We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS

BOOK CITATION INDEX

INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Imitation Learning Based Talking Heads in Humanoid Robotics

Enzo Mumolo and Massimiliano Nolich
DEEI, Universitá degli Studi di Trieste
Italy

1. Introduction

The main goal of this Chapter is to describe a novel approach for the control of Talking Heads in Humanoid Robotics.

In a preliminary section we will discuss the state of the art of the research in this area. In the following sections we will describe our research results while in the final part some experimental results of our approach are reported. With the goal of controlling talking heads in mind, we have developed an algorithm which extracts articulatory features from human voice. In fact, there is a strong structural linkage between articulators and facial movements during human vocalization; for a robotic talking head to have human-like behavior, this linkage should be emulated. Exploiting the structural linkage, we used the estimated articulatory features to control the facial movements of a talking head. Moreover, the articulatory estimate is used to generate artificial speech which is - by construction - synchronized with the facial movements.

Hence, the algorithm we describe aims at estimating the articulatory features from a spoken sentence using a novel computational model of human vocalization. Our articulatory features estimator uses a set of fuzzy rules and genetic optimization. That is, the places of articulation are considered as fuzzy sets whose degrees of membership are the values of the articulatory features. The fuzzy rules represent the relationships between places of articulation and speech acoustic parameters, and the genetic algorithm estimates the degrees of membership of the places of articulation according to an optimization criteria. Through the analysis of large amounts of natural speech, the algorithm has been used to learn the average places of articulation of all phonemes of a given speaker.

This Chapter is based upon the work described in [1]. Instead of using known HMM based algorithms for extracting articulatory features, we developed a novel algorithm as an attempt to implement a model of human language acquisition in a robotic brain. Human infants, in fact, acquire language by imitation from their care-givers. Our algorithm is based on imitation learning as well.

Nowadays, there is an increasing interest in service robotics. A service robot is a complex system which performs useful services with a certain degree of autonomy. Its intelligence emerges from the interaction between data gathered from the sensors and the management algorithms. The sensorial subsystem furnishes environment information useful for motion tasks (dead reckoning), auto-localization and obstacle avoidance in order to introduce

reactiveness and autonomy. Humanoid robotics has been introduced for enabling a robot to give better services. A humanoid, in fact, is a robot designed to work with humans as well as for them. It would be easier for a humanoid robot to interact with human beings because it is designed for that. Inevitably, humanoid robots tend to imitate somehow the form and the mechanical functions of the human body in order to emulate some simple aspects of the physical (i.e. movement), cognitive (i.e. understanding) and social (i.e. communication, language production) capabilities of the human beings. A very important area in humanoid robotics is the interaction with human beings, as reported in [2]. Reference [2] describes the Cog project at MIT and the related Kismet project which have been developed under the hypothesis that *humanoid intelligence requires humanoid interactions with the world.* In this chapter we deal with human-humanoid interaction by spoken language and visual cues, i.e. with talking heads in humanoid robotics. In fact, human-like artificial talking heads can increase a person's willingness to collaborate with a robot and helps create the social aspects of the human-humanoid relationship. The long term goal of the research in talking heads for a humanoid is to develop an artificial device which mechanically emulates the human phonatory organs (i.e. tongue, glottis, jaw) such that unrestricted natural sounding speech is generated. The device will be eventually contained in an elastic envelop which should resemble and move as a human face. Several problems have to be addressed towards this goal. First of all the complex phenomena in the human vocal organs should be mechanically emulated to produce a good artificial speech. Second, the control of the mechanical organs must be temporally congruent with human vocalization and this can be very complex to manage. The result is that at the state of the art the quality obtained with mechanical devices is only preliminar, yet interesting. For these reasons, and waiting that the mechanical talking heads reach a sufficient quality, we just emulate a talking head in a graphical way while the artificial speech is algorithmically generated.

It is worth emphasizing now the objective of this Chapter, which is the description of a novel algorithm to the control of a humanoid talking head and to show some related experimental results. This means that we estimate a given set of articulatory features to control the articulatory organs of a humanoid head, either virtual or mechanical. Two applications are briefly described: first, a system which mimicry human voice and, second, a system that produces robotic voice from unrestricted text, both of them with the corresponding facial movements.

Although almost all the animals have voices, only human beings are able to use words as mean of verbal communication. As a matter of fact, voice and the related facial movements are the most important and effective method of communication in our society. Human beings acquire control methods of their vocal organs with an auditory feedback mechanism by repeating trials and errors of hearing and uttering sounds. Humans easily communicate each other using vocal languages. Robotic language production for humanoids is much more difficult. At least three main problems must be solved. First, concepts must be transformed into written phrases. Second, the written text must be turned into a phonemic representation and, third, an artificial utterance must be obtained from the phonemic representation. The former point requires that the robot is aware of its situational context. The second point means that graphemic to phonemic transformation is made while the latter point is related to actual synthesis of the artificial speech.

Some researchers are attempting to reproduce vocal messages using mechanical devices. For instance, at Waseda University researchers are developing mechanical speech production systems for talking robots called WT-1 to WT-5, as reported in [3, 4, 5, 6, 7, 8, 9, 10, 11]. The authors reported that they can generate Japanese vowels and consonants (stops, fricatives

and nasal sounds) reasonably clearly, although not all the utterances sound natural yet. On the other hand, the researchers of the robot Kismet [12] are expanding their research efforts on naturalness and perception of humanness in robots. An important step toward talking heads development is to estimate accurate vocal tract dynamic parameters during phonation. It is known, in fact, that there is a very high correlation between the vocal tract dynamic and the facial motion behavior, as pointed out by Yehia et al. [13]. For a mechanical talking robot, the artificial head should have human like movements during spoken language production by the robot, provided that the artificial head is tied to the vocal tract by means of some sort of elastic joint. In any case, the mechanical vocal tract should be dynamically controlled to produce spoken language. This requires enough knowledge of the complex relations governing the human vocalization. Until now, however, there has been no comprehensive research on the speech control system in the brain, and thus, speech production is still not clearly understood. This type of knowledge is pertaining to articulatory synthesis, which includes the methods to generate speech from dynamic configuration of the vocal tract (articulatory trajectory).

Our algorithm is based on imitation learning, i.e. it acquires a vocalization capability in a way similar to human development; in fact, human infants learn to speak through interaction by imitation with their care-givers. In other words, the algorithm tries to mimic some input speech according to a distance measure and, in this way, the articulatory characteristics of the speaker who trained the system are learned. From this time on, the system can synthesize unrestricted text using the articulatory characteristics estimated from a human speaker. The same articulatory characteristics are used to control facial movements using the correlation between them. When implemented on a robot, the audio-synchronized virtual talking head give people the sense that the robot is talking to them. As compared to other studies, our system is more versatile, as it can be easily adapted to different languages provided that some phonetic knowledge of that language is available. Moreover, our system uses an analysis-by-synthesis parameter estimation and it therefore makes available an artificial replica of the input speech which can be useful in some circumstances.

The rest of this chapter is organized as follows. In Section 2 some previous work in graphical and mechanical talking heads is briefly discussed. In Section 3 the imitation learning algorithm based on fuzzy model of speech is presented, and the genetic optimization of articulatory parameters is discussed. In Section 4 some experimental results are presented; convergence issues, acoustical and articulatory results are reported. In this Section also some results in talking head animation are reported. Finally, in Section 5 some final remarks are reported.

2. Previous work on talking heads

The development of facial models and of virtual talking heads has a quite long history. The first facial model was created by F.Parke in 1972 [14]. The same author in 1974 [15] produced an animation demonstrating that a single model would allow representation of many expressions through interpolated transitions between them. After this pioneer work, facial models evolved rapidly into talking heads, where artificial speech is generated in synchrony with animated faces. Such developments were pertaining to the human-computer interaction field, where the possibility to have an intelligent desktop agent to interact with, a virtual friend or a virtual character for interacting with the web attracted some attention. As regards these last points, Lundeberg and Beskow in [16] describe the creation of a talking

head for the purpose of acting as an interactive agent in their dialogue system. The purpose of their dialogue system is to answer questions on chosen topics using a rich repertoire of gestures and expressions, including emotional cues, turntalking signals and prosodic cues such as punctuators and emphasisers. Studies of user reactions indicated that people had a positive attitude towards the agent. The FAQBot describes in [17] a talking head which answers questions based on the topics of FAQs. The user types in a question, the FAQBot's AI matches an answer to the question, and the talking head speaks, providing the answer to the user.

Other applications of talking head have been envisaged in many other field, such as the improvement of language skills, education and entertainment as reported in [18]. As regards entertainment, interactive input devices (e.g. Facial Animation, instrumented body suits, data gloves and videobased motion-tracking systems) are often used to drive the animation. In [19, 20] approaches for acquiring the expressions of the face of a live actor, and to use that information to control facial animation are described. Also the MIT Media Laboratory Perceptual Computing Section has developed systems that allow realtime tracking and recognition of facial expressions as reported in [21, 22].

The field of assistive technology has been also explored: in [23] a set of tools and technologies built around an animated talking head to be used in daily classroom activities with profoundly deaf children has been described. The students enters commands using speech, keyboard and mouse while the talking head responds using animated face and speech synthesis. On the other hand, if accurate face movements are produced from an acoustic vocal message uttered by a human, important possibilities of improving a telephone conversation with added visual information for people with impaired hearing conversation are introduced [24].

2.1 Social implication of a talking head in humanoid robotics

A brief description of social implication of talking heads is worth of because many current research activities are dealing with that. Socially-situated learning tutors with robot-directed speech is discussed in [25]. The robot's affective state and its behavior are influenced by means of verbal communication with a human care-giver via the extraction of particular cues typical of infant-directed speech as described in [26]. Varshavskaya in [27] dealt with the problem of early concept and vocal label acquisition in a sociable robot. The goal of its system was to generate "the kind of vocal output that a prelinguistic infant may produce in the age range between 10 and 12 months, namely emotive grunts, canonical babblings, and a formulaic proto-language". The synthesis of a robotic proto-language through interaction of a robot either with human or a robotic teacher was also investigated in [28].

Other authors (for example [29, 30, 31, 32]) have investigated the underlying mechanisms of social intelligence that will allow it to communicate with human beings and participate in human social activities. In [33] it was described the development of an infant-like humanoid robot (Infanoid) for situating a robot in an environment equivalent to that experienced by a human infant. This robot has a human-like sensori-motor systems, to interacts with its environment in the same way as humans do, implicitly sharing its experience with human interlocutors, sharing with humans the same environment [32]. Of course, talking heads have a very important role in this developments.

2.2 Graphical talking heads

In achieving the above goals, facial animation synthesis often takes two approaches: 3D mesh based geometry deformations and 2D image manipulations. In a typical 3D mesh approach, a mesh model is prepared which contains all the parameters necessary for the subsequent animations. Noh and Neumann describe in [34] several issues of graphical talking heads. The model is animated by mesh node displacements based on motion rules specified by deformation engine such as vector muscles [35, 36], spring muscles [37, 38], free form deformations [39], volume morphing [40], or simple interpolation [41]. If only the frontal movements are required, like in application based on mouth animation only, a 2D image-based approach is sufficient. 2D based approaches are also attractive for lip reading.

Ezzat et al. described in [42] a text to audiovisual translator using image warping and morphing between two viseme images. Gao et al. described in [43] new mouth shapes by linear combinations of several base images. Koufakis et al. describe in [44] how to use three basis images captured from different views and synthesize slightly rotated views of the face by linear combination of these basis images. Cosatto et al. in [45] describe their algorithm based on collecting various image samples of a segmented face and parameterize them to synthesize a talking face. By modeling different parts of the face from different sample segments, synthesized talking faces also exhibit emotions from eye and eyebrow movements and forehead wrinkles. Methods that exploit a collection of existing sample images must search their database for the most appropriate segments to produce a needed animation.

Other work, in particular that described in [43, 44, 46] used Mesh based texture mapping techniques. Such techniques are advantageous because warping is computed for relatively few control points.

Finally, there have been attempts to apply Radial Basis Functions (RBF) to create facial expressions. In [47] one of these approaches is described. Most approaches warped a single image to deform the face. However, the quality obtained from a single image deformation drops as more and more distortions are required. Also, single images lack information exposed during animation, e.g., mouth opening. Approaches without RBF using only single images have similar pitfalls.

2.3 Mechanical talking heads in humanoid robotic

When applied to a robot, mechanical talking heads give people a compelling sense that the robot is talking to them at a higher level as compared to virtual ones. At Waseda University the talking robots WT-1 to WT-5 [3, 4, 5, 6, 7, 8, 9, 10, 11] have been reported to the scientific community starting from 2000. The WT1 to 5 talking heads have been developed for generated human vocal movements and some human-like natural voices. For emulating the human vocalization capability, these robots share human-like organs as lungs, vocal cords, tongue, lips, teeth, nasal cavity and soft palate. The robots have increasing features, as an increasing number of degree of freedom (DOF) and the ability to produce some human-like natural voices. The anthropomorphic features were further improved in WT-4 and WT-5. WT-4 had a human-like body to make the communication with a human more easily, and has an increased number of DOF. This robot aimed to mimic continuous human speech sounds by auditory feedback by controlling the trajectory and timing. The mechanical lips and vocal cords of WT-5 have similar size and biomechanical structure as humans. As a result, WT-5 could produce Japanese vowels and consonant sounds (stops, fricatives and nasals) of 50 Japanese sounds for human-like speech production. Also at Kagawa University

researchers dealt with talking heads from about the same years [48, 49, 50, 51]. They developed and improved mechanical devices for the construction of advanced human vocal systems with the goals to mimicry human vocalization and for singing voice production. They also developed systems for open and closed loop auditory control.

3. An algorithm for the control of a talking head using imitation learning

The block diagram of the algorithm described in this paper is reported in Fig. 1. According to the block diagram, we now summarize the actions of the algorithm.
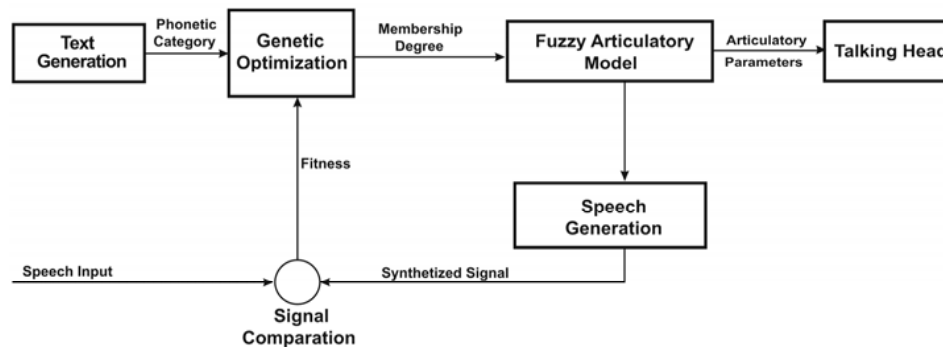

Figure 1. Block diagram of the genetic-fuzzy imitation learning algorithm

First, the operator is asked to pronounce a given word; the word is automatically selected from a vocabulary defined to cover all the phonemes of the considered language. Phonemes are described through the 'Locus Theory' [52].

In particular, the transition between two phonemes is described using only the target one. For example we do not consider that in the transition, say, 'no', the phoneme /o/ comes from the phoneme /n/ but only an average target configuration of phoneme /o/ is considered.

	Round	Open	Anterior	Sonorant	Bilaial	Labiodental	Alveolar	Prepalatal	Palatal	Vibrant	Dental	Velar
/sil/												
/a/												
⋮						⋮						
/n/	0	0	0	0.8	0.6	0	0	0	0.8	0	0.4	0
/o/	0.6	0.8	0.5	0.9	0	0	0	0	0	0	0	0
⋮						⋮						
/z/												

Figure 2. Membership degrees of phoneme transitions coming from 'any' phoneme. The membership degrees for the utterance 'no' are shown.

Each phoneme is therefore described in terms of articulatory as described in Fig. 2. The number corresponding to the articulatory feature is the degree of membership of that

feature. These degrees of membership are obtained through genetic optimization, as described shortly. For example, in Fig. 3 a string of membership values for the utterance /no/ obtained with genetic optimization is reported.

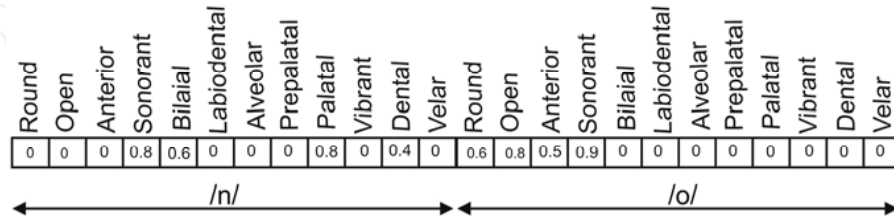| Round	Open	Anterior	Sonorant	Bilaial	Labiodental	Alveolar	Prepalatal	Palatal	Vibrant	Dental	Velar	Round	Open	Anterior	Sonorant	Bilaial	Labiodental	Alveolar	Prepalatal	Palatal	Vibrant	Dental	Velar
0	0	0	0.8	0.6	0	0	0	0.8	0	0.4	0	0.6	0.8	0.5	0.9	0	0	0	0	0	0	0	0

/n/ /o/

Figure 3. String of membership degrees for the utterance 'no'

To summarize, the learning mechanism of the articulatory parameters works as follows: the operator, acting as care-giver, pronounces a word and the robot generates an artificial replica of the word based on the articulatory and acoustic estimation. This process iterates until the artificial word matches the original one according to the operator's judgement. At this point the robot has learnt the articulatory movements of the phoneme contained in the word. The operator must repeat this process for a number of words. After these phases, the speech learning process is completed.

The synthesis of the output speech is performed using a reduced Klatt formant synthesizer [53], whose block diagram is represented in Fig. 4.
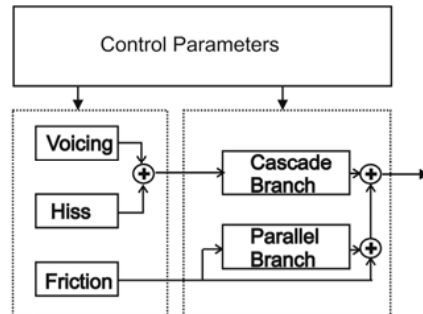

Figure 4. Simplified Klatt model used in this work

time	AV	AH	AF	F1	B1	F2	B2	F3	B3	A2F	A3F	A4F	A5F	A6F	AB
0	0	0	0	500	60	1500	90	2500	150	0	0	0	0	0	0
5	0	0	0	505	60	1441	90	2440	150	0	0	0	0	0	0
10	0	0	0	510	60	1381	91	2379	150	0	0	0	0	0	0
15	0	0	0	516	60	1322	91	2319	150	0	0	0	0	0	0
20	0	0	0	521	60	1262	92	2258	151	0	0	0	0	0	0

Figure 5. Acoustic parameters of a vowel sound for the first 20 ms

This system is basically composed by a parallel filter bank for the vocal tract modeling for unvoiced sounds and a cascade of filters for the vocal tract modeling for voiced sounds. It is controlled by fifteen parameters, namely the first three formants and bandwidths, the bypass AB, the amplitude AV for voiced sounds and the amplitudes AF, AH and A2F-A6F

for the fricative noise generator, updated every 5 ms. For instance, in Fig. 5 we report the parameters of a vowel sound in the very first interval, 20 ms long.

Since the fuzzy rules, however, describe the locus of the acoustical parameters, a model of the parameters profiles has been introduced. The profile of each synthesis parameter 'p' is described with four control features, namely the initial and final intervals $I(p)$ and $F(p)$, the duration $D(p)$ and the locus $L(p)$, as reported Fig. 6.
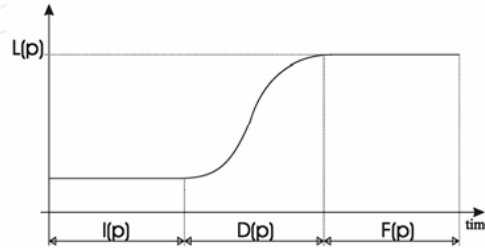

Figure 6. Synthesis parameters profiling in terms of Initial, Final, Duration and Locus fuzzy variables, $I(p)$, $F(p)$, $D(p)$ and $L(p)$, respectively

The $I(p)$ control feature determines the length of the starting section of the transition, whose slope and target values are given by the $D(p)$ and $L(p)$ features. The parameter holds the value specified by their locus for an interval equal to $F(p)$ ms; however, if other parameters have not completed their dynamic, the final interval $F(p)$ is prolonged. The $I(p)$, $F(p)$, and $D(p)$ parameters are expressed in milliseconds, while the target depends on what synthesis control parameter is involved; for example, for frequencies and bandwidths the locus is expressed in Hz, while for amplitudes in dB. It is worth noting that the initial values of the transition depend on the previous target values.

3.1 Phoneme and Control Parameters Fuzzification

As mentioned above, the phonemes are classified into broad classes by means of the manner of articulation; then, the place of articulation is estimated by genetic optimization. Therefore, each phoneme is described by an array of nineteen articulatory features, six of them are boolean variables and represent the manner of articulation and the remaining thirteen are fuzzy and represent the place of articulation.

Representing the array of features as (vowel, plosive, fricative, affricate, liquid, nasal | any, rounded, open, anterior, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar), the /a/ phoneme, for example, can be represented by the array:

$$[1,0, 0, 0,0,0|1, 0.32,0.9, 0.12,1,0,0,0, 0, 0,0,0,0]$$

indicating that /a/ is a vowel, with a degree of opening of 0.9, of rounding of 0.32, and it is anterior at a 0.12 degree. The /b/ phoneme, on the other hand, can be considered a plosive sonorant phoneme, bilabial and slightly velar, and therefore it can be represented by the following array:

$$[0,1, 0, 0,0,0|1, 0,0,0,0.8, 0.9,0, 0, 0,0,0,0, 0.2].$$

The arrays reported as an example have been partitioned for indicating the boolean and the fuzzy fields respectively. Such arrays, defined for each phoneme, are the membership values of the fuzzy places of articulation of the phonemes.

On the other hand, the I, D, F and L fuzzy variables, defined in a continuous universe of discourse, can take any value in their interval of definition. The fuzzy sets for these variables have been defined as follows:
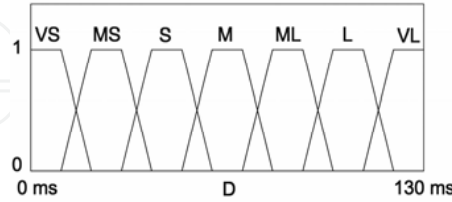

Figure 7. Fuzzy sets of the D(p) fuzzy variable

- Duration D(p). The global range of this fuzzy variable is 0-130 ms, with trapezoidal membership functions as shown in Fig. 7. In Fig. 7 such values are indicated as follows:

$$\text{Very Short, Medium Short, Short, Medium,} \\ \text{Medium Long, Long, Very Long} \tag{1}$$

- Initial Interval I(p). As D(p), this fuzzy variable is divided into trapezoidal membership functions in a 0-130 ms interval. The fuzzy values are indicated, in this case:

$$\text{Instantaneous, Immediate, Quick, Medium,} \\ \text{Medium Delayed, Delayed, Very Much Delayed} \tag{2}$$

- Final Interval F(p). The numeric range is $0-130$ ms and the fuzzy values are the same as indicated for the Initial Interval I(p).
- Locus L(p). The fuzzy values of this variable depend on the actual parameter to be controlled. For AV, AH and AF the fuzzy values are:

$$\text{Zero, Very Low, Low, Medium Low, Medium,} \\ \text{Medium High, High, Very High} \tag{3}$$

and their membership functions are equally distributed between 12 and 80 dB with the trapezoidal shape shown in Fig. 7. The other gain factors, namely A2F-A6F and AB, take one of the following values:

$$\text{Very Low, Low, Medium Low, Medium,} \\ \text{Medium High, High, Very High} \tag{4}$$

in the range 0-80 dB with the same trapezoidal shape as before. The values of L(F1), L(F2) and L(F3) are named as in (4), with trapezoidal membership functions uniformly distributed from 180 to 1300 Hz, 550 to 3000 Hz and 1200 to 4800 Hz for the first, second and third formant respectively. For example, the fuzzy sets of L(F1) are shown in Fig. 8.
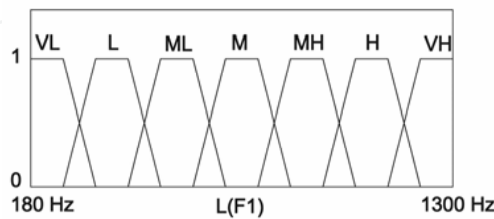

Figure 8. Fuzzy sets of the F1 locus

Finally, the loci of the bandwidths Bl, B2 and B3 take one of the fuzzy values described in (4), and their trapezoidal membership functions are regularly distributed in the intervals 30-1000 Hz for Bl, 40-1000 Hz for B2 and 60-1000 Hz for B3.

3.2 Fuzzy Rules and Defuzzification

By using linguistic expressions which combine the above linguistic variables with fuzzy operators, it is possible to formalize the relationship between articulatory and acoustic features.

We report as an exemplification, the semplified fuzzy rule of the transitions Vowel-Fricative.

```
IF  PO  IS  ANY AND  PI  IS  SONORANT THEN  {
                                        B(AV)   IS  MEDIUM-HIGH
                                              }
IF  PO  IS  ANY AND  PI  IS  ^SONORANT THEN {
                                        B(AV)   IS  ZERO
                                              }
IF PO IS ^LAB AND PI IS ANY THEN {    D(F1) IS MEDIUM-LONG ;
                                      D(F2) IS MEDIUM-LONG ;
                                      D(F3) IS MEDIUM-LONG
                                              }
IF PO IS LAB*SON AND PI IS ANY THEN  { I(AV) IS IMMEDIATE ;
                                       D(F1) IS MEDIUM ;
                                       D(F2) IS MEDIUM ;
                                       D(F3) IS MEDIUM
                                              }
IF PO IS LAB*^SON AND PI IS ANY THEN { D(F1) IS MEDIUM-SHORT ;
                                       D(F2) IS MEDIUM-SHORT ;
                                       D(F3) IS MEDIUM-SHORT
                                              }
IF PO IS ANY AND PI IS OPEN THEN { B(F1) IS MEDIUM
                                              }
IF PO IS ANY AND PI IS ^OPEN THEN { B(F1) IS VERY-SHORT
                                              }
IF PO IS ANY AND PI IS ANTERIOR THEN { B(F2) IS MEDIUM-HIGH
                                              }
IF PO IS ANY AND PI IS ^ANTERIOR THEN { B(F2) IS LOW
                                              }
IF PO IS ANY AND PI IS ROUND THEN { B(F3) IS LOW
                                              }
IF PO IS ANY AND PI IS ^ROUND THEN { B(F3) IS MEDIUM
                                              }
```

Since the manner of articulation well partitions the phonemes in separated regions, the rules have been organized in banks, one for each manner.

That is, calling PO and PI the actual and the future phonemes respectively, the set of rules is summarized in Fig. 9. The rule decoding process is completed by the defuzzification operation, which is performed with the fuzzy centroid approach.

Concluding, as shown in Fig. 9, there are several transitions which are performed with the same set of rules. For example, all the transition toward fricatives and liquid phonemes are realized with the same bank of rules. This is because the related transitions can be approximated with a strong discontinuity, and thus they can be considered independent from the starting phonemes; the symbol 'CO' used in these banks stands, in fact, for a generic consonant sounds. Other banks are missing; this is because they are concerned with transitions which occur very rarely in Italian language.

3.3 Genetic optimization of articulatory and acoustic parameters

Let us take a look at Fig. 1. Genetic optimization estimates the optimum values of the degrees of membership for the articulatory features used to generate an artificial replica of the input signal by comparing the artificial with the real signal. The optimal membership degrees of the articulatory places minimize the distance from the uttered signal; the inputs are the number of phonemes of the signal and their classification in terms of manner of articulation.

One of the most important issues of the genetic algorithm is chromosome coding. The chromosome used for the genetic optimization of a sequence of three phonemes is shown in Fig. 10.

Figure 9. Outline of the bank of fuzzy rules. P0 and P1 represent the actual and target phonetic categories. CO denotes a generic consonant

It represents the binary coding of the degrees of membership. Typical values of mutation and crossover probability are around 0.1 and 0.7 respectively.

An important aspect of this algorithm is the fitness computation, which is represented by the big circle symbol in Fig. 1. For the sake of clarity of the Section, we now briefly summarize the mel-cepstrum distance measure.
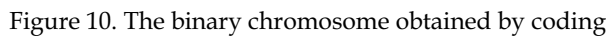
3.3.1 Mel-Cepstrum distance measure

Our distance measure uses the well known band-pass Mel-scale distributed filter bank approach, where the output power of each filter is considered. We can interpret the output of a single band-pass filter as the k-th component of the DFT of the input sequence x(n):

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}nk} \tag{5}$$

Since we are interested in the center frequencies of the band-pass filters, the k-th frequency is moved, with respect to eq. (5), by ir/N. Therefore:

$$\hat{X}(k) = \sum_{n=0}^{N-1} x(n)e^{-j\left(\frac{2\pi}{N}k + \frac{\pi}{N}\right)n} = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}n(k-0.5)} \tag{6}$$

Since $x(n)$ is real, the sequence $X(n)$ is symmetric; hence only the first $N/2$ points are sufficient. The inverse transform of the last relation is thus:

$$x(k) = \frac{2}{N} \sum_{n=0}^{N/2-1} \hat{X}(k) \cos\left(\frac{2\pi}{N}(k-0.5)n\right) \tag{7}$$

Figure 10. The binary chromosome obtained by coding

In conclusion, setting $N = N/2$, the Mel-cepstrum coefficients can be computed with the following relation:

$$c_i = \sum_{n=0}^{\overline{N}-1} X(k) \cos\left(\frac{2\pi}{N}(k-0.5)i\right) \quad i = 0...M-1 \tag{8}$$

where $X(k)$ is the logarithm of the output energy for the k-th triangular filter, N is the number of filters and M is the number of Mel-cepstrum coefficients. As reported above, the computation of the cepstral coefficients is performed using a band-pass filter bank distributed on a Mel frequency scale. The Mel scale is a non linear scale, motivated by perceptual studies of the frequency response characteristics of the human auditory system. Within the audible frequency range, a pure tone or the fundamental frequency of a complex sound produces a psychological effect on the listener. The psychological effect of a given frequency is measured in [Mel] according to the following definition: A 40 dB tone at l000Hz produces a 1000 Mel effect. Different frequencies produce different effects. It was experimentally evaluated that a 2000 Mel response is obtained for a 3100 Hz tone, while a 500 Mel response is obtained at 400 Hz. The relation between frequency in Hz and Mel it was found to be well described by a linear scale in Hz below 1000 Hz and a logarithmic scale above 1000 Hz: $Mel = 1000 \, log_2\left(1 + \frac{Hz}{1000}\right)$.

The mel-cepstral distance measure is then defined by:

$$D = \sum_{i=0}^{M-1} (c_i^a - c_i^o)^2 \tag{9}$$

where c_i^a and c_i^o are the i-th mel cepstral coefficients of the artificial and original phrase.

3.3.2 Fitness computation and articulatory constraints

The fitness, which is the distance measure between original and artificial utterances and is optimized by the genetic algorithm, is an objective measure that reflects the subjective quality of the artificially generated signal. The mel-cepstrum measure is used to compare the artificial signal generated by the fuzzy module and the speech generation module against the original input signal. The original and the artificial utterances are first aligned and then divided into frames and the average squared Euclidean distance between spectral vectors obtained via critical band filters is computed. The alignment between the original and artificial utterances is performed by using dynamic programming [54], with slope weighting as described in [55] and shown in Fig. 11.

Therefore, using the mapping curve between the two signals obtained with dynamic programming, the mel-cepstral distance D between original and artificial utterances represented respectively with X and Y is computed as follows:

$$D(X,Y) = \frac{1}{L_\Phi} \sum_{k=0}^{T} \left[\sum_{j=0}^{K} [c_x(\Phi_x(k),j) - c_y(\Phi_y(k),j)]^2 \, m(k) \right]$$

Figure 11. Slope weighting performed by dynamic programming

where T is the number of frames, K is the number of cepstrum coefficients, $\Phi = (\Phi_x, \Phi_y)$ is the non-linear mapping obtained with dynamic programming, L_Φ is the length of the map, $c_x(i,j)$ is the j-th Mel-cepstrum of the i-th frame of the original utterance, $c_y(i,j)$ is the j-th Mel-cepstrum of the i-th frame of the artificial utterance, and $m(k)$ are the weights as shown in Fig. 11.

The fitness function of the Place of Articulation (PA), i.e. the measure to be maximized by the genetic algorithm, is then computed as:

$$Fitness\,(PA) = \frac{1}{D(X,Y)}$$

Therefore, the goal of the genetic optimization is to find the membership values that lead to a maximization of the fitness, i.e. the minimization of the distance $D(X, Y)$, namely $PA = argmax \{Fitness\,(PA)\}$, $PA = \bigcup PA_i$, $i = 1, \ldots, N$, where PA_i is the degree of membership of the i-th place of articulation, N is the number of phonemes of the input signal. However, in order to correctly solve the inverse articulatory problem, the following constraints, due to the physiology of the articulations, have been added to the fitness:

- it is avoided that a plosive phoneme is completely dental and velar simultaneously;
- it is avoided that a nasal phoneme is completely voiced;

- it is avoided that all the membership degrees are simultaneously less than a given threshold;
- it is avoided that two or more degrees of membership are simultaneously greater than another threshold.

The fitness is therefore given by:

$$Fitness\,(PA) = \frac{1}{D(X,Y) + \sum_{j=1}^{N_c} P_j}$$

where P_j is the j-th penalty function and N_c is the number of constraints.

In conclusion, the optimization of places of articulation (PA) can be expressed as follows:

$$PA = argmax\,\{Fitness\,(PA)\}$$

4. Experimental results

Our imitation learning algorithm has been tested considering several different aspects. Initially, we have considered the convergence issue. In Fig. 12 a typical convergence behavior is represented, where $1/Fitness\ (PA)$ against number of generation is shown. Basing on the results shown in Fig. 12, the experimental results presented in the following are obtained with a population size of 500 elements, mutation rate equal to 0.1 and crossover probability of 0.75.
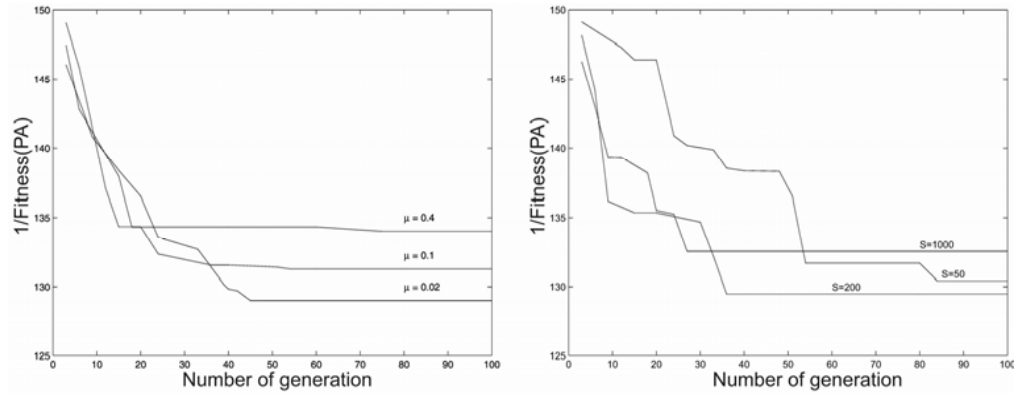
Figure 12. Convergence diagram, i.e. $1/Fitness\ (PA)$ versus number of generation for six different learning experiments. In the left panel the mutation rate μ is varied and the population size is maintained constant to 200 elements. In the right panel the population size S is varied and the mutation rate is maintained constant to 0.02

As outlined in Section 3, the speech learning mechanism works as follows: the operator pronounces one word and the talking robot generates an artificial replica of the word based on the articulatory and acoustic estimation. This process iterates until the artificial word matches the original one according to the operator judgement. More precisely, the robot learns how to pronounce a word in terms of articulatory movements using several utterances from the same talker. In Fig. 13 is shown a typical fitness error behavior considering different utterances of the same word and both non-iterative and iterative

learning: in non-iterative learning (left panel), the optimization process starts randomly in each optimization process; in iterative learning (right panel) the optimum obtained using the previous utterance of the same word is used as a starting point for the new optimization process using the new utterance of the word, obtaining usually a better fitness error. Iterative learning has another important feature: it allows us to obtain a mean behavior of the articulatory parameters for a given word and a given talker. The operator must repeat this iterative learning process for a number of words. After these phases, the speech learning process is completed.

Coming back to Fig. 13, it is worth noting that the three descending curves are related to three subsequent different pronunciations of the same word.
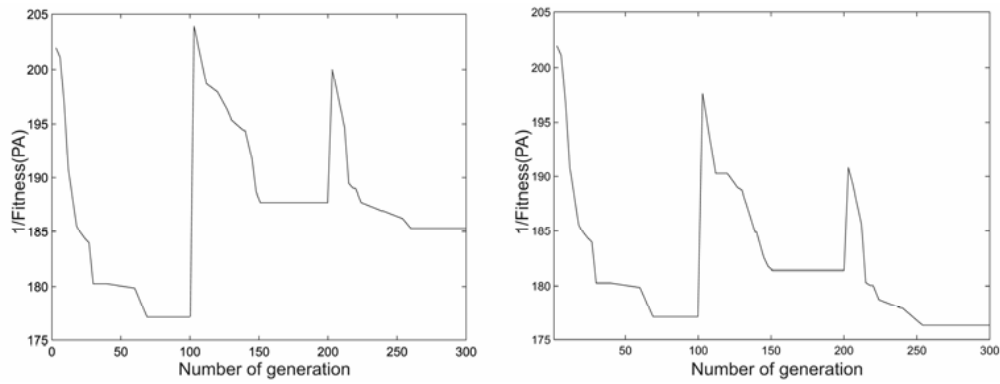
Figure 13. Convergence diagram for non-iterative (left panel) and iterative (right panel) learning algorithm. Using non-iterative learning, each learning process starts from random initial conditions. Using iterative learning, the new learning process starts from the optimal parameters obtained from the learning of another utterance of the same word. Each curve of this figure is related to the Italian word 'nove' pronounced three times; averaged values are depicted

In Fig. 14 and in Fig. 15 are reported some experimental results related to the analysis of the Italian word 'gentile' ('kind'). In the upper panel of Fig. 14 the dynamic behavior of the first three formant frequencies is reported. The vertical lines denote the temporal instants of the stationary part of each phoneme. It is worth noting that this segmentation is done on the synthetic signal but it can be related to the original signal using the non−linear mapping between the original and synthetic word obtained by dynamic programming. In the lower panel of Fig. 14 the behavior of low and high frequencies amplitudes are shown.

The trajectories of the places of articulation, estimated with the algorithm, and reported as an example in Fig. 15, can be used to shape the various organs of a mechanical vocal tract, and consequently, by means of a mechanical linkage, of the mechanical talking head.

Since the facial motion can be determined from vocal tract motion by means of simple linear estimators as shown by Yehia et al. in [13], we used the same parameters to control a graphical talking head. Yehia et al. in [13] built an estimator to map vocal-tract positions to facial positions. Given a vector y of vocal-tract marker positions to a vector x of facial positions, an affine transformation is defined by

$$\tilde{x} - \mu_x = T_{yx}(X - \mu_y) \tag{10}$$

with $\mu_x = E[x]$ and $\mu_y = E[y]$ the expected values of x and y. Arranging all the frames of vocal-tract and facial data training sets in singles matrices Y and X, where M_{tr} is the number of vectors contained in the training set, an estimation of T_{yx} is given by

$$T_{yx} \approx X_0 Y_0^T (Y_0 Y_0^T)^{-1}$$

(11)

where Y_0 and X_0 are given by Y and X subtracting the corresponding expected value from each row respectively. Using this linear estimation and the articulatory parameters of our algorithm, we animated a graphical talking head.

To generate an artificial face, a synthetic 3D model was used and visualized in OpenGL under controlled illumination as shown in Fig. 16. The Geoface 3-D facial mesh [56] was used for the experiment and tesselated to a high degree for utmost smoothness in the surface normal computation. Some muscles around the mouth have been added in order to obtain the correct facial movements during articulation of vowels and consonants.
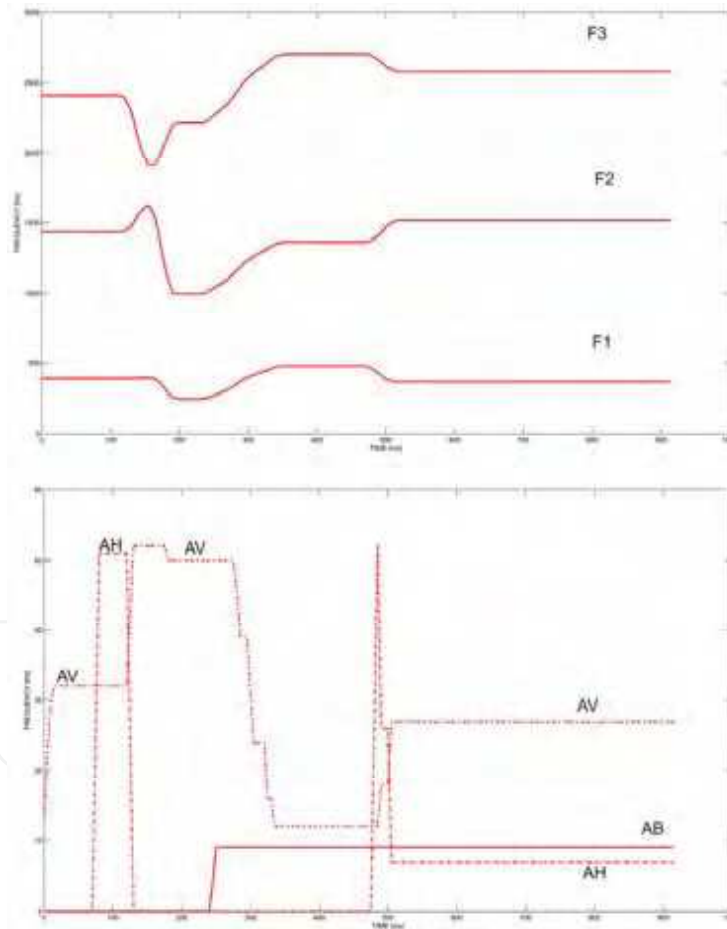
Figure 14. Acoustic analysis of the Italian word 'gentile' obtained with the genetic-fuzzy imitation learning algorithm
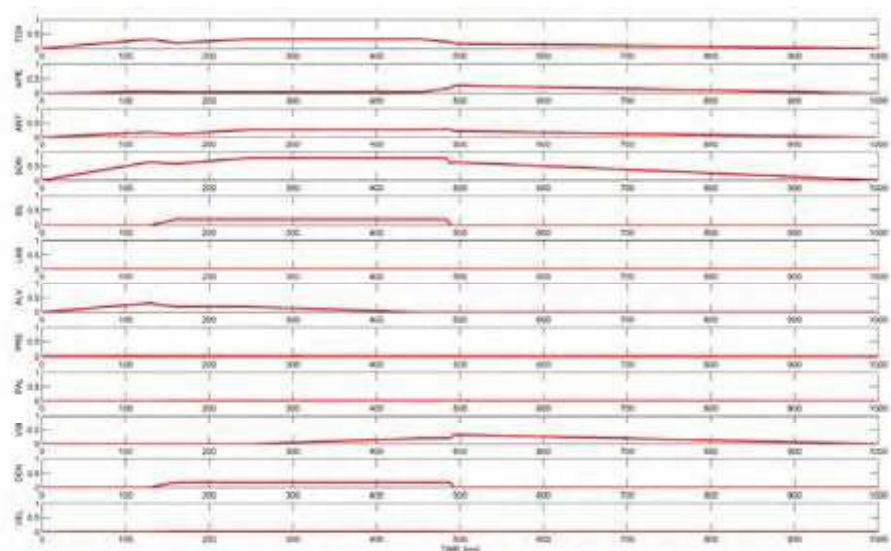
Figure 15. Places of articulation of the Italian word 'gentile' estimated with the genetic-fuzzy imitation learning algorithm

Figure 16. In the left panel is shown the mesh model of the talking head. In the right panel, a skin has been added to the model

Figure 17. Some frames of the Italian utterance "tanto gentile e tanto onesta pare" as pronounced by the talking head

The linear model between the articulatory movements and the facial movements, as described by eq. (11), has been estimated on the basis of the MOCHA-TIMIT data base. MOCHA-TIMIT [57] is a database of articulatory features that considers a set of 460 sentences designed to include the main connected speech processes in English, pronounced by two speakers, a male and a female. This database includes articulatory and acoustic data recorded in studies on speech production. Some instruments, namely EMA and EPG, have been used for the production of MOCHA-TIMIT. EMA (electromagnetic articulography) is a technique which allows articulatory movements to be monitored by means of small electromagnetic coils attached to vocal-tract structures in the mid-sagittal plane. Possible measurement points are situated on the tongue, the upper and lower lips, the mandible, and the velum. In addition, coils are generally attached to fixed structures such as the bridge of the nose and the upper central incisors to provide a maxillary frame of reference. Alternating magnetic fields generated by transmitter coils make it possible to measure the position of each receiver coil relative to two orthogonal axes in the midsagittal plane, with a measurement bandwidth typically ranging from DC up to about 250 Hz [58]. EPG (electropalatography) is a technique for recording the timing and location of tongue-palate contact patterns during speech [59]. It involves the subject wearing an artificial plate moulded to fit the upper palate and containing a number of electrodes mounted on the surface to detect lingual contacts.

A set of common phonemes between English and Italian language pronounced by the male speaker has been extracted from MOCHA-TIMIT database, and the movements of the lips related to the articulatory features has been recorded in the form of eq. (11).

In Fig. 17 are shown three picture of the talking head pronouncing the Italian utterance: "Tanto gentile e tanto onesta pare" . Informal audio-visual tests show that the algorithm is able to produce correct results.

5. Discussion and Final remarks

The estimation of articulatory static and dynamic configuration is one of the most difficult problems in voice technology. Several approaches have been attemped during the past years, and most of the difficulties are due to the fact that the articulatory-acoustic relation is not unique: different articulatory configurations can produce the same signal. The problem can be faced with suitable constraints which are needed to avoid the unrealistic configurations. One on the last work in this area has been reported in [60], which estimated articulatory parameters by finding the maximum a posteriori estimate of articulatory parameters for a given speech spectrum and the state sequence of a HMM—based speech production model.

This model consists of HMMs of articulatory parameters for each phoneme and an articulatory-to-acoustic mapping that transforms the articulatory parameters into the speech spectrum for each HMM state. The authors constructed the model by using simultaneously observed articulatory and acoustic data for sentence utterances, which was collected by using an electro-magnetic articulo-graphic (EMA) system.

Our system does not use EMA at all, trying to get the right articulations with several penalty factors. For this reason, the audio-video result is not always correct and it needs a judgment by the operator.

We emphasize the following final remarks.

- Our algorithm generates allophonic variations of the phonemes. In other words, since the described procedure performs a sort of interpolation among the acoustical parameters, the actual phonemic realization depends on the phonetic context.
- Our fuzzy model can be easily modified and tuned because the fuzzy rules are expressed in a linguistic manner.
- Many further optimizations are possible, in terms of genetic algorithm and in terms of the fuzzy rules which could be automatically estimated instead of being defined from phonetic knowledge on its own, as we did.
- Our algorithm could be used as a design tool of mechanical talking heads because the artic ulatory parameters can be easily added or deleted from the fuzzy rules and their effects of the modification can be immediately verified.
- Finally, is has to be noted that in this paper we implemented only the rules pertaining to the Italian language. This does not limit the generality of our method: if a different language has to be considered, new banks of fuzzy rules could be added and the previous banks could be modified.

6. Conclusions

In this Chapter we have dealt with the articulatory control of talking heads, which can be either graphical or mechanical. A novel approach for the estimation of articulatory features from an input speech signal is described. The approach uses a set of fuzzy rules and a genetic algorithm for the optimization of the degrees of membership of the places of articulation. The membership values of the places of articulation of the spoken phonemes have been computed by means of genetic optimization. Many sentences have been generated on the basis of this articulatory estimation and their subjective evaluations show that the quality of the artificially generated speech is quite good. As compared with other works in acoustic to articulatory mapping, which generally compute the vocal tract area functions from actual speech measurements, our work present a method to estimate the place of articulation of input speech through the development of a novel computational model of human vocalization.

7. References

E. Mumolo, M. Nolich, and E. Menegatti. A genetic-fuzzy algorithm for the articulatory imitation of facial movements during vocalization of a humanoid robot. In *Proceeding of the 2005 IEEE Int. Conf. on Humanoid Robotics,* pages 436-441, 2005. [1]

R. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. Wiliamson. The Cog Project: Building a Humanoid Robot. *Lecture Notes in Artificial Intelligence, Springer-Verlag,* 1998. [2]

K. Nishikawa, K. Asama, K. Hayashi, H. Takanobu, and A. Takanishi. Development of a Talking Robot. In *Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems,* pages 1760−1765, 2000. [3]

K. Nishikawa, K. Asama, K. Hayashi, H. Takanobu, and A. Takanishi. Mechanical Design of a Talking Robot for Natural Vowels and Consonant Sounds. In *International Conference on Robotics and Automation,* pages 2424-2430, May 2001. [4]

K. Nishikawa, A. Imai, T. Ogawara, H. Takanobut, T. Mochidal, and A. Talcanishi. Speech Planning of an Anthropomorphic Talking Robot for Consonant Sounds Production. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation,* pages 363-368, 2002. [5]

K. Nishikawa, H. Takanobu, T. Mochida, M. Honda, and A. Takanishi. Speech Production of an Advanced Talking Robot based on Human Acoustic Theory. In *Proceadlngr dthe 2004 IEEE International Conference on Robotics and Automation,* pages 363-368, 2004. [6]

K. Nishikawa, T. Kuwae, H. Takanobu, T. Mochida, M. Honda, and A. Takanishi. Mimicry of Human Speech Sounds using an Anthropomorphic Talking Robot by Auditory Feedback. In *Proceeding of the 2004 IEEE International Conference on Robotics and Automation,* pages 272-278, 2004. [7]

K. Fukui, K. Nishikawa, S. Ikeo, E. Shintaku, K. Takada, H. Takanobu, M. Honda, and A. Takanishi. Development of a Talking Robot with Vocal Cords and Lips Having Humanlike Biological Structures. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems,* pages 2023-2028, 2005. [8]

K. Fukui, K. Nishikawa, T. Kuwae, H. Takanobu, T. Mochida, M. Honda, and A. Takanishi. Development of a New Human-like Talking Robot for Human Vocal Mimicry. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation,* pages 1437-1442, 2005. [9]

K. Fukui, K. Nishikawa, S. Ikeo, E. Shintaku, K. Takada, A. Takanishi, and M. Honda. New Anthropomorphic Talking Robot having Sensory Feedback Mechanism and Vocal Cords based on Human Biomechanical Structure. In *The First IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics,* BioRob 2006, pages 1095—1100, 2006. [10]

K. Fukui, K. Nishikawa, S. Ikeo, M. Honda, and A. Takanishi. Development of a Human-like Sensory Feedback Mechanism for an Anthropomorphic Talking Robot. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation,* pages 101—106, 2006. [11]

C. Breazeal. *Designing Sociable Robots.* MIT Press, 2004. [12]

H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication,* V.26:23-43, 1998. [13]

F. I. Parke. Computer generated animation effaces. Master's thesis, University of Utah, 1972. [14]

F. I. Parke. A Parame.te.ric Model for Human Faces, *PhD thesis,* University of Utah, 1974. [15]

M. Lundeberg and J. Beskow. Developing a SDagent for the august dialogue system. In *Proceedings of the Auditory Visual Speech Processing '99 Conference (AVSP'99),* 1999. [16]

Grossman B. Cechner P. Beard, S. and A. Marriott. Faqbot. In *Sydney Area Workshop on Visual Information Processing.,* 1999. [17]

Cohen M. A. Massaro, D. W. and M. A. Berger. Creating talking faces: Applying talking faces. In *In Perceiving Talking Faces: From Speech Perception to a Behavioral Principle,* 1998. [18]

L. Williams. Performance-driven facial animation. In *Computer Graphics,* 1990. [19]

D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1993. [20]

I. A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proceedings of IEEE Computer Vision Pattern Recognition Conference,* 1994. [21]

I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. In *International Conference on Computer Vision,* 1995. [22]

Cole R. et al. New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. In *Proceedings of ESCA/SOCRATES Workshop on Methods and Tool Innovations for Speech Science Education.* London, UK., 1999. [23]

J.J. Williams and A.K. Katsaggelos. An HMM-Based Speech-to-Video Synthesizer. *IEEE Trans, on Neural Networks,* V.13(N.4):900-915, 2002. [24]

C. Breazeal and L. Aryananda. Recognition of affective communicative intent in robot-directed speech. In *Autonomous Robots,* 2002. [25]

A. Fernald. Four-month-old infants prefer to listen to motherese. In *Infant Behavior and Development,* 1985. [26]

P. Varshavskaya. Behavior-based early language development on a humanoid robot. In *2nd Int. Conf. on Epigenetics Robotics,* 2002. [27]

K. Dautenhahn and A. Billard. Studying robot social cognition within a developmental psychology framework. In *3rd Int. Workshop on Advanced Mobile Robots,* 1999. [28]

C. Breazeal and B. Scassellati. Infant-like social interactions between a robot and a human caretaker. In *Adaptive Behavior,* 2000. [29]

K. Dautenhahn. I could be you: The phenomenological dimension of social understanding. In *Cybernetics and Systems Journal,* pages 417—453, 1997. [30]

B. Scassellati. Theory of mind for a humanoid robot. In *Proceedings of the 2000 IEEE/RSJ International Conference on Humanoid Robotics,* 2000. [31]

J. Zlatev. The epigenesis of meaning in human beings, and possibly in robots. In *Lund University Cognitive Studies,* 1999. [32]

H. Kozima and J. Zlatev. An epigenetic approach to human-robot communication. In *International Workshop on Robot and Human Interactive Communication ROMAN-2000,* 2000. [33]

Ulrich Neumann Jun-yong Noh. Talking Faces. In *IEEE International Conference on Multi-media and Expo (II),* 2000. [34]

K. Waters. A muscle model for animating threedimensional facial expression. In *Computer Graphics,* 1987. [35]

J. Frisbie K. Waters. A Coordinated Muscle Model for Speech Animation. In *Graphics Interface,* 1995. [36]

K. Waters Y. C. Lee, D. Terzopoulos. Realistic face modeling for animation. In *Siggraph proceedings,* 1995. [37]

N. Badler S. Platt. Animating facial expression. In *Computer Graphics,* 1981. [38]

N. M. Thalmann D. Thalmann P. Kalra, A. Mangili. Simulation of Facial Muscle Actions Based on Rational Free From Deformations. In *Eurographics,* 1992. [39]

T. Kim R. Enciso U. Neumann D. Fidaleo, J-Y. Noh. Classification and Volume Morphing for Performance-Driven Facial Animation. In *Digital and Computational Video,* 1999. [40]

D. Lischinski R. Szeliski D. H.Salesin F. Pighin, J. Hecker. Synthesizing Realistic Facial Expressions from Photographs. In *Siggraph proceedings,* 1998. [41]

T. Poggio T. Ezzat. Mike Talk: A Talking Facial Display Based on Morphing Visemes. In *Computer Animation 1998,* 1998. [42]

Y. Ohta L. Gao, Y. Mukaigawa. Synthesis of Facial Images with Lip Motion from Several Real Views. In *Automatic Face and Gesture Recognition,* 1998. [43]

B.F. Buxton I. Koufakis. Very low bit rate face video compression using linear combination of 2D face view and principal components analysis. In *Image and Vision Computing,* 1999. [44]

H. P. Graf E. Cosatto. Sample-Based Synthesis of Photo Realistic Talking Heads. In *Computer Animation 1998,* 1998. [45]

M. Ouhyoung W. Perng, Y. Wu. Image Talk: A Real Time Synthetic Talking Head Using One Single Image with Chinese Text-To-Speech Capability. In *IEEE,* 1998. [46]

D. Reisfeld Y Yeshurun N. Arad, N. Dyn. Image Warping by Radial Basis Functions: Application to Facial Expressions. In *Graphical Models and Image Processing,* March 1994. [47]

T. Higashimoto and H. Sawada. Speech Production by a Mechanical Model Construction of a Vocal Tract and its Control by Neural Network. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation* (ICRA 2002), pages 3858-3863, 2002. [48]

H. Sawada, M. Nakamura, and T. Higashimoto. Mechanical Voice System and Its Singing Performance. In *Proceedings of 2004 1EEE/RSJ International Conference on Intelligent Robots and Systems,* October 2004. [49]

H. Sawada, M. Nakamura, and T. Higashimoto. Mechanical voice system and its singing performance. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems,* 2004 (IROS 2004), pages 1920-1925, October 2004. [50]

M. Nakamura and H. Sawada. Talking Robot and the Analysis of Autonomous Voice Acquisition. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems,* pages 4684-4689, October 2006. [51]

P.C. Delattre, A.M. Liberman, and F.S. Cooper. Acoustic Loci and Transitional Cues for Consonants. *Journal of the Acoustical Society of America,* V.27(N.4), 1955. [52]

D. H. Klatt. Review of Text-to-Speech Conversion in English. *J. Acous. Soc. Am.,* pages 737-793, 1987. [53]

H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. Acoust. Speech Signal Processing,* V.26:43-49, 1978. [54]

L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice—Hall, 1993.[56]

F. Parke and K. Waters. *Computer Facial Animation.* A.K. Peters, 1996. [55]

A. Wrench. MOCHA-TIMIT http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html, November 1999. [57]

J.S. Perkell, M.H. Cohen, M.A. Svirsky, M.L. Matthies, I. Garabieta, and M.T.T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America,* V.92:3078-3096, 1992. [58]

W.J. Hardcastle, W. Jones, C. Knight, A. Trudgeon, and G. Calder. New developments in electropalatography: a state—of—the—art report. *Clinical Linguistic and Phonetics,* V.3:1—38, 1989. [59]

S. Hiroya and M. Honda. Estimation of Articulatory Movements From Speech Acoustics Using an HMM-Based Speech Production Model. *IEEE Transactions on Speech and Audio Processing,* Vol. 12(N. 2):175, March 2004. [60]

Humanoid Robots, Human-like Machines

Edited by Matthias Hackel

In this book the variety of humanoid robotic research can be obtained. This book is divided in four parts: Hardware Development: Components and Systems, Biped Motion: Walking, Running and Self-orientation, Sensing the Environment: Acquisition, Data Processing and Control and Mind Organisation: Learning and Interaction. The first part of the book deals with remarkable hardware developments, whereby complete humanoid robotic systems are as well described as partial solutions. In the second part diverse results around the biped motion of humanoid robots are presented. The autonomous, efficient and adaptive two-legged walking is one of the main challenge in humanoid robotics. The two-legged walking will enable humanoid robots to enter our environment without rearrangement. Developments in the field of visual sensors, data acquisition, processing and control are to be observed in third part of the book. In the fourth part some "mind building" and communication technologies are presented.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

INTECH
open science | open minds