

A Model for Mapping Speech to Head Gestures in Human-Robot Interaction

Amir Aly, Adriana Tapus

▶ **To cite this version:**

Amir Aly, Adriana Tapus. A Model for Mapping Speech to Head Gestures in Human-Robot Interaction. T. Borangiu and A. Thomas and D. Trentesaux. Service Orientation in Holonic and Multi-Agent Manufacturing Control, Springer, Heidelberg, pp.183-196, 2012, Studies in Computational Intelligence. hal-02501845

HAL Id: hal-02501845
https://hal.science/hal-02501845

Submitted on 8 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Model for Mapping Speech to Head Gestures in Human-Robot Interaction

Amir Aly and Adriana Tapus
Cognitive Robotics Lab, ENSTA-ParisTech, France
{*amir.aly, adriana.tapus*}*@ensta-paristech.fr*

***Abstract*— In human-human interaction, para-verbal and non-verbal communication are naturally aligned and synchronized. The difficulty encountered during the coordination between speech and head gestures concerns the conveyed meaning, the way of performing the gesture with respect to speech characteristics, their relative temporal arrangement, and their coordinated organization in a phrasal structure of utterance. In this research, we focus on the mechanism of mapping head gestures and speech prosodic characteristics in a natural human-robot interaction. Prosody patterns and head gestures are aligned separately as a parallel multi-stream HMM model. The mapping between speech and head gestures is based on Coupled Hidden Markov Models (CHMMs), which could be seen as a collection of HMMs, one for the video stream and one for the audio stream. Experimental results with Nao robot are reported.**

***Index Terms*— Coupled HMM, audio-video signal synchronization, signal mapping**

I. INTRODUCTION

Robots are more and more present in our daily lifes and the new trend is to make them behave more natural so as to obtain an appropriate social behavior and response. The work described in this paper presents a new methodology that allows the robot to automatically adapt its head gestural behavior to the user's profile (e.g. the user prosodic patterns) and therefore to produce a personalizable interaction. This work is based on some findings in the linguistic literature that show that head movements (e.g., nodding, turn taking system) support the verbal stream. Moreover, in human-human communication, prosody express the rhythm and intonation of speech and reflect various features of the speakers. These two communication modalities are strongly linked together and synchronized. Humans use gestures and postures as a communicative act. McNeill in [1] defines a gesture as a movement of the body synchronized with the flow of speech. The mechanism of the human natural alignment of the verbal and non-verbal characteristic patterns based on the work described in [2] shows a direct relationship between prosody features and gestures/postures, and constitute an inspiration for our work.

Recently, there has been a growth of interest in socially intelligent robotic technologies featuring flexible and customizable behaviors. Based on the literature in linguistics and psychology that suggests that prosody and gestural kinematics are synchronous and therefore strongly linked together, we posit that is important to have a robot behavior that integrates this element. Therefore, in this paper, we describe a new methodology for speech prosody and head gesture mapping for human-robot social interaction. The gesture/prosody modeled patterns are aligned separately as a parallel multi-stream HMM model and the mapping between speech and head gestures is based on Coupled Hidden Markov Models (CHMMs). A specific gestural behavior is estimated according to the incoming voice signal's prosody of the human interacting with the robot. This permits to the robot to adapt its behavior to the user profile (e.g. here the user prosodic patterns) and therefore to produce a personalizable interaction.

To the best of our knowledge, very little research has been dedicated to this research area. An attempt is described by the authors in [3] that present a robotic system that uses dance so as to explore the properties of rhythmic movement in general social interaction. Most of the existing works are related to computer graphics and interactive techniques. A general correlation between head gestures and voice prosody had been discussed in [4], [5]. The emotional content of the speech can also be correlated to some bodily gestures. In [6], it is discussed the relation between voice prosody and hand gestures, while [7] discusses the relation between the verbal and semantic content and the gesture. In [8], which is somehow closed to the discussed topic on this research, presents the relation between prosody changes and the orientation of the head (Euler angles). Moreover, authors in [9], proposed a mechanism for driving a head gesture from speech prosody.

Our work presents a framework for head gesture and prosody correlation for an automatic robot gesture production from interacting human user speech. The system is validated with the Nao robot in order to find out how naturalistic will be the driven head gestures from a voice test signal with respect to an interacting human speaker. The rest of the paper is organized as following: section II presents the applied algorithm for extracting the pitch contour of a voice signal; section III illustrates the detection of head poses and Euler angles; section IV describes speech and gesture temporal segmentation; section V presents the speech to head gesture coupling by using CHMMs; section VI resumes the results obtained; and finally, section VII concludes the paper.

II. PROSODIC FEATURES EXTRACTION

In human-robot interaction applications, the human voice signal can convey many messages and meanings, which should be understood appropriately by the robot in order to interact properly. Next, we describe the methodology used for pitch extraction.

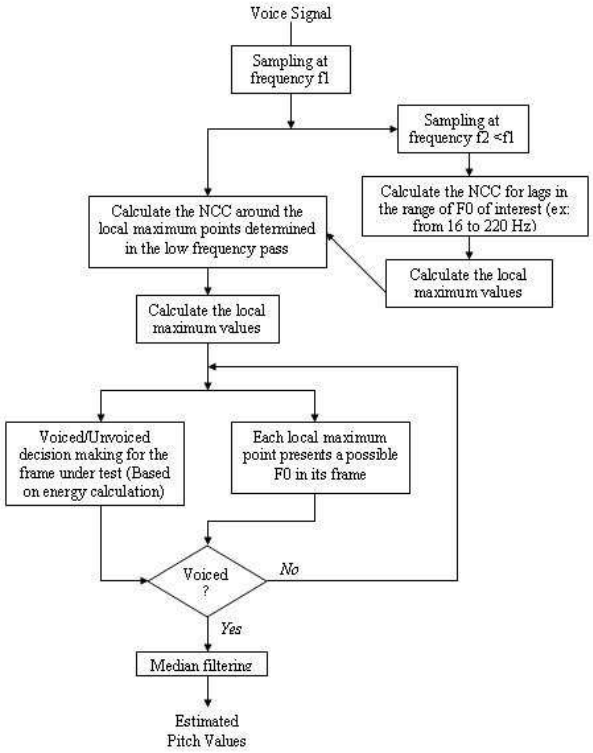


Fig. 1: Pitch Tracking

Talkin [10] defined the pitch as the auditory percept of tone, which is not directly measurable from a signal. Moreover, it is a nonlinear function of the signal's spectral and temporal energy distribution. Instead, another vocal characteristic, the fundamental frequency $F0$, is measured as it correlates well with the perceived pitch. Voice processing systems that estimate the fundamental frequency $F0$ often have 3 common processes: (1) Signal Conditioning; (2) Candidate Periods Estimation and (3) Post Processing.Signal preconditioning process is concerned by removing interfering signal components like noise and DC offset, while post processing process chooses the more likely candidate period in order to precisely estimate the fundamental frequency $F0$. Talkin in [10] developed the traditional (NCC) method in order to estimate reliably the voicing periods and the fundamental frequency $F0$ by considering all candidates simultaneously in a large temporal context. This methodology uses two pass normalized cross correlation (NCC) calculation for searching the fundamental frequency $F0$ which reduces the overall computation load with respect to the traditional (NCC) method. The procedures of the algorithm are illustrated in Figure 1.

In this work, we choose to express the characterizing vector of the voice signal in terms of the pitch and the intensity of the signal.

III. Head Pose Estimation

During social human-robot interaction, robots should be able to estimate the human head pose. This can help the robot to understand the human focus of attention and/or the meaning of the spoken message. The authors in [11] present a survey on the different existing algorithms for head poses estimation and detection.

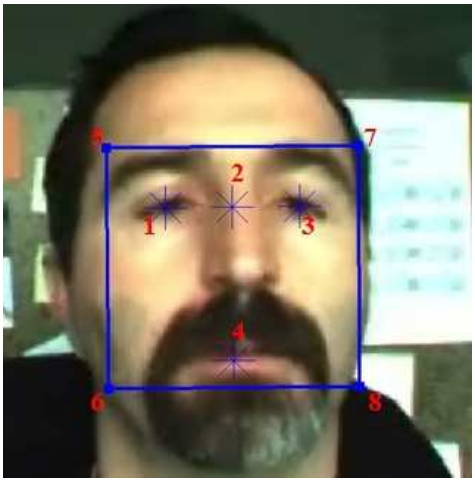


Fig. 2: Detecting the face rectangle that contains all salient points

Our methodology used for the detection of head poses is Viola and Jones algorithm [12]. After extracting the head region, the eyes are detected by the valley points' detection algorithm [13]. After detecting the location of eyes, it is possible to detect the location of the other salient points of the face using the geometry of the face [14].

For example, if the distance between the two eyes points $(1\&3)$ equals to D (see Figure 2), and point 2 is the midpoint between the eyes, then the mouth point 4 is located at a distance = $1.1D$ downwards from point 2.

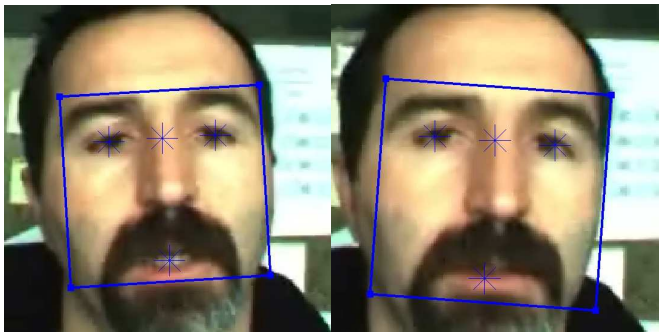
The X-Y coordinates of the rectangle surrounding the salient points of the face (points 5, 6, 7, and 8) (see Figure 2) could be precised as following:

- The difference between the Y-coordinates of points $(5\&1$ or $3\&7)$ = $0.2 * 1.8D$
- The difference between the X-coordinates of points $(5\&1$ or $3\&7)$ = $0.225 * 1.8D$

After calculating the coordinates of points $(5, 7)$, the coordinates of points $(6, 8)$ are directly calculated based on the vertical distance between points $(7\&8$ or $5\&6)$, which is equal to $1.8D$.

One of the problems that may appear when detecting the surrounding rectangle of the facial salient points is the rotation of the head clockwise and counterclockwise (see Figure 3). Therefore, the (X, Y) coordinates of the eyes has to be rotated first to (X^-, Y^-) before following the previous steps in order to precise the points of the surrounding rectangle, because the above metioned relations are valid when the eyes coordinates are in the same plane of the face (i.e., if the face is rotated, the coordinate of the eyes have also to be located in the rotated plane). The direction of rotation will be detected by calculating the slope (i.e., rotation angle θ) of the line passing by the two eyes using their (X, Y) coordinates. The rotation of the axes is described by the following equations:

$$X^- = Xcos\theta - Ysin\theta \quad (1)$$



(a) (b)

Fig. 3: Tracking Salient Face Details in Different Cases of Rotation: (a) Rotation clockwise; (b) Rotation counterclockwise

$$Y^{-} = X \sin\theta + Y \cos\theta \tag{2}$$

After calculating the coordinates of a face salient point in the rotated (X^{-}, Y^{-}) plane while the head is rotating clockwise or counterclockwise, it is important to make the inverse rotation each time to detect the location of this point in the (X, Y) plane. The pose of the head is calculated in terms of Euler angles: Pitch, Yaw, and Roll. These angles are calculated using the previously detected 8 salient facial points and their relative positions based on the geometry and symmetry of faces (Figure 2) as following:

$$\begin{aligned} Yaw_i = b_1[&\frac{(P2_{x,i} - C1_{x,i}) + (P2_{x,i} - C2_{x,i})}{2D_eyes_0} - \\ &\frac{(P2_{x,0} - C1_{x,0}) + (P2_{x,0} - C2_{x,0})}{2D_eyes_0}] + \\ b_2[&\frac{(P4_{x,i} - C1_{x,i}) + (P4_{x,i} - C2_{x,i})}{2D_eyes_0} - \\ &\frac{(P4_{x,0} - C1_{x,0}) + (P4_{x,0} - C2_{x,0})}{2D_eyes_0}] \end{aligned} \tag{3}$$

$$\begin{aligned} Pitch_i = b_3[&\frac{(P2_{y,i} - C3_{y,i}) + (P2_{y,i} - C4_{y,i})}{2D_eyes_0} - \\ &\frac{(P2_{y,0} - C3_{y,0}) + (P2_{y,0} - C4_{y,0})}{2D_eyes_0}] + \\ b_4[&\frac{(P4_{y,i} - C3_{y,i}) + (P4_{y,i} - C4_{y,i})}{2D_eyes_0} - \\ &\frac{(P4_{y,0} - C3_{y,0}) + (P4_{y,0} - C4_{y,0})}{2D_eyes_0}] \end{aligned} \tag{4}$$

where:

- $P2_{x,i}$, $P4_{x,i}$: the x coordinates of the midpoint between eyes, mouth point, respectively (see Figure 2), in frame i of the video.
- $P2_{y,i}$, $P4_{y,i}$: the y coordinates of the midpoint between eyes, mouth point, respectively (see Figure 2), in frame i of the video.
- $P2_{x,0}$, $P4_{x,0}$: the x coordinates of the midpoint between eyes, mouth point, respectively (see Figure 2), in frame 0 which is the reference frame in the video (1st frame).
- $P2_{y,0}$, $P4_{y,0}$: the y coordinates of the midpoint between eyes, mouth point, respectively (see Figure 2), in frame 0 which is the reference frame in the video (1st frame).

	Frame1 (reference)	Frame 2	Frame 3
Yaw	0	0.0016	0.0034
Pitch	0	0.00255	0.0075

TABLE I: Yaw and Pitch Initial Angles (Frames 1-3) Used for Calculation of Regression Values

- $C1_{x,i}$: the x coordinates of the center point between point 5 and point 6 (Figure 2), in frame i.
- $C2_{x,i}$: the x coordinates of the center point between point 7 and point 8 (Figure 2), in frame i.
- $C3_{y,i}$: the y coordinates of the center point between point 5 and point 7 (Figure 2), in frame i.
- $C4_{y,i}$: the y coordinates of the center point between point 6 and point 8 (Figure 2), in frame i.

The regression values b_1, b_2, b_3, and b_4 are constants throughout all the video frames. They are calculated by fixing the absolute values of Yaw and $Pitch$ angles in the second and third frames (according to empirical test) as shown in Table I. The substitution of the second and third values of $Pitch$ and Yaw in the equations 3 and 4, leads directly to the computation of the values of the constants b_1, b_2, b_3, and b_4.

The calculation of $Roll$ angle is straightforward, it depends on the coordinates of the midpoint between eyes (point 2) in frame i with respect to the reference frame [15], and it is clear that the value of $Roll$ angle in the first reference frame equals to 0.

$$Roll_i = \tan^{-1}(\frac{P2_{y,i}}{-P2_{x,i}}) - \tan^{-1}(\frac{P2_{y,0}}{-P2_{x,0}}) \tag{5}$$

IV. SPEECH AND HEAD GESTURE SEGMENTATION

The mapping between speech and head gestures is done by using the Coupled Hidden Markov Models (CHMMs), which could be seen as a collection of HMMs, one for the video stream and one for the audio stream. The advantage of this model over a lot of other topologies is its ability to capture the dual influences of each stream on the other one across time (see Figure 6). In the beginning, speech and head gestures streams are aligned separately as a parallel multi-stream HMM model.

The mapping between speech and head gestures is performed in 2 main steps: (1) the first is modeling the gesture sequence and the associated voice prosody sequence (in terms of their characteristic vectors) into two separate HMMs; (2) then after training both models, a correlation between the two HMMs is necessary so as to estimate a final head gesture states sequence given a speech test signal.

The HMM structure used in analyzing gestures (and similarly voice prosody) is indicated in Figure 4. It is composed of N parallel states, where each one represents a gesture composed of M observations. The goal of the transition between states S_{END} to S_{START} is to continue the transition between states from 1 to N (e.g., after performing gesture state 1, the model transfers from the transient end state to the start state to perform any gesture state from 2 to N in a sequential way and so on). In order to be able to model gestures/prosody, it is necessary to make a temporal segmentation of the video

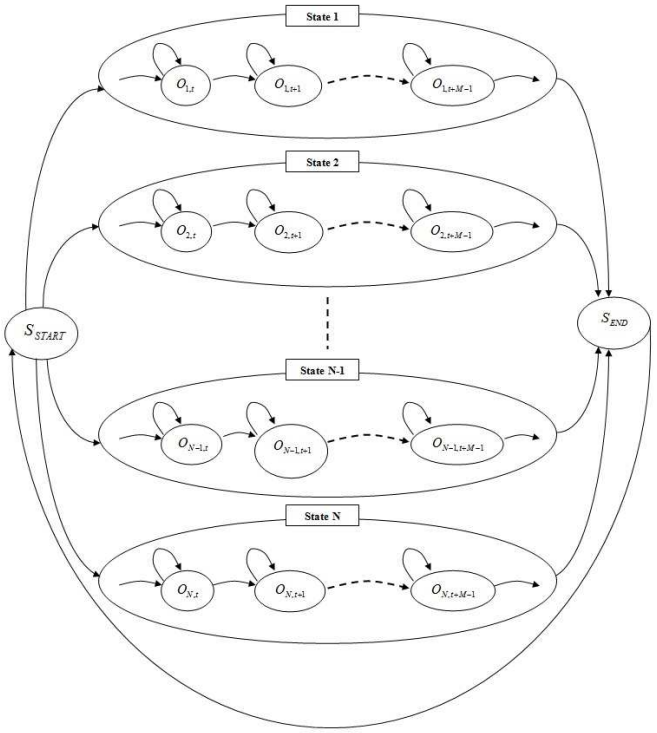


Fig. 4: HMM structure for gesture and prosody analysis

Trajectory Class	Trajectory State
1	pitch ↑ & intensity ↑
2	pitch ↑ & intensity ↓
3	pitch ↓ & intensity ↑
4	pitch ↓ & intensity ↓
5	Unvoiced segment

TABLE II: Voice Signal Segmentation Labels

content to detect the M number of observations in each state and the total number of states N.

A. Speech Temporal Segmentation

Speech is segmented as syllables presented by the states from 1 to N as indicated in Figure 4. The segmentation is performed by intersecting the inflection points (zeros crossing points of the rate of change of the curve) for both the pitch and intensity curves, beside the points that separate between the voiced and unvoiced segments of the signal (see Figure 5 for an example of pitch and intenisty curves). When comparing the two curves together, 5 different trajectory states can result [16] (see Table II).

The goal is to code each segment of the signal with its corresponding pitch-intensity trajectory class (e.g., a voice signal segment coding could be: $5, 3, 4, 2$, etc.). This segmental coding is used as label for CHMM training. The next step corresponds to segmenting the voice signal with its corresponding trajectory labeling into syllables. Arai and Greenberg in [17] defined the average duration of a syllable as 200 ms and this duration can increase or decrease according to the nature of the syllable as being short or long. Practical tests proved that

Trajectory Class	Trajectory State (Rate of Change)
1	Yaw ↑ & Pitch ↑
2	Yaw ↑ & Pitch ↓
3	Yaw ↓ & Pitch ↑
4	Yaw ↓ & Pitch ↓
5	No change

TABLE III: Gesture Segmentation Labels

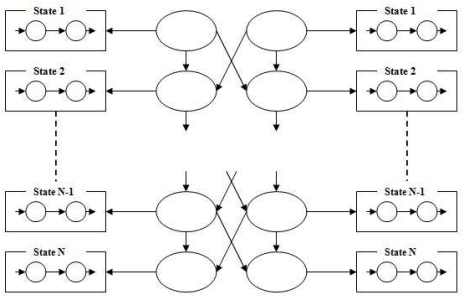


Fig. 6: Coupled Hidden Markov Model CHMM lag-1 Structure

within a syllable of duration varying from 180 ms to 220 ms, the average number of trajectory classes in its corresponding pitch and intensity curves is around 5. Therefore, given the voice signal with its segments coded by the corresponding pitch-intensity trajectory labels, each 5 segments of the signal will create a syllable state (from 1 to N) and the corresponding 5 labels will be the observations M within the syllable state.

B. Gestures Temporal Segmentation

The average duration for making gestures, in general, varies between 0.1 to 2.5 seconds according to the speed and the performed gesture as being pointing or head gesture for example. In case of head gestures, the average duration of performing a gesture will be limited to 0.4 seconds [18, 19]. In our case, the camera used to capture the gestures had the ability of capturing 30 frames/second, and therefore we can estimate to 12 frames the average number of frames sufficient to characterize a gesture.

Similarly to the speech temporal segmentation (see Section IV-A), gesture temporal segmentation is performed by comparing the 9 trajectory classes according to the sinusoidal evolution of the extracted angles curves. However, the mechanical characteristics of our platform (NAO robot) are limited only to pitch and yaw movements, therefore introducing only 5 trajectory classes (see Table III). In the context of the CHMM model each group of 12 frames will form a complete gesture state from 1 to N, and the corresponding coding labels will constitute the observations within the gesture state.

V. SPEECH TO HEAD GESTURE COUPLING

A typical CHMM structure is shown in Figure 6, where the circles present the discrete hidden nodes/states while the rectangles present the observable continuous nodes/states, which contain the observation sequences of voice and gestures characteristics.

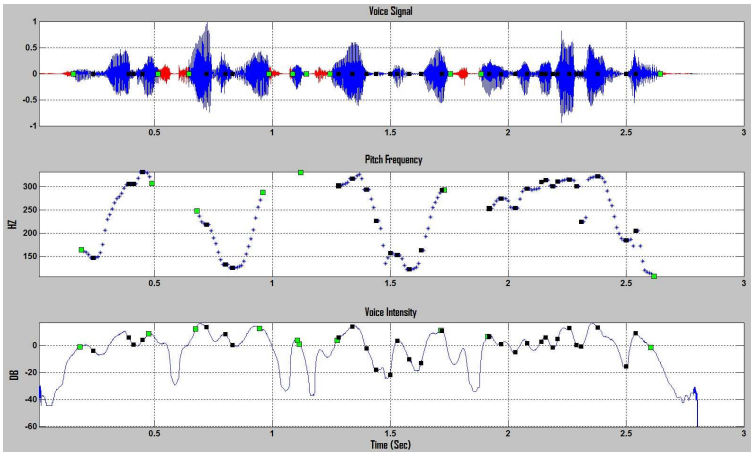


Fig. 5: Speech, Pitch and Intensity Curves (The red parts in the voice signal are the unvoiced parts, while blue parts are the voiced parts of the signal. The black points depict the inflection points of the signal, while green points represent the separating points between the unvoiced and the voiced segments.)

According to the sequential nature of gestures and speech, the CHMM structure is of type lag-1 in which couple (backbone) nodes at time t are conditioned on those at time $t-1$ [20, 21, 22]. A CHMM model λ_C is defined by the following parameters:

$$\pi_0^C(i) = P(q_1^C = S_i) \tag{6}$$

$$a_{i|j,k}^C = P(q_t^C = S_i | q_{t-1}^{audio} = S_j, q_{t-1}^{video} = S_k) \tag{7}$$

$$b_t^C(i) = P(O_t^C | q_t^C = S_i) \tag{8}$$

where $C \in \{audio, video\}$ denotes the audio and visual channels respectively, and q_t^C is the state of the coupling node in the c_{th} stream at time t [23, 24].

The training of this model is based on the maximum likelihood form of the expectation maximization (EM) algorithm. Supposing there are 2 observable sequences of the audio and video states $O = \{A_{1..N}, B_{1..N}\}$ where $A_{1..N} = \{a_1, \cdots, a_N\}$ is the set of observable states in the first audio sequence, and similarly $B_{1..N} = \{b_1, \cdots, b_N\}$ is the set of observable states in the second visual sequence, and $S = \{X_{1..N}, Y_{1..N}\}$ is the set of states of the couple nodes at the first audio chain and the second visual chain respectively [21, 22]. The expectation maximization algorithm finds the maximum likelihood estimates of the model parameters by maximizing the following function [22]:

$$f(\lambda_C) = P(X_1)P(Y_1)\prod_{t=1}^{T} P(A_t|X_t)P(B_t|Y_t) \tag{9}$$
$$P(X_{t+1}|X_t, Y_t)P(Y_{t+1}|X_t, Y_t), 1 \leq T \leq N$$

where:

- $P(X_1)$ and $P(Y_1)$ are the prior probabilities of the audio and video chains respectively
- $P(A_t|X_t)$ and $P(B_t|Y_t)$ are the observation densities of the audio and video chains respectively
- $P(X_{t+1}|X_t, Y_t)$ and $P(Y_{t+1}|X_t, Y_t)$ are the couple nodes transition probabilities in the audio and video chains.

The training of the CHMM differs from the standard HMM in the expectation step (E) while they are both identical in the maximization step (M) which tries to maximize equation 9 in terms of the expected parameters [25].The expectation step of the CHMM is defined in terms of the forward and backward recursion. For the forward recursion we define a variable for the audio and video chains at $t = 1$:

$$\alpha_{t=1}^{audio} = P(A_1|X_1)P(X_1) \tag{10}$$

$$\alpha_{t=1}^{video} = P(B_1|Y_1)P(Y_1) \tag{11}$$

Then the variable α is calculated incrementally at any arbitrary moment t as follows:

$$\alpha_{t+1}^{audio} = P(A_{t+1}|X_{t+1}) \int\int \alpha_t^{audio}\alpha_t^{video} P(X_{t+1}|X_t, Y_t)dX_t dY_t \tag{12}$$

$$\alpha_{t+1}^{video} = P(B_{t+1}|Y_{t+1}) \int\int \alpha_t^{audio}\alpha_t^{video} P(Y_{t+1}|X_t, Y_t)dX_t dY_t \tag{13}$$

Meanwhile, for the backwards direction there is no split in the calculated recursions which can be expressed as follows:

$$\beta_{t+1}^{audio,video} = P(O_{t+1}^N|S_t) = \int\int P(A_{t+1}^N, B_{t+1}^N|X_{t+1}, Y_{t+1}) P(X_{t+1}, Y_{t+1}|X_t, Y_t)dX_{t+1}dY_{t+1} \tag{14}$$

After combining both forward and backwards recursion parameters, an audio signal will be tested on the trained model, generating a synthesized equivalent gesture that most likely fit the model. The generated gesture sequence is determined when the change in the likelihood is below a threshold.

VI. EXPERIMENTAL RESULTS

The experimental testbed used in this study is the humanoid robot Nao developed by Aldebaran Robotics. For the training and testing, we used the MVGL-MASAL gesture-speech

Synthesized/Real Gesture Classes	1	2	3	4	5
1	25	13	13	6	36
2	3	29	5	3	28
3	2	6	20	8	25
4	4	2	5	40	33
5	20	18	30	43	351

TABLE IV: Confussion matrix of the original and synthesized trajectories' classes

Turkish database [9]. The database is composed of 4 videos of different durations that go from 6 to 8 minutes. It contains the audiovisual information of different subjects instructed to tell stories to children audience. We use one part of the database for the training of the models and the other part for the testing. The audio signals are extracted and then they are processed in order to extract the relevant prosodic characteristics. The proposed speech to gesture mapping methodology was tested on the database using cross validation algorithm. The system was trained on the audio/visual sequences of 3 videos from the database, and then tested on the audio sequence of the 4th video. The corresponding generated gestures are compared to the natural gesture sequence in the video of test and an average score of 62% was found in terms of the similarity of trajectory classes.

Table IV depicts the confussion matrix between the original and synthesized gesture labels trajectories. The confussion matrix reveals that the trajectory state 5 in which there would be no change in the Yaw and $Pitch$ angles is the dominant trajectory class. This can be a result of the smoothing processes and/or of the precision of Euler angles extracting algorithm; however this will not cause unnaturaleness when the robot and the human are interacting in long conversations.

After calculating the score of similarity between the trajectory labels of the original and the synthesized signals, it is important to generate the corresponding Yaw and $Pitch$ curves for the head motion and compare them to the original curves by calculating the total average root mean square (RMS) error between the corresponding curves points. The RMS errors found between the generated Yaw and $Pitch$ curves with respect to the original curves are 10% and 12% respectively.

In fact, the obtained score 62% and the RMS errors between the original and the synthesized curves can be considered a reasonable result, because the duration and the surrounding environement conditions of the test video and the training videos set were similar. Also, the speaker's tonality in all training and test videos were similar. However, we don't know yet the score we will obtain in real applications where the robot will be tested under different condiditions. The performed head gestures could differ in the amplitude or the direction from one person to another without hindering the transfer of the meaning of the gesture message between interacting humans and similarly, between the interacting robot and human. Figures 7 and 8 show a comparison between a part of the original and synthesized pitch and yaw curves (after being smoothed by a median filter) of the test video from the database.
A video of the speech-gesture mapping system with Nao robot is available at: `http://www.ensta-paristech.fr/~tapus/HRIAA/media`.

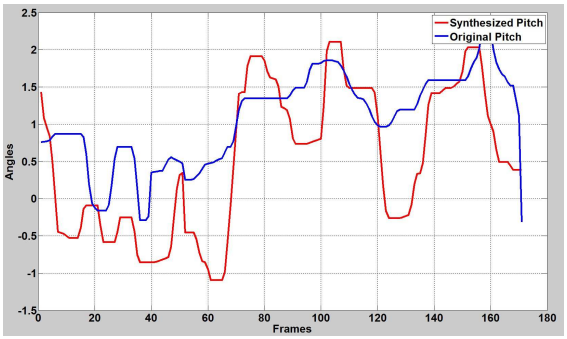


Fig. 7: View of the original (blue curve) and synthesized (red curve) Pitch angles of a part of the test video

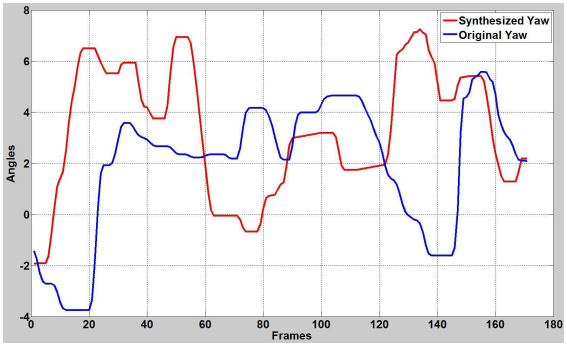


Fig. 8: View of the original (blue curve) and synthesized (red curve) Yaw angles of a part of the test video

VII. CONCLUSIONS

This research focuses on synthesizing head gestures based on speech characteristics (e.g., pitch and intensity of the signal). Our mapping system is based on the Coupled Hidden Markov Model (CHMM) that tries to find a coupling joint between audio and visual sequences. The audio sequence is composed of parallel states presenting the syllables and each syllable is composed of a specific number of observations (M=5, in our case). Meanwhile, the video sequence has the same parallel construction where the states present the gestures and each state is composed of another specific number of observations determined experimentally (M=12, in our case). After training the CHMM on audio-visual sequences from a database, and when a test audio signal is generated, the system tries to find a corresponding sequence of gestures based on its own experience learnt during the training phase. The generated gesture sequence is the sequence that achieves the maximum likelihood estimation with the speech test signal. Our system shows a score of 62%, which measures the similarity between the original gesture sequence labels and the synthesized gesture sequence labels, over a test video of 8 minutes. This can be considered a good score. The proposed system is able to generate appropriate robot head gesture from speech input, which allows it to produce an automatic natural robot behavior that is almost completely absent from present-day

human-robot interactions. Further work will focus on creating a triadic alignment between the speech, head gestures, and hand gestures in different human-robot interactional contexts that will allow the robot to interact naturally under different conditions.

REFERENCES

[1] D. McNeill, *Hand and mind : what gestures reveal about thought*. Chicago, USA: Chicago : University of Chicago Press, 1992.

[2] F. P. Eyereisen and J. D. D. Lannoy, *Gestures and Speech: Psychological Investigations.* Cambridge University Press, 1991.

[3] M. P. Michalowski, S. Sabanovic, and H. Kozima, "A dancing robot for rhythmic social interaction," in *Proceedings of the Human-Robot Interaction Conference*, Arlington, USA, mar 2007, pp. 89–96.

[4] K. 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.

[5] T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999, pp. 1279–1282.

[6] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," in *Proceedings of the European Signal Processing Conference(EUSIPCO)*, vol. 1, 2005, pp. 75–78.

[7] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K. McCullough, "Gesture cues for conversational interaction in monocular video," in *Proceedings of the ICCV*, 1999, pp. 64–69.

[8] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: facial movements accompanying speech," in *Proceedings of IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 381–386.

[9] M. E. Sargn, Y. Yemez, E.Erzin, and A. M. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, 2008.

[10] D. Talkin, "A robust algorithm for pitch tracking," in *Speech Coding and Synthesis*. W B Kleijn, K Paliwal eds, Elsevier, 1995, pp. 497–518.

[11] E. M.Chutorian and M.M.Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[13] K. Wong, K. Lam, and W. Siu, "A robust scheme for live detection of human faces in color images," *Signal Processing: Image Communication*, vol. 18, no. 2, pp. 103–114, 2003.

[14] K. W. Wong, K. I. Lam, and W. Siu, "An efficient algorithm for human face detection and facial feature extraction under different conditions," *Pattern Recognition*, vol. 34, no. 10, pp. 1993–2004, 2000.

[15] B. Yip, W. Y. Siu, and S. Jin, "Pose determination of human head using one feature point based on head movement," in *Proceedings of IEEE Int. Conf. on Multimedia and Expo (ICME)*, vol. 2, 2004, pp. 1183–1186.

[16] F. Ringeval, J. Demouy, G. S. andM. Chetouani, L. Robel, J. Xavier, and D. C. Plaza, "Automatic intonation recognition for the prosodic assessment of language impaired children." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 99, pp. 1–15, 2010.

[17] T. Arai and S. Greenberg, "The temporal properties of spoken japanese are similar to those of english," in *Proceedings of Eurospeech*, Rhodes, Greece, 1997, pp. 1011–1114.

[18] K. Nickel and R. Stiefelhagen, "Real-time recognition of 3d-pointing gestures for human-machine-interaction," in *Proceedings of DAGM-Symposium*, Magdeburg, Germany, 2003, pp. 557–565.

[19] S. A. Moubayed and J. Beskow, "Effects of visual prominence cues on speech intelligibility," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, Norwich, UK, 2009.

[20] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, vol. 77, no. 2, 1989, pp. 257–286.

[21] I.Rezek, P. Sykacek, and S. Roberts, "Coupled hidden markov models for biosignal interaction modelling," in *Proceedings of the International Conference on Advances in Medical Signal and Information Processing (MEDSIP)*, 2000.

[22] I. Rezek and S. J. Roberts, "Estimation of coupled hidden markov models with application to biosignal interaction modelling," in *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing (NNSP)*, Sydney, Australia, 2000.

[23] A. V. Nean, L. Liang, X. Pi, X. Liu, and C. Mao, "A coupled hidden markov model for audio-visual speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Orlando, USA, 2002, pp. 2013–2016.

[24] L. Liang, X. Liu, X. Pi, Y. Zhao, and A. V. Nean, "Speaker independent audio-visual continuous speech recognition," in *Proceedings of the International Conference on Multimedia and Expo (ICME)*, vol. 2, Lausanne, Switzerland, 2002, pp. 25–28.

[25] W. Penny and S. Roberts, "Gaussian observation hidden markov models for eeg analysis," in *Technical Report TR-98-12*, Imperial College London, UK, 1998.