to the next. A good match is determined if the segments have

1) Similar pitch segments (both shape and length wise)
2) Similar length in the voiceless region following the voiced region, and
3) Pair-wise matching boundaries for the head motions between successive segments, i.e. successive head-motion segments should have matching boundaries such as position and velocity.
4) Consecutive segments in the original database are highly encouraged, since it results in the most natural motion.
5) The same segment is discouraged, since it produces repetitive motion.

The first 2 items define the cost of each segment, while the last 3 terms define the transitional cost from one segment to the next. They are combined into 2 terms, S_k and $G_{k,k+1}$ in the following equation as weighted sum.

$$\text{PathCost} = \sum_k (\alpha \cdot S_k + \beta \cdot G_{k,k+1}) \qquad (2)$$

If we only have the S_k term, we could individually match each input sequence and have a linear-time complexity search problem. Since we also have transitional terms that couple the search for each successive segment pair, we need to use a more complex search procedure. We used the Viterbi algorithm [Viterbi67] that's related to the search for Hidden Markov Models (HHM). It is computational efficient due to its recursive nature. It also looks at the whole sequence before deciding on the most likely final state, and then `backtracking' through the pointers to indicate how it might have arisen. Figure 3 illustrates this search technique.

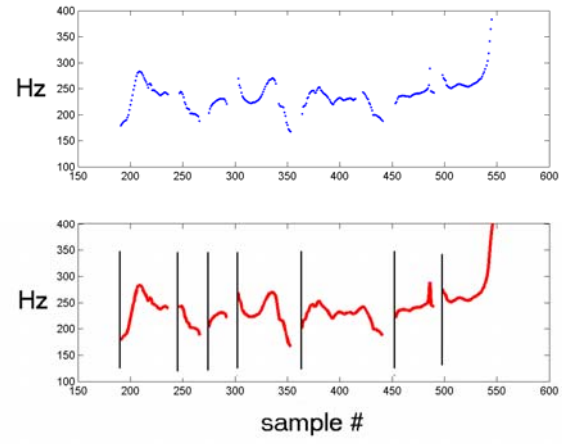

Figure 2: Top, example pitch contours output of Praat. Bottom, segmented pitch data.

Final Joining: The final head motion is computed using the best path chosen by the above algorithm as shown in figure (3). First, the motion data is re-sampled, since the lengths of the path segments do not match exactly with the length of the test segment. Although this alters the frequency content of the motion data, the length of the segment is part of the matching criteria, thus the selected segments should have length close to the test segment. Next, the starting position of each motion segment is moved to match the ending position of the previous segment. This could lead to potential drifting for long sentences. When that occurs, we break long sequences into smaller parts, and perform path search for each sub-sequence. The ends are joined via simple interpolation. The partitioning of long sequences is currently done manually. Finally, the end positions where the segments are connected are smoothed to minimize high frequency artifact caused by mismatch in the velocity.

4 Experiments

We recorded a collection of 67 phrases. Each phrase consists of 2 or 3 sentences. The actress was informed that highly expressive motions are desired, therefore the footage consists of a wide variety of motion. The test audio is analyzed with techniques described in the pre-processing section, and used to synthesize new head motion. A ball-head is to used to visualize the result. Figure (4) shows sub-sampled frames from the test data.

We found that it is difficult to judge the quality of the synthesized result, if we only animate head-motions, but do not render and animation facial expressions. Most people do not isolate head motion from the rest of the face. Therefore, we also apply the head motion to an animated 3D head model containing facial expressions. We used motion capture techniques to generate the lip motion and the rest of the facial expression [Ref]. The actress was told not to move her head around for our testing purpose. The final head motion is synthesized using the our technique. We found that by integrating the lip motion, the facial expression, and the head motion together, the face really comes alive.

We asked users to look at the animation with no head motion, head motion generated with random noise, head motion taken randomly from other sentences, and head motion synthesized by our system. We asked the users the following question and have them give scores to each of them:

(1) Does the character look lively.
(2) Is the character convincing.
(3) Does the character have interesting personality.

Figure (x) shows the result of this comparison. In comparison with head motion synthesized with random noise, the data-driven head motion seems to convey much more personality. Figure (5) shows some sub-sampled frames of the 3D facial animation.

Please consult the accompanying video to see the comparison between no head motions, random head motions, and the output of our technique.

5 Conclusion and future work

We have demonstrated a new technique that is able to generate head-motions for facial speech animation. In deriving these head-motions from example motion capture data, we are able to convey specific style and idiosyncrasies. This increases the realism of the animation. To the best of our knowledge, this is one of the first systems that generate head-motions in a data-driven way. We found, that it is possible to create new head-motions from only the audio-pitch information.

We are aware that the mapping between pitch and head-motions there is one-to-many. We can disambiguate some aspects with our global matching technique. But we also notice that several aspects of head-motions depend on the higher level semantics. Many non-verbal cues are used by a speaker to highlight the intended interpretation of the utterance. Most of those cues co-occur in pitch and head-motions, which is addressed in our technique. But several cues are in head-motions only. For example, if the speaker says: “and this on the right and that on the left”, she most certainly will also move her to the right and to the left. We can not infer those kind of head-motions only from pitch.

In future work, we propose to study further, which aspects of head-motions depend explicitly on those semantic cues, and are not contained in audio information. Those cues should be an additional input of our technique. Ultimately this technique could be used to augment a rule-based technique that covers those semantic cues, but will add more subtle idiosocratic features.

References

[Arikan2002] Arikan O., Forsyth D. A. Interactive motion generation from examples, SIGGRAPH 2002.

[Arun87] Arun KS, Huang TS, Blostein SD, Least-Squares Fitting of Two 3-D Point Sets”, IEEE Transaction on Pattern Analysis and Machine Intelligence, 9,5, 1987, pp698-700.

[Dellaert96] Dellaert F., Polzin T., and Waibel A.. *Recognizing emotion in speech*. In Proc. Int. Conf. on Spoken Language Processing, volume 3, Philadelphia, USA, 1996.

[BOERSMA93] BOERSMA, P., 1993, ”Accurate short-term analysis of the fundamental frequency and the harmonics0to-noise ration of a sampled sound”, *Proceedings of the Institute of Phonetic Sciences,***17**:97-110, University of Amsterdam.

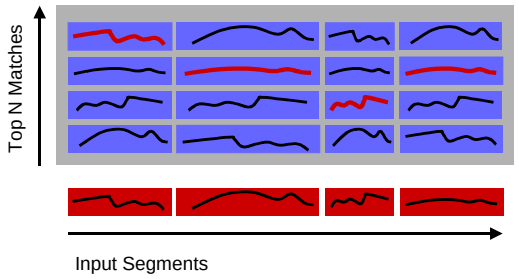

Figure 3: The best path is chosen via dynamic programming and joined together.

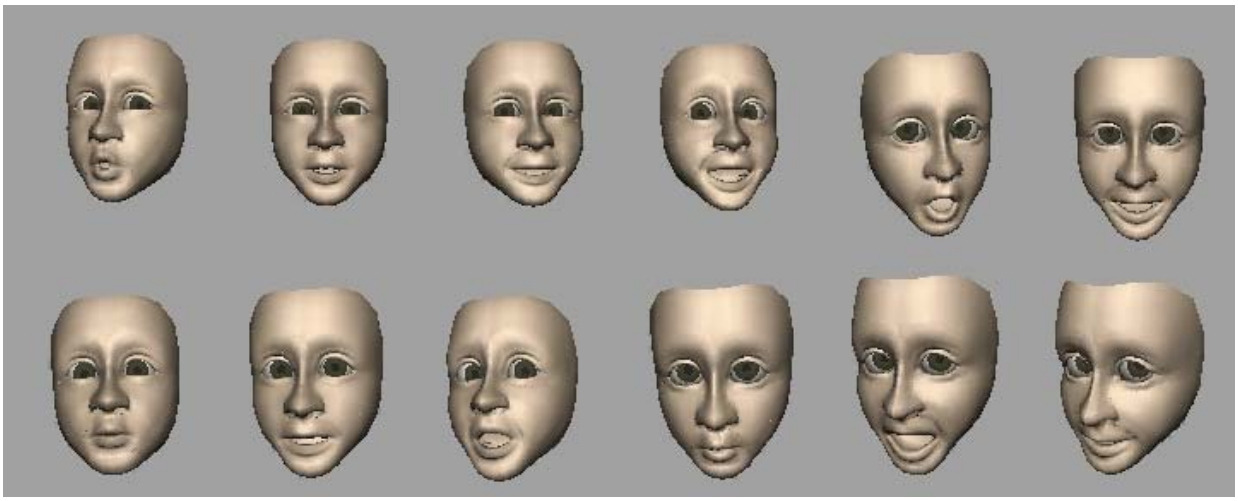

Figure 5. Head animated with lip motion, facial expression, and synthesized head motion.

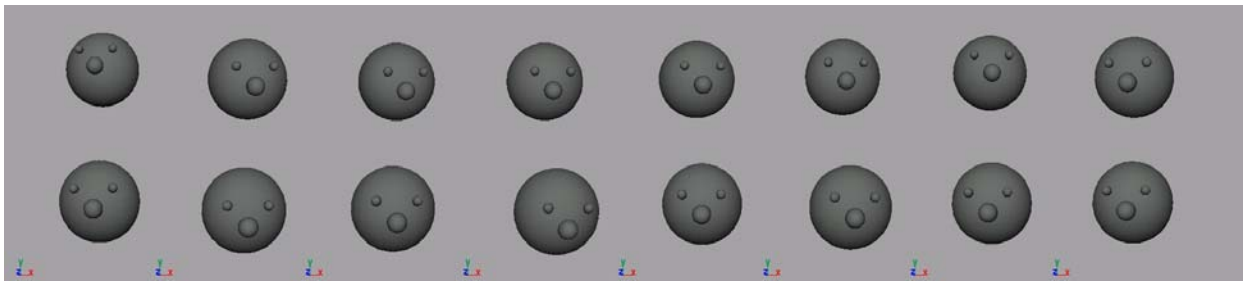

Figure 4. Visualization of synthesized head motion sequence.

[Bourland90] Bourlard H. and Morgan N.. A continuous speech recognition system embedding MLP into HMM. *Advances in Neural Information Processing Systems,* 2:413–416, 1990.

[BRAND99] Brand M. *Voice Puppetry, SIggraph 1999*

[BREGLER97] BREGLER, C, Covell M., and Slaney M., *Video rewrite: driving visual speech with audio*, in SIGGRAPH, 1997.

[Cassell94] Cassell J., Pelachaud C., Badler N., Steedman M., Achorn B., Becket T., Douville, B. Prevost S. and Stone M., *Animated Conversation: Rule-based Generation of Facial Expression*, Gesture and Spoken Intonation, for Multiple Conversational Agents, Proceedings Siggraph'94

[Cosatto2002] Cossato, E. "Sample-Based Talking-Head Synthesis", Signal Processing Lab, Swiss Federal Institute of Techology, Lausanne, Switzerland, October 2002. PhD. Thesis.

[DeCarlo2002] DeCarlo D., Revilla C., Stone M. and Venditti J. Making discourse visible: Coding and animating conversational facial displays *Computer Animation 2002*, pp. 11-16

[Ezzat2002] Ezzat T., Geiger G., and Poggio T., Trainable Videorealistic Speech Animation *Proceedings of ACM SIGGRAPH 2002, San Antonio, Texas, July 2002.*

[Honda2000] Honda, K. (2000) Interactions between vowel articulation and F0 control. In *Proceedings of Linguistics and Phonetics: Item Order in Language and Speech (LP'98), pp. 517-527.* O. Fujimura, B. D. Joseph & B. Palek (editors), Hungary: Kakrolinum Press, Prague

[Kovar2002] Kovar L., Gleicher M., Pighin F., Motion graphs, SIGGRAPH, 2002

[Lee2002] Lee J., Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, Nancy S. Pollard, Interactive control of avatars animated with human motion data, SIGGRPAH 2002. [Li2002] Li Y., Wang T., Shum H**.** Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis, SIGGRAPH 2002.

[Li2002] Li Y, Wang T., H-Y. Shum, Motion Textures: A Two-Level Statistical Model for Character Motion Synthesis, SIGGRAPH 2002.

[Maya] Alias|Wavefront.

[Parke75] A model for human faces that allows speech synchronized animation. Journal of Computers and Graphics, 1(1), 1-4.

[Parke2000] Parke, F. I., and Waters, K. 2000. *Computer Facial Animation*. A. K. Peters.

[Poggi2000] Poggi, C. Pelachaud, and F. de Rosis. *Eye communication in a conversational 3D synthetic agent*. Special Issue on Behavior Planning for Life-Like Characters and Avatars of AI Communications. 2000.

[Praat] Boersma P. and Weenink D., htto://www.praat.org.

[Pullen2002] Pullen K., C. Bregler C., "*Motion Capture Assisted Animation: Texturing and Synthesis*", SIGGRAPH 2002.

[Schödl2000] Schödl, Szeliski R., Salesin D., and Essa I.. Video textures. *Proceedings of SIGGRAPH 2000*, pages 489-498, July 2000.

[Sorin94] Sorin, C. (1994). *Towards high-quality multilingual text-to-speech*. Proceedings on the CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology, Munich, pp. 53--62.

[Takeuchi93] Takeuchi A. and Nagao K., Communicative facial displays as a new conversational modality. In ACM/IFIP INTERCHI '93, Amsterdam, 1993.

[Viterbi67] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans. Informat. Theory, IT-13:260-269, 1967. ISSN 0018-9448.

[Yehia2000] Yehia H, Kuratate, T. and Vatikiotis-Bateson, E. Facial animation and head motion driven by speech acoustics. In P. Hoole (ed). 5th Seminar on Speech Production: Models and Data, (pp. 265-268), Kloster Seeon, Germany, May 1-4, 2000.