Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots

# A genetic-fuzzy algorithm for the articulatory imitation of facial movements during vocalization of a humanoid robot

Enzo Mumolo and Massimiliano Nolich
*DEEI, University of Trieste, Trieste, Italy*
*{mumolo,mnolich}@units.it*

Emanuele Menegatti
*DEI, University of Padova, Padova, Italy*
*emg@dei.unipd.it*

***Abstract*— In human heads there is a strong structural linkage between vocal tract and facial behavior during speech. For a robotic talking head to have a human-like behavior, this linkage should be emulated. One way to do that is to compute an estimate of the articulatory features which produce a given utterance and then to transform them into facial animation. We present a computational model of human vocalization which is aimed at describing the articulatory mechanisms which produce spoken phonemes. It uses a set of fuzzy rules and genetic optimization. The former represents the relationships between places of articulations and speech acoustic parameters, while the latter estimates the degrees of membership of the places of articulation. That is, the places of articulation are considered as fuzzy sets whose degrees of membership are the articulatory features. The trajectories of articulatory parameters can be used to control a graphical or mechanical talking head. We verify the model presented here by generating and listening to artificial sentences. Subjective listening tests of artificially generated sentences from the articulatory description resulted in an average phonetic accuracy of about 79 %. Through the analysis of a large amount of natural speech, the algorithm can be used to learn the places of articulation of all phonemes of a given speaker.**

***Index Terms*— Speech imitation, talking head, imitation learning, articulatory model, fuzzy-genetic optimization.**

## I. Introduction

Nowadays, there is an increasing interest in humanoid robotics. A humanoid autonomous 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 devices generally furnish environment information useful for motion tasks, auto-localization and obstacle avoidance in order to introduce reactiveness and autonomy. The goal of humanoid robotics is to create 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 would be designed for that purpose. 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) capabilities of human beings. Human-robot interaction and dialogue modalities have been widely studied in recent years in robotics and AI communities. Significant contributions have been made within the field of advanced robotics and humanoids like the Cog project at MIT [1], and the related Krismet project. In these works the idea of 'conveying intentionality' is connected to human children's social interactions with caregivers.

A very important area in humanoid robotics is the interaction with the human operators. As speech is the most natural communication means for humans, conversational interfaces with humanoids are very promising to facilitate and improve the way people interact with a robot. Basically, conversational interfaces are built around two technologies: speech synthesis from unrestricted text and speech recognition. Besides the well-known problems of accuracy and speaker dependance, a major problem in speech recognition for interacting with a robot using a conversational interface is that the human operator is distant from the microphone, which is fitted to the robot; a way to overcome this problem is to implement beamforming algorithms with microphone arrays [2].

Generally speaking, there are three main problems in robotic language production. 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 first point requires that the robot is aware of its situational context [3]. The second point means that graphemic to phonemic transformation is made while the latter point is related to actual synthesis of the artificial speech [4]. In humanoid robotics, however, it is very important, if not necessary, to produce speech using a mechanical talking head for the reason we describe shortly.

The motivation of mechanical talking robot development, as described in [5], can be to conduct research into the speech motor control system in the brain and to create a dynamic mechanical model to reproduce human speech. Besides the applications in telecommunications, medical training devices and learning devices mentioned in [5] we think that a mechanical talking head derived from articulatory features would eventually lead to a mechanical robotic face with natural behavior. 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. in [6]. This fact has been used in [7] to develop natural animation of a talking head. Moreover, as far as the mechanical talking robot is concerned, if a mechanical vocal tract is embedded into an artificial head which emulates a human head, the artificial head should have natural movements during spoken language

0-7803-9320-1/05/$20.00 ©2005 IEEE

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 control system in the brain, and thus, speech production is still not clearly understood [8]. This type of knowledge pertains to articulatory synthesis, which includes the methods of generating speech from a given movement of the vocal tract (articulatory trajectory).

In this work, we developed fuzzy rules which relate places of articulation with the corresponding acoustic parameters. The degrees of membership of the places of articulation are adapted to those of the actual speaker who trained the system.

Human infants learn to speak through interaction with their care-givers. The aim of our study is to build a robot that acquires a vocalization capability in a way similar to human development. In this paper, we considered a human-robot interaction scenario where the robot has a humanoid talking head. This does not necessarily mean a robotic head; in fact, we considered a software talking head, which is an image plotted in graphical form on the computer screen. Besides the obvious differences between these two alternatives, the common part is that the same articulatory model should be developed in both cases. This model is responsible for the estimation of articulatory parameters from natural speech. Such parameters can be used to control the talking head that is plotted on the computer screen on one hand or the articulations of the mechanical head on the other.

An original contribution of this work is that a novel articulatory model based on imitation learning is presented. In other words, the algorithm tries to reproduce some input speech and, in this way, the articulatory characteristics of the speaker who trained the system are learned. From this point, the system can synthesize unrestricted text using the articulatory characteristics of the given speaker. The articulatory parameters can then be used to control a humanoid head or, as in our case, a talking head displayed on a monitor, which can be put on the top of the service robot.

As compared with other works in acoustic to articulatory mapping, which generally compute the vocal tract area functions from actual speech measurements, our work presents a method to estimate the place of articulation of input speech through the development of a novel computational model of human vocalization.

It is worth noting, however, that in our implementation of the model only the rules pertaining to the Italian phonemes were considered. This does not limit the generality of the method: if other phonemes are considered, new fuzzy rules must be added.

The rest of this paper is organized as follows. In Section II the mechanical robot, talking heads and their control mechanism developed so far are introduced, together with an introduction to articulatory models. In Section III the problem of adaptive learning of human vocalization is presented. In Section IV the proposed fuzzy model of speech is introduced, and the genetic optimization of articulatory parameters is discussed in Section V. In Section VI some experimental results are presented and some examples of talking head commands are described in Section VII. Finally, in Section VIII some final remarks are reported.

## II. Related work

Research in robotic talking head is important in the face to face communication between humans and humanoids robots, as reported in [9]. In this section the main studies on artificial vowel and consonant sound production using a mechanical vocal tract pursued by various research groups are briefly summarized.

At Waseda University, WT-3, WT-2 and WT-1R have been developed for the production of Japanese vowels and some consonant sounds. WT-1R [10] has articulators (the tongue, lips, teeth, nasal cavity and soft palate) and vocal organs (the lungs and vocal cords), it can reproduce human vocal movement and it has 15 degrees of freedom. The vocal movement of WT-1R for vowels is steady. However, the vocal movement for consonant sounds is transient. Therefore, because the Japanese voice generally consists of two phonemes of the first consonant sound and the last vowel, they proposed the speech planning of WT-1R by considering the phenomenon of the voice as three parts (steady consonant sound, transient consonant sound and vowel). WT-2 [11] is an improvement of WT-1R; it has lungs, vocal cords, a vocal cavity and nasal cavity and it aims at reproducing the vocal tract of an adult male. WT-3 [8] is based on human acoustic theory for the reproduction of human speech; it consists of lungs, vocal cords and articulators and could reproduce human-like articulatory motion. The oral cavity was designed based on MRI images of a human sagittal plane.

At Kagawa University, a mechanical model of a human vocal tract system based on mechatronic technology has been developed [12]. It implements a neural network based mapping between motor positions and the produced vocal sound obtained by an auditory feedback in the learning phase.

In [13] the mechanical vocal tract is excited by a mechanical vibrator that oscillates at particular frequencies and acts as artificial larynges. The vocal tract shape is obtained using a neural network.

The development of a mechanical anthropomorphic talking robot for unrestricted text is very difficult as shown in [13] [12] [1]. On the other hand, graphical talking heads graphically displayed have been developed for many years. Generally, talking heads are connected to a speech synthesizer which plays a signal synchronized with the movements.

## III. Problem formulation

The goal of this work is to learn automatically the place of articulation of spoken phonemes in such a way that the robot learns how to speak through articulation movements. The block diagram of the system for adaptive learning of human vocalization is depicted in Fig. 1.

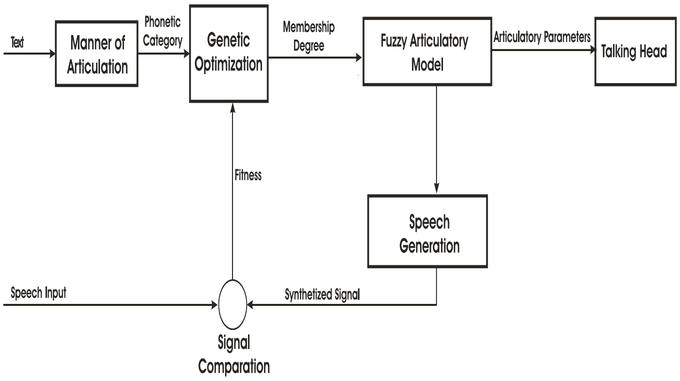


Fig. 1. Block diagram of the genetic-fuzzy optimization algorithm.

More precisely, the speech learning mechanism using our algorithm 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 judgement. At this point the robot has learnt how to pronounce those words in terms of articulatory movements. The operator must repeat this process for a large number of words. After these phases, the speech learning process is completed.

The algorithm is based on the following assumptions: the degrees of membership of a place of articulation estimated by means of the genetic optimization process are directly related to the physical configuration of the phonatory organs. For example, if a phoneme is characterized by a degree of opening equal to 0.6, it is assumed that the mouth is opened at a 60% degree of the maximum opening width. Even if no direct experimental evidence of that is given in this paper, this assumption can be indirectly verified: the overall model produces good synthetical vocalizations.

The first block on the left on Fig. 1, i.e. the 'manner of articulation' block, is fed with the text corresponding to the phrase the operator has pronounced. This block computes the phonetic categories of the sentence. For example, if the word is 'nove' in Italian language ('nine' in English), the phonetic categories are: nasal, vowel, fricative, vowel. This information is used to build the chromosome and to select the correct rules by the fuzzy articulatory module.

It has to be noted that the estimation of the broad phonetic categories could be done directly from speech [14], thus opening to the field of language recognition. However, in this paper we address the problem of talking head leaving the issues of language recognition to future works. Hence, the phonetic categories are given by the operator - who writes the sentence in a graphemic form - in the learning phase.

## IV. The fuzzy articulatory module

Usually, phonemes are classified in terms of manner and place of articulation. The manner of articulation is concerned with the degree of constriction imposed by the vocal tract on the airflow. The following six categories of the manner of articulation have been considered in this work: vowel, in which air flows throw the vocal tract without constrictions; liquid, similar to the vowels but that use the tongue as an obstruction; nasal, which is characterized by a lowering of the velum, allowing airflow out of the nostril; fricative, which employ a narrow constriction in the vocal tract which introduces turbulence in the air flow; plosive, involving a complete closure and subsequent release of a vocal obstruction; affricate, which is a plosive followed by a fricative.

The places of articulation are concerned with the location of the constriction imposed on the air flow in the vocal tract. The following places of articulation for the vowels have been used: open, anterior, voiced and rounded. A vowel opening is related to the quantity of airflow which flows between the tongue and the palate, while the anteriority is related to the position of the tongue in the vocal tract. The roundness of a vowel is related to the opening of the mouth during its generation. In the same way, the places of articulation for the consonants are related to the place where the vocal tract is narrower. So, for the bilabial consonants the lips are closed, while for labiodental consonants the constriction is produced by the tongue which is close to the superior teeth and so forth.

Using manner and place of articulation, any phoneme can be fully characterized in binary form. However, a certain degree of fuzzyness, due to the lack of knowledge, is involved in this characterization, which thus should be fuzzy rather than strictly binary. For example, it may be that the /b/ phoneme, classically described as plosive, bilabial and voiced, involve also a certain degree of anteriority and rounding, as well as some other features.

The possibility of facing the vagueness involved in the interpretation of phonetic features using methods based on fuzzy logic has been realized in the past, when approaches to speech recognition via phonetic classification were proposed [15], [16].

The following phonemes in Italian were considered in this work. Their classification in terms of the manner of articulation is as follows (using IPA symbols):

| | |
|---|---|
| vowel : | /a/, /e/, /i/, /o/, /u/, /SIL/, /$/ |
| liquid : | /l/, /$\lambda$/, /r/ |
| nasal : | /m/, /n/, /$\eta$/ |
| fricative : | /f/, /v/, /s/, /z/, /$\int$/ |
| plosive : | /p/, /b/, /t/, /d/, /k/, /g/ |
| affricate : | /d$\int$/, /d$\zeta$/, /dz/, /ts/ |

Clearly, all the quantities involved, namely phonemes and control parameters, are fuzzified, as described in the following.

### A. Phoneme and Control Parameters Fuzzification

As it was 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. In this way, the phonetic description appears as an extension of the classical binary definition described for instance by Fant in [17], and a certain vagueness in the definition of the place of articulation of the phonemes is introduced.

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. Similarly, the /i/ phoneme can be described by:

$$[1, 0, 0, 0, 0, 0|1, 0, 0.06, 0.9, 1, 0, 0, 0, 0, 0, 0, 0, 0.1]$$

The /b/ phoneme, on the other hand, can be considered a plosive sonor 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.

The output of the phonetic module, described in the following, is given in terms of these four parameters; hence the translation to the synthesis parameters trajectories required by the synthesizer must be performed.

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:

- Duration D(p);
- Initial Interval I(p);
- Final Interval F(p);
- Locus L(p).

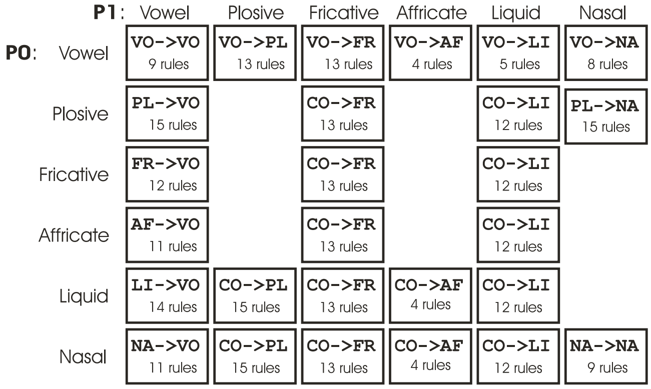
### B. Fuzzy Rules and Defuzzification

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

In general, the rules involve the actual and the future phonemes. Moreover, the fuzzy expressions involve the fuzzy operators AND, NOT and OR. Since the manner of articulation splits the phonemes into well-defined separate regions, the rules have been organized in banks, one for each manner.

That is, calling P0 and P1 the actual and the future phonemes respectively, the set of rules is summarized in Fig. 2. The rule decoding process is completed by the defuzzification operation, which is performed with the fuzzy centroid approach.

A rule example is the following:



| P0 \ P1 | Vowel | Plosive | Fricative | Affricate | Liquid | Nasal |
|---|---|---|---|---|---|---|
| Vowel | VO->VO<br>9 rules | VO->PL<br>13 rules | VO->FR<br>13 rules | VO->AF<br>4 rules | VO->LI<br>5 rules | VO->NA<br>8 rules |
| Plosive | PL->VO<br>15 rules | | CO->FR<br>13 rules | | CO->LI<br>12 rules | PL->NA<br>15 rules |
| Fricative | FR->VO<br>12 rules | | CO->FR<br>13 rules | | CO->LI<br>12 rules | |
| Affricate | AF->VO<br>11 rules | | CO->FR<br>13 rules | | CO->LI<br>12 rules | |
| Liquid | LI->VO<br>14 rules | CO->PL<br>15 rules | CO->FR<br>13 rules | CO->AF<br>4 rules | CO->LI<br>12 rules | |
| Nasal | NA->VO<br>11 rules | CO->PL<br>15 rules | CO->FR<br>13 rules | CO->AF<br>4 rules | CO->LI<br>12 rules | NA->NA<br>9 rules |

Fig. 2. Outline of the bank of fuzzy rules. P0 and P1 represent the actual and target phonetic categories. CO denotes a generic consonant.

```
IF P0 IS any and P1 IS open
  THEN { L(F1) IS medium }
IF P0 IS any and P1 IS anterior
  THEN { L(F2) IS medium_high }
IF P0 IS any and P1 IS rounded
  THEN { L(F3) IS low }
```

## V. Genetic optimization of articulatory and acoustic parameters

Let us take a look at Fig. 1. Genetic optimization aims at computing the optimum values of the degrees of membership for the articulatory features used to generate an artificial replica of the input signal.

### A. Genetic optimization module

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 main parts of the genetic algorithm is coding. The chromosome used for genetic optimization of a sequence of three phonemes is shown in Fig. 3. It represents the binary coding of the degrees of membership. The genetic algorithm

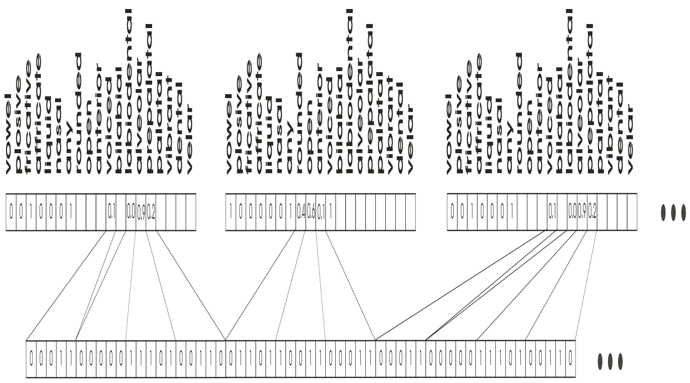


Fig. 3. The binary chromosome obtained by coding.

uses only mutations of the chromosome. This means that each bit of the chromosome is changed at random and the mutation rate is constant at 2%.

### B. Fitness computation and articulatory constraints

An important aspect of this algorithm is the fitness computation, which is represented by the big circle symbol in Fig. 1. 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 Modified Bark Spectral Distortion (MBSD) measure has been used [18], [19]. Such measure is based on the computation of the pitch loudness, which is a psycho-acoustical term defined as the magnitude of the auditory sensation. In addition to this, a noise-masking threshold estimation is considered. This measure is used to compare the artificial signal generated by the fuzzy module and the speech generation module against the original input signal.

The MBSD measure is frame-based. That is, 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 [20], with slope weighting as described in [21].

In conclusion, our optimization problem can be formalized as follows:

$$AP = argmax \left\{ \frac{1}{D(X,Y)} + \sum_{j=1}^{N_c} P_j \right\}$$

where AP are the articulatory parameters, $P_j$ is the penalty function and $N_c$ is the number of constraints.

## VI. Experimental results

The experimental results presented in the following are obtained with a population size of 200 elements and a mutation rate equal to 0.02.

In Fig. 4 and in Fig. 5 some experimental results related to the analysis of the Italian word ‘nove’ (‘nine’) are shown. In the upper panel of Fig. 4 the dynamic behavior of the first three formant frequencies, which are the most important acoustic features obtained by the algorithm, 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. 4 the behavior of low and high frequency amplitudes are shown. In Fig. 5 the dynamic of the membership degrees of the articulatory places of articulation is reported.

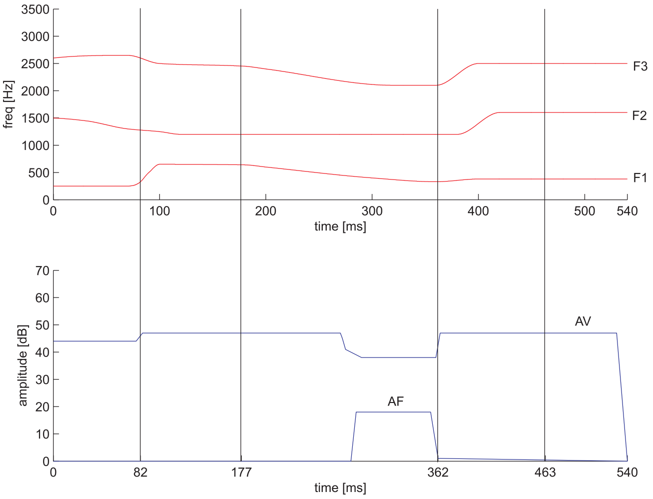


Fig. 4. Acoustic analysis of the Italian word ‘nove’ obtained with fuzzy model and genetic optimization.

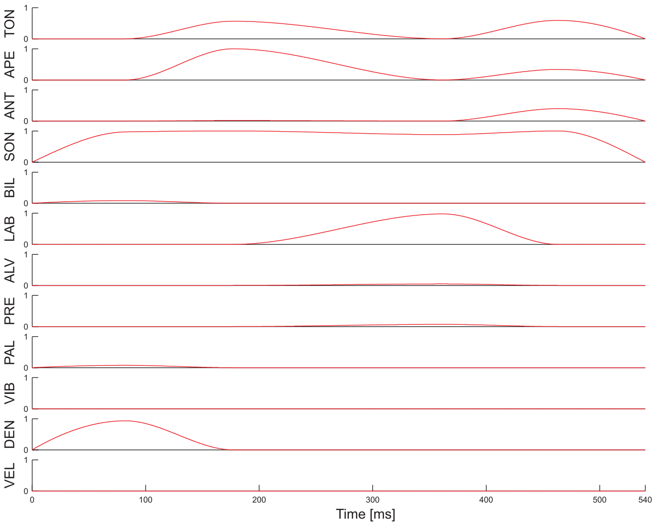


Fig. 5. Articulatory places of articulation of the Italian word ‘nove’ estimated with genetic optimization.

In Fig. 6, finally, some subjective evaluation results related to a phonetic listening test are shown: the phonetic categories used in this test are quite critical from a correct comprehension point of view. However, the subjective rate ranges from 70% to 85% and therefore it is quite promising for future developments.

## VII. Towards control of talking heads

The trajectories of the places of articulation, estimated with the algorithm, can be used to shape an animated 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 [6]. As a preliminary step towards this goal, and to verify the algorithm, we animated a midsagittal model of a human head using the articulatory output of the algorithm. In this way, the movements of the animation are automatically synchronized with the speech produced with the same algorithm.

| Phonetic Categories | Number of signals | Exact recognitions [%] |
|---|---|---|
| plosive | 193 | 76 |
| fricative | 105 | 86 |
| affricate | 47 | 76 |
| liquid | 83 | 78 |
| Total | 428 | 79 |

Fig. 6. Subjective evaluation results.

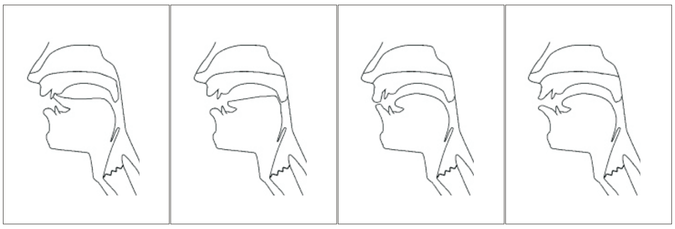


Fig. 7. Sagittal section of a talking head driven by the fuzzy-genetic articulatory model emitting the four phonemes of the Italian word 'nove'.

Informal audio-visual tests show that the algorithm is able to produce correct results. In Fig. 7and in Fig. 8 a sequence of four frames of the animation related to the Italian word 'nove' ('nine') is reported. The four frames correspond to the four phonemes of the word. The algorithm described can be eventually used to drive a mechanical vocal tract for producing natural facial configuration during speech production.

## VIII. Final remarks and conclusions

In this paper we have dealt with the articulatory control of talking heads, which can be simply graphical or even 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. One interesting property of the fuzzy model is that the fuzzy rules can be quite easily modified and tuned. The membership values of the place 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.

## References

[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] Changkyu Choi, Donggeon Kong, Jaywoo Kim, and Seokwon Bang. Speech Enhancement and Recognition Using Circular Microphone Array For Service Robotics. In *Proc. of the 2003 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 3516–3521, October 2003.

[3] Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. Mental Imagery for a Conversational Robot. *IEEE Transactions on System, Man and Cybernetics - Part B: Cybernetics*, 34(3):1374–1383, June 2004.

[4] Jonathan Allen, M. Sharon Hunnicutt, and Dennis Klatt. *From text to speech: The MITalk system*. Cambridge University Press, 1987.

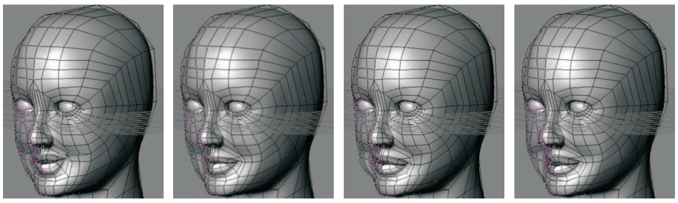


Fig. 8. Facial movements of a talking head driven by the fuzzy-genetic articulatory model emitting the four phonemes of the Italian word 'nove'.

[5] Kazufumi Nishikawa, Hideaki Takanobu, Takemi Mochida, Masaaki Honda, and Atsuo Takanishi. Modeling and Analysis of Elastic Tongue Mechanism of Talking Robot for Acoustic Simulation. In *Proceedings of the 2003 IEEE lntemational Conference on Robotics and Automation*, pages 2107–2114, 2003.

[6] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26:23–43, 1998.

[7] Eric Vatikiotis-Bateson, Christian Kroos, Kevin G. Munhall, and Michel Pitermann. Task Constraints on Robot Realism: The Case of Talking Heads. In *Proceedings of the 2000 IEEE lntemational Workshop on Robot and Human Interactive Communication*, pages 352–357, 2000.

[8] Kazufumi Nishikawa, Hideaki Takanobu, Takemi Mochida, Masaaki Honda, and Atsuo Takanishi. Speech Production of an Advanced Talking Robot based on Human Acoustic Theory. In *Proceedings of the 2004 IEEE lntemational Conference on Robotics and Automation*, pages 3213–3219, 2004.

[9] M. Shiomi, T. Kanda, N. Miralles, T. Miyashita, I. Fasel, J. Movellan, and H. Ishiguro. Face-to-face interactive humanoid robot. In *Proc. of the 2004 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1340–1346, September 2004.

[10] Kazufumi Nishikawa, Akihiro Imai, Takayuki Ogawara, Hideaki Takanobu, Takemi Mochida, and Atsuo Takanishi. Speech Planning of an Anthropomorphic Talking Robot for Consonant Sounds Production. In *Proceedings of the 2002 IEEE lntemational Conference on Robotics and Automation*, pages 1830–1835, 2002.

[11] Kazufumi Nishikawa, Hideaki Takanobu, Takemi Mochida, Masaaki Honda, and Atsuo Takanishi. Development of a New Human-like Talking Robot Having Advanced Vocal Tract Mechanisms. In *Proceedings of the 2003 IEEE/RSJ Intl. Conferece on Intelligent Robots and Systems*, pages 1907–1013, 2003.

[12] Toshio Higashimoto and Hideyuki Sawada. Speech Production by a Mechanical Model Construction of a Vocal Tract and its Control by Neural Network. In *Proceedings of the 2002 IEEE lntemational Conference on Robotics and Automation*, pages 3858–3863, 2002.

[13] Y. Yoshikawa, J. Koga, M. Asada, and K. Hosoda. Primary Vowel Imitation between Agents with Different Articulation Parameters by Parrot-like Teaching. In *Proc. IEEE–RSJ Int. Conf. on Intelligent Robots and System*, pages 149–154, 2003.

[14] J. P. Martens and L. Depuydt. Broad phonetic classification and segmentation of continuous speech by means of neural networks and dynamic programming. *Speech Communication*, 10:81–90, 1991.

[15] Renato De Mori. *Computer Models of Speech Using Fuzzy Algorithms*. Plenum Publishing Corporation, New York, 1983.

[16] R. De Mori and P. Laface. Use of Fuzzy Algorithms for Phonetic and Phonemic Labeling of Continuous Speech. *IEEE Transactions on Pattern Anal. and Machine Intell.*, Vol.2:136–148, 1980.

[17] G. Fant. *Speech Sounds and Features*. MIT Press, 1973.

[18] S. Wang, A. Sekey, and A. Gersho. An Objective Measure for Predicting Subjective Quality of Speech Coders. *IEEE J. on Select. Areas in Comm.*, Vol.10, 1992.

[19] Whonho Yang, M. Dixon, and R. Yantorno. A modified bark spectral distortion measure which uses noise masking threshold. In *IEEE Workshop on Speech Coding For Telecommunications Proceeding*, pages 55–56, 1997.

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

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