




Article

An Effective and Efficient Genetic-Fuzzy Algorithm for Supporting Advanced Human-Machine Interfaces in Big Data Settings

Alfredo Cuzzocrea ^{1,*}, Enzo Mumolo ^{2,†} and Giorgio Mario Grasso ^{3,†}

¹ iDEA Lab, University of Calabria, 87036 Rende, Italy

² DIA Department, University of Trieste, 34127 Trieste, Italy; mumolo@units.it

³ COSPECS Department, University of Messina, 98121 Messina, Italy; gmgrasso@unime.it

* Correspondence: alfredo.cuzzocrea@unical.it

† These authors contributed equally to this work.

Received: 21 November 2019; Accepted: 18 December 2019; Published: 31 December 2019



Abstract: In this paper we describe a novel algorithm, inspired by the mirror neuron discovery, to support automatic learning oriented to advanced man-machine interfaces. The algorithm introduces several points of innovation, based on complex metrics of similarity that involve different characteristics of the entire learning process. In more detail, the proposed approach deals with an humanoid robot algorithm suited for automatic vocalization acquisition from a human tutor. The learned vocalization can be used to multi-modal reproduction of speech, as the articulatory and acoustic parameters that compose the vocalization database can be used to synthesize unrestricted speech utterances and reproduce the articulatory and facial movements of the humanoid talking face automatically synchronized. The algorithm uses fuzzy articulatory rules, which describe transitions between phonemes derived from the International Phonetic Alphabet (IPA), to allow simpler adaptation to different languages, and genetic optimization of the membership degrees. Large experimental evaluation and analysis of the proposed algorithm on synthetic and real data sets confirms the benefits of our proposal. Indeed, experimental results show that the vocalization acquired respects the basic phonetic rules of Italian languages and that subjective results show the effectiveness of multi-modal speech production with automatic synchronization between facial movements and speech emissions. The algorithm has been applied to a virtual speaking face but it may also be used in mechanical vocalization systems as well.

Keywords: genetic optimization; fuzzy algorithms; advanced human-machine interfaces; humanoid robotics

1. Introduction

1.1. Preliminary Insights: Advanced Human-Machine Interfaces and the Emerging Big Data Trend

Nowadays, *big data* (e.g., References [1–5]) is the emerging trend that is pervading our life. Among the so many topics in big data research, *human-machine interfaces combined with big data processing* (e.g., References [6–8]) is a critical area with several interesting and challenging aspects in both the research and industrial application context.

Basically, this line of discipline aims at integrating the well-known *big data analytics* area (e.g., References [9–12]) with the enormous size of information coming from typical human-machine interfaces (e.g., References [13–17]). Indeed, these sources of information generate massive amounts of data, so that analyzing these (big) data repositories with big data analytics plays a critical role, in order to feedback the state model underlying the target human-machine interface, for optimization purposes.

Figure 1 reports the reference architecture of the integration between human-machine interfaces and big data, where our work co-locates. Here, several layers are identified:

- *Text-To-Speech Algorithms*: the layer where text-to-speech algorithms (like ours) execute;
- *Big Data Understanding*: the layer where big data are derived and understood;
- *Big Data Repository*: the layer where derived big data are stored;
- *Big Data Analytics*: the layer where knowledge is extracted from big data;
- *Big Knowledge*: the layer where the final knowledge is made available to users/applications.

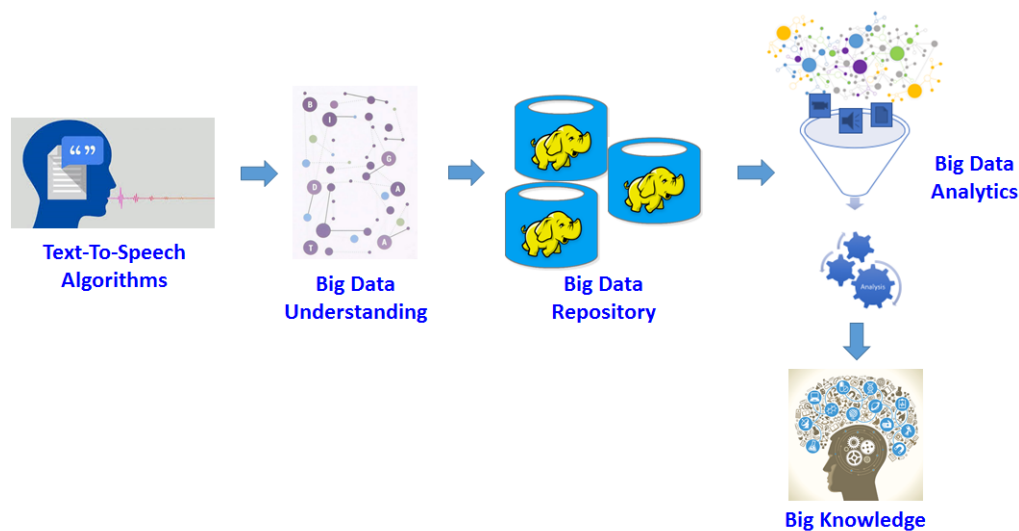


Figure 1. Human-machine interfaces and big data integration.

Our research mainly focuses on this area. Indeed, in more detail, we consider a special “piece” of work represented by the problem of supporting *text-to-speech and articulation* of a target humanoid, which can be reasonably considered as a basic step of a major system for achieving the interaction between advanced human-machine interfaces and big data management/analytics. Therefore, our proposed genetic-fuzzy algorithm can be reasonably considered for as a component of the wider goal of supporting advanced human-machine interfaces.

1.2. Motivations

A humanoid is a robot designed to work with humans and for them. Humanoid robots have been designed to allow a robot to offer services to humans in difficult or repetitive situations. Therefore, a fundamental issue in humanoid robotics is the interaction with humans. In Reference [18] the Cog project at MIT is described, as well as its Kismet project, which were developed in the hypothesis that a humanoid intelligence requires humanoid interactions with the world.

Indeed, nowadays there is growing interest in humanoid robotics. An autonomous humanoid robot is a complex system with a certain degree of autonomy and provides useful services to humans. Its intelligence emerges from the interaction between data collected by the sensors and the management algorithms. Sensor devices provide information about the environment useful for motion, self-localization, and avoidance of obstacles in order to introduce responsiveness and autonomy.

It would be easier for a humanoid robot to interact with humans if it was designed specifically for that purpose. For this reason, humanoid robots tend to imitate the form and the mechanical functions of the human body in some way to emulate some simple aspects of physical (i.e., movement), cognitive (i.e., understanding) and social (i.e., communication) skills of humans. Man-robot interactions and dialogue modes have been extensively studied in recent years in the robotics and AI communities.

A very important area in humanoid robotics is interaction with human beings. Since language is the most natural means of communication for humans, human-humanoid vocal interfaces should be

directed to facilitate and improve the way people interact with a robot. Basically, conversational interfaces are built around speech synthesis, voice recognition, and semantic inference engines technologies. In addition to the known problems of accuracy in presence of noise, one of the main problems in voice recognition for interaction with a robot is that the human operator is generally distant from the microphone, which is mounted on the robot itself; one way to overcome this problem is to use beamforming [19] algorithms.

The human face is an important communication channel in face-to-face communication. Through the face, many information are displayed: verbal, emotional or conversational. All of them should be carefully modeled in order to have natural looking facial animation. Speech driven facial animation has been researched much during the past decade (e.g., References [20–22]). For human-like behavior of talking head, all facial displays need to be accurately simulated in synchrony with speech. Most efforts, so far, have focused on the synchronization of lip and tongue movements, since those movements are necessary in order to have speech intelligibility.

As for voice synthesis, it is worth noting that in Humanoid robotics is very important, if not necessary, to produce speech using a mechanical speaking head for the reason we describe briefly. The motivation of the mechanical development of the speaking robot, as described in Reference [23], may be to conduct research on the vocal motor brain control system and to create a dynamic mechanical model for reproducing human language. In addition to telecommunications applications, medical training devices and learning devices mentioned in Reference [23], we believe that a mechanical head derived from articulate features can ultimately lead to a robotic mechanical face with natural behavior. It is known, in fact, that there is a very high correlation between the dynamic of the vocal tract and the behavior of the facial movement, as emphasized by Yehia et al. in Reference [24]. This was used in Reference [25] to develop the natural animation of a talking head. In addition, as regards the talking mechanical robot, if a mechanical vocal stretch is embedded in an artificial head that emulates a human head, the artificial head should have natural movements during the production of speech spoken by the robot, provided that the artificial head is tied to the articulators by means of a kind of elastic joint.

In any case, the mechanical vocal tract must be dynamically controlled to produce a speech language, since human speech requires transitions between phonemes. This requires sufficient knowledge of the complex relationships that govern human vocalization. So far, however, no enough research has been carried out on the brain control system, so the production of the words is not yet perfectly comprehensible [26]. This type of knowledge relates to the articulation synthesis of voice, which describes how to generate a word from a given movement of articulators (articulation trajectory). By exploiting our current knowledge, we developed fuzzy rules that link the places of articulation with corresponding acoustic parameters. The degrees of membership of the places of articulation are adapted to those of the human being who trained the system.

The purpose of this study is to build a robot that acquires a vocalization capability similar to human language development. In this paper, we consider a human-robot interaction scenario in which the robot has a humanoid speaking head. This does not necessarily mean a robotic head; in fact, we considered a software speaking head, which is a graphically image drawn on the computer screen. In addition to the obvious differences between these two alternatives, the common part is that the same articulation model should be developed in both cases. This model is responsible for estimating articulate parameters from natural language.

Our original contribution is a new articulation model based on automatic learning. The algorithm attempts to reproduce a voice signal by optimizing an articulator synthesizer and, in this way, learns the articulator characteristics of the speaker who trained the system. The idea is inspired by the theory of the *mirror neurons* that there are neurons in the human brain, linked to the control of human articulators, which are activated after listening to vocal stimuli [27,28]. This means that human perception of speech is related to its generation. Our system can produce an artificial utterance from arbitrary written text imitating the articulation characteristics of a given speaker.

The articulators movements estimated by speech could be used to control a mechanical speaking head or, as in our case, a talking head drawn on a monitor placed on a service robot. Compared to other works on acoustic and articulation mapping, we present a method to estimate the places of articulation of an input voice through a new computational model of human vocalization. The result is that accurate articulation movements are estimated from those of the human who trained the system. The ability to produce individualized facial communication is becoming more important in modern user-computer interfaces, for example in virtual storyboards used for remote conferences. However, the proposed system does not imitate the whole face of the trainer, but only the face movements, due to movements of the articulatory organs that the trainer uses to pronounce vocal sentences. In our model implementation, only the rules for Italian phonemes have been considered. This does not limit the generality of the method—if other languages are considered, new fuzzy rules must be added. We remark that we automatically extract the fuzzy rules with a genetic programming algorithm.

1.3. Goals of Our Research

This paper describes an algorithm that automatically learns how to generate artificial voice by reproducing human voice through articulation synthesis and an optimization scheme. Our algorithm is therefore suitable for giving vocalization capabilities to talking robots by imitating the human learning process. With “vocalization ability” we refer not only to the generation of artificial voice, but also to the facial movements that accompany the generation. The algorithm requires that a human operator, as a caregiver, read a list of words aloud. The spoken words, analyzed at an acoustical and articulatory level, are automatically divided into pre-defined speech units and a knowledge base of such units is automatically built. Therefore, the algorithm can generate vocalizations from unrestricted text by concatenating the basic speech units.

Other systems to generate artificial speech together with the related facial movements have been published (e.g., References [20–22]). Our algorithm, however, automatically generates artificial speech using the places of articulation estimated during a training phase. Optimization is realized using a genetic algorithm, on the basis of the similarity of the synthetic utterance with the original speech. In this work we estimate the face movements from the optimized degrees of membership of the places of articulation. In fact, there is a strong structural link between places of articulation and facial behavior during vocalization. The facial movements obtained were used to move a virtual face naturally during vocalization. It is worth noting that both artificial expressions and facial movements seem to clone that of the human operator who served as a virtual face caregiver, thus giving the opportunity to personalize the virtual head. It is important to note that we implemented the training phase on a GPU device so that the training phase is very fast. Time needed for training is only a few tenths of minutes needed to read a list of words.

A general model of verbal/facial communication is shown in Figure 2. By adapting the Figure 2 to our talking head, the concept and message generation refer to a cognitive level, and the phonatory control refer to the vocalization level we are describing in this paper. We propose a novel design of a vocalization module which is the module that converts a written message into artificial speech and facial movement. Therefore, the issues we describe here are the following:

- articulatory vocalization of unrestricted text;
- generation of facial movement;
- synchronization of artificial speech with facial movements.

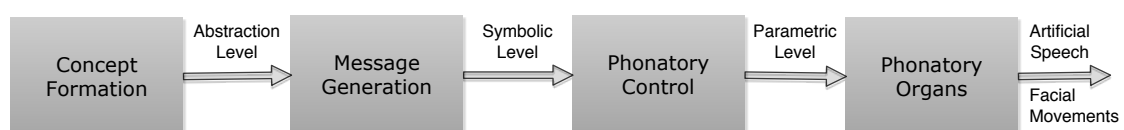


Figure 2. General scheme of communication that is used by our algorithm.

In talking head the articulatory configuration during vocalization is fundamental, because it could be used to control a mechanical talking face, such as described in Reference [29].

Our algorithm is similar to the process of human language acquisition. Indeed, in the early days of life, human infants try to imitate the speech they heard; this process is recognized as a language acquisition process driven by a mirror neuron system [30]. Similarly, our algorithm acquires vocalization through interaction with a human. It is important to note that it is not necessary to produce high quality artificial speech. Since our speech is used in a similarity comparison process, only low quality, yet intelligible speech, is enough. For this reason we use the simplest approach for artificial speech generation, that is the concatenation approach [31]. A number of basic vocal units is defined, and a base of knowledge of articulatory and acoustic parameters for their generation is built. Training iterates as long as the knowledge base is not full. The acoustic parameters and corresponding articulators are obtained with the fuzzy-genetic articulation model proposed in Reference [32], whose parameters are optimized to produce a language similar to that of a human user. In addition to guiding the synthesis of artificial language, knowledge of articulate movements is critical to providing facial movements since an important link between the movements of the internal organs of the vocal tract and the facial movements has been recognized by many researchers [33,34]. An important characteristic of our algorithm is that a perfect synchronization of facial movement with artificial utterances is obtained.

1.4. Paper's Contributions

In this paper, we provide the following contributions:

- an effective and efficient genetic-fuzzy algorithm for supporting automatic vocalization acquisition from a human tutor, by delivering artificial speech and facial movements (both synchronized);
- a comprehensive review of state-of-the-art proposals in the related scientific area;
- an extensive experimental evaluation of the proposed algorithm where we stress quantitative and qualitative parameters.

1.5. Paper Organization

The structure of this paper is the following. In Section 4, a brief overview of our proposed algorithm is provided. Section 2 deeply presents and discusses research that is relevant for our work. In Section 3, some preliminary considerations on speech synthesis are outlined. In Section 5, we provide the fuzzy-genetic articulatory model that is at the basis of our proposed algorithm. Section 6 reports on our proposed algorithm, which is the main result of our research. Section 7 describes some further optimizations we developed in order to further improve the effectiveness and the efficiency of our proposed algorithm. In Section 8, we discuss a possible exploitation of our proposed algorithm, which concerns with the control of talking heads. Section 9 describes our comprehensive experimental campaign along with derived results. Finally, in Section 10 we provide final comments and future work for our research.

2. Related Work

In this Section, we briefly describe some previous research that is related to our work.

The research activities carried out in the last few years for the face-to-face communication between humans and humanoid robots, as reported in Reference [35], also involved the creation of mechanically speaking heads, as we shall describe below.

At Waseda University, several prototypes for producing Japanese vowels and some consonant sounds, called WT-3, WT-2 and WT-1R, have been developed. WT-1R [36] has articulators (tongue, lips, teeth, nasal cavities and soft palates) and vocal organs (lungs and vocal cords). It can reproduce the human voice movement and has 15 degrees of freedom. The vocal movement of the WT-1R for vocals is constant. However, consonant sounds are produced by dynamic movements of the

vocal tract, to generate transitions of resonant frequencies (formant frequencies). Therefore, since Japanese voice generally consists of two phonemes, where the first is a consonant sound and the last a vowel, researchers at Waseda University have proposed considering the concatenation of three types of sounds (constant consonant sound, transient and vocal consonant sound). WT-2 [37] is an improvement of WT-1R; has lungs, vocal cords, vocal cavity and nasal cavity and aims to reproduce the vocal trait of an adult male. WT-3 [26] is based on human acoustic theory for the reproduction of human language; it consists of lungs, vocal cords and articulators and it can reproduce human-like articulation movements. The oral cavity was designed based on the MRI images of a human sagittal plan.

At the Kagawa University, a mechanical model of the human vocal apparatus was developed [38] using mechatronics technology. It implements a neural network mapping between the motor positions and the voice sound produced by auditory feedback in the learning phase.

In Bib:Asada the mechanical vocal tract is excited by a mechanical vibrator that oscillates at certain frequencies and acts as an artificial larynx. The shape of the vocal tract is controlled by a neural network.

However, the development of an anthropomorphic mechanical speaking robot is very difficult as shown in References [18,38]. On the other hand, graphic representations of human talking heads have been developed since many years. Generally, these graphics systems are connected to a speech synthesizer that plays a signal synchronized with the facial movements. Many solutions to this synchronization problem have been proposed since now. However, very few of them go through the articulatory movements of the phonatory organs.

In addition to controlling the mechanical phonemic organ of a mechanical speaking head, knowledge of articulating movements is essential to drive the movements of the face. In fact, many researchers have experimentally demonstrated that there is an important link between the movements of the internal organs of the vocal tract and the facial movements.

Similar to the approach proposed in this document, Nishikawa et al. [23,26] developed an algorithm that estimates articulator parameters using Newton's optimization.

H. Kanda et al., describe a computational model that explains the development process of human infants in the first acquisition period of the language [39]. The model is based on recurrent neural networks.

Sargin et al., propose a joint analysis of the prosody and head movements to generate natural movements [40].

Albrecht et al., introduce a method to add nonverbal communication to the talking head. This communication is realized by automatic generation of different facial expressions from the prosodic parameters extracted from the voice [41].

The SyncFace [42] system uses articulation-oriented parameters to recreate the visible articulation of a talking head during phonation.

The systems described so far require a pre-processing step. Real-time facial animation is described in Reference [43].

Further Comparison with Mechanical-Aware Approaches

Some researchers address the problem of vocalization from a mechanical point of view, for example, Reference [26]. They try to realize a mechanical replication of the human vocal tract, of the larynx and tongue. In addition, they develop various open and closed loop algorithms to control the movements of the mechanical organs for the production of artificial words. For example, in closed loop strategy, voice control parameters extraction is optimized by minimizing the distance between the original and the generated voice. Our work is similar to the work described in References [23,26] in the sense that we estimate from the input voice the articulating movements using a closed loop strategy. These parameters are then used to generate artificial replicas of the input statements on the one hand and to control the artificial phonatory organs on the other. The algorithm uses the fuzzy tone pattern introduced in Reference [32]. However, we must note that this paper represents a breakthrough point

compared to the algorithms described in Reference [32]. The main points of improvement, which are the contributions of this paper, are as follows:

- while the rules in Reference [32] refer to the Italian language, the rules developed in this paper use the International Phonetic Alphabet (IPA) [44] thus enabling the system to be extended to other languages;
- while in Reference [32] we use single words and short phrases, in this paper we describe a system that produces unrestricted vocalizations;
- while in References [32,45] we control only four points around the lip, that is four face muscles to produce facial movements, in this paper we extend the number of controlled muscles to the left and right cheek, thus leading to better facial movements;
- the algorithm described here uses the average pitch extracted from the tutor as the basis of the artificial prosody;
- finally, we include in this paper extensive subjective evaluations of the proposed system that were impossible to execute earlier.

Vocalization of unrestricted text is performed by acquiring, at an initial stage of training, a database of basic voice units appropriately defined. The acquisition is performed automatically using the speech of a human tutor. The tutor is asked to pronounce a number of statements that are analyzed and automatically segmented into the defined small units. A database is then created and the desired vocalization is obtained by concatenating the basic vocal units.

3. Preliminaries

In this Section, we describe and discuss some preliminary considerations that are founding our research.

In this work, we generate synthetic speech with the well known Cascade-Parallel formant synthesizer [46]. The synthesizer is driven by the following minimum set of eight acoustic parameters: $(AV, AF, F1, F2, F3, B1, B2, B3)$, namely the amplitudes for voicing and frication, three formant frequencies with the respective bands. It is well known that this is a quite old yet very simple approach to produce artificial speech. Indeed, the speech obtained in this way does not sound natural but rather, it is perceived as highly robotic.

The talking head produces artificial speech from unrestricted text. One of the first attempts to generate speech from arbitrary text is synthesis from concatenation approach [47], where sets of basic speech units are defined. Each speech unit contains at least a transition between phonemes (diphone).

In conclusion, our unrestricted text is generated by diphone concatenation where each diphone (transition) is modeled with the minimum set of eight acoustic parameters. The motivation of this simple approach is that, as regards the small number of parameters which drive the synthesizer to produce transitions, the small number of acoustic parameters simplify the fuzzy rule development process. As regards the diphone concatenation approach, by concatenation of basic speech units, the motivation is that the goal of this work is not to produce good artificial speech but to estimate good articulatory parameters, and concatenated speech is enough for that purpose.

In Figure 3 we show an example of transition of the first formant frequency between the two phonemes in a speech unit. For simplicity, each of the speech units we selected contain a single transition, thus leading to the minimum set of speech units for speech generation. In our set, we consider consonant-vowel and vowel-consonant transitions, affricate-vowel and vowel-affricate transitions and diphthongs transitions. The total number of basic speech units is 192, namely 140 consonant-vowel transitions and viceversa, 20 affricate-vowels transition and viceversa, 10 vowel-silence transitions and viceversa, 8 diphthongs, that is, vowel-vowel transitions and 14 transitions between silence and a consonant.

As shown in Figure 3, any transition is described by the *Initial*, *Duration*, *Final* and *Locus* values, we call $[I(\cdot), D(\cdot), F(\cdot), L_1(\cdot), L_2(\cdot)]$.

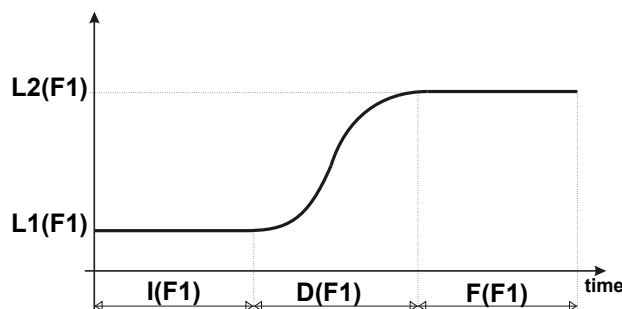


Figure 3. F1 transition trajectory example.

Using the $[I(\cdot), D(\cdot), F(\cdot), L_1(\cdot), L_2(\cdot)]$ parameters, the transitions can be easily concatenated. An example of concatenation of two trajectories is shown in Figure 4.

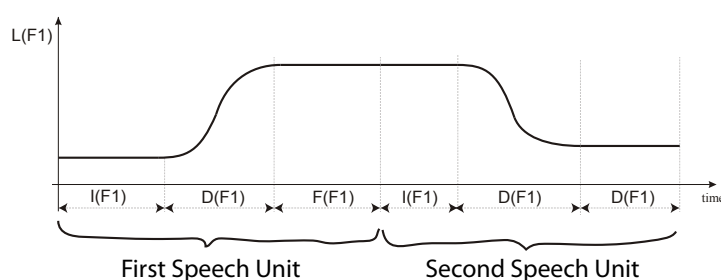


Figure 4. Formant trajectories concatenation example.

Along with the acoustic parameters that are used to synthesize the signal, we also store in the knowledge base the corresponding articulation configurations. The acoustic parameters are the transitions of formants and amplitudes (i.e., $[I(\cdot), D(\cdot), F(\cdot), L_1(\cdot), L_2(\cdot)]$). We considered the following twelve places of articulation: (*rounded, open, front, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar*).

Articulatory parameters are the genetically optimized degrees of membership of the places of articulation. The knowledge base of articulatory parameters is reported in Figure 5. Therefore, the knowledge base describes the membership values of all the possible transitions in the selected language for the operator who acts as trainer. Please note that the symbol *Sil* means silence (therefore */Sil-a/* is an initial part of a sentence starting with the */a/* vowel). Moreover, note that the symbols on the vertical axis are selected from the IPA alphabet.

	Round	Open	Anterior	Sonorant	Bilabial	Labiodental	Alveolar	Prepalatal	Palatal	Vibrant	Dental	Velar
/sil-a/												
/sil-e/												
/b-a/	0	0	0	0.8	0.6	0	0	0	0.8	0	0.4	0
/b-e/	0.6	0.8	0.5	0.9	0	0	0	0	0	0	0	0
/z-w/												

Figure 5. Basic speech units knowledge base.

4. Algorithm Overview

Our proposed algorithm is reported in Figure 6.

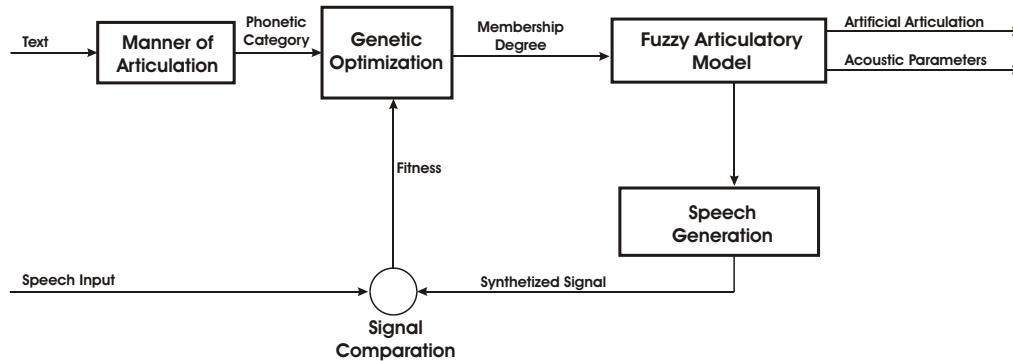


Figure 6. Block diagram of the optimization algorithm.

The proposed algorithm is divided into a first training phase and a second phase of facial movements generation. In the training phase (see Figure 7), the algorithm automatically learns the acoustic and articulation characteristics of the basic vocal units from the tutor’s voice. At the end of this phase, a knowledge base is produced. In the second phase, the algorithm uses the knowledge base to generate speech and face movements. We note that the training phase should be done once for each different trainer. Of course, if you do not want to personalize facial movements, just consider a standard speaker.

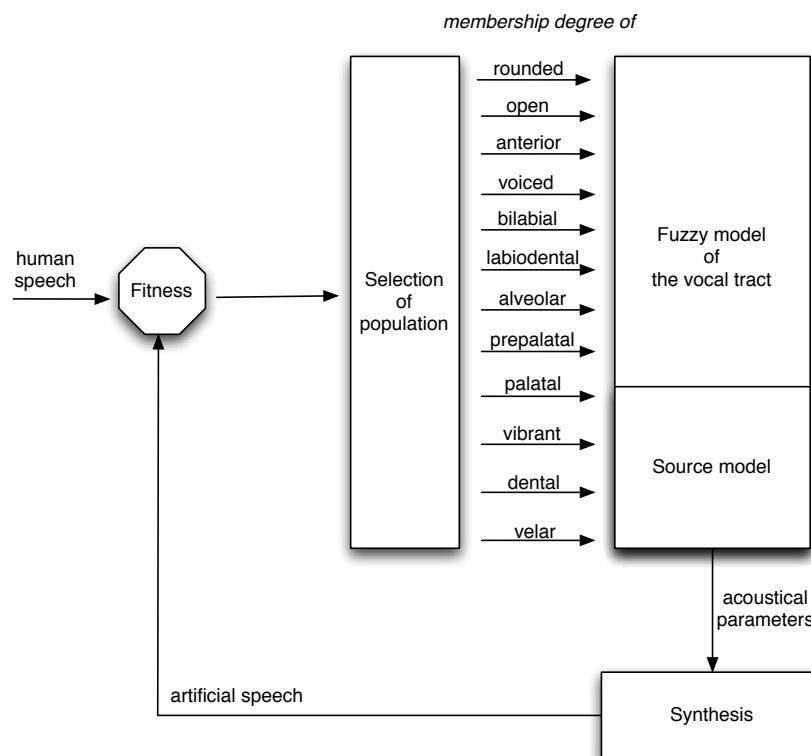


Figure 7. Training phase scheme.

It should be noted that, in this paper, we refer to the fuzzy rules defined for the Italian language. However, extension to other languages is simplified because we use IPA to build fuzzy rules.

As explained above, the knowledge base describes the membership degrees of the basic vocal units which are automatically extracted from the set of words spoken aloud in the training phase. Each vocal unit contains a transition between phonemes. Each vocal unit is stored along with its contextual information, that is, the previous and subsequent phonemes. This information is used in the concatenation phase to choose the segment that best suits the actual phonetic environment of the unit.

Under a more general umbrella, the final goal of our 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, and, indeed, Figure 6 can be seen as the block diagram of the system for adaptive learning of human vocalization.

We now include some basic considerations to further clarify the optimization process, which is central in our work. According to the classical Distinctive Feature Theory [48], phonemes are described in terms of *Manner of articulation* and *Place of articulation*, both described in binary form. The former is related to the constraints imposed to the air flow in the vocal tract while the latter refers to the point where the constraint is located. We consider the following six categories of manner of articulation: *vowel*, *liquid*, *nasal* for the sonorant sounds, and *fricative*, *affricate*, *stops* for obstruent sounds. In the sonorant sounds, the air flows without any obstruction or under moderated obstruction represented by the tongue or the velum. Obstruent sounds are generated under tight or complete closure in the vocal tract represented by tongue or lips. The purpose of the genetic algorithm is to find the degrees of membership of the places of articulation that optimize the similarity between input and artificial voice. Our assumption is that the optimum degrees of membership are related to the physical configuration of the human phonatory organs. For example, if we found that a phoneme is an open vowel with a membership degree of openness of 0.6, we assume that the mouth is open at a level of 60% of its maximum opening size. If a vowel has a membership degree of roundness 0.8, we set the amount of the amount of lip rounding to 80%. This assumption can be verified indirectly by the quality of the obtained results, reported in Section 9.4.

The optimization scheme is shown in Figure 6. The first left block, that is, the “manner of articulation” block, is fed by the text corresponding to the sentence pronounced by the operator. This block evaluates the manner of articulation of the phonemes in the sentence. For example, if the word is “nove” in Italian (“nine” in English), the manners are: nasal, vocal, fricative, vocal. This information is used to construct the chromosome and to select the correct rules from the fuzzy articulation module.

To summarize the objective of the whole system, in order to generate facial movements of a written arbitrary text, we compute first the corresponding articulatory description and we map articulation to facial movements. Articulatory description is obtained through articulatory synthesis of the voice corresponding to the written input text.

In the next sections, we focus on the main components of our proposed algorithm.

5. Fuzzy-Genetic Articulatory Model

We describe here the basis of our algorithm, which is developed upon a fuzzy-genetic articulatory model, proposed in Reference [32]. This model estimates, with a genetic optimization algorithm, the places of articulation membership degrees of a user speech. The parameters are used to produce synthetic speech which is compared to the input signal using a genetic approach. In this paper, the approach has been extended to cover the IPA phonetic symbols rather than the phonetic rules of a particular language, and the mimicking system has been extended to cover unrestricted text.

According to our point of view, however, the place of articulation is characterized in fuzzy form. Each phoneme is described with eighteen variables, six of which are Boolean and represent the manner of articulation while the remaining twelve are fuzzy and represent the place of articulation. Therefore, our phonetic description appears as an extension of the classical Distinctive Feature Theory, because a

certain vagueness is introduced in the definition of the place of articulation. For example, with a fuzzy description, an /a/ phoneme can be represented as:

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

which means that /a/ is a vowel with a frontness of 0.12, an openness of 0.9 and a roundness of 0.32.

It is important to remark that :

- All the fuzzy sets for acoustic parameters have trapezoidal shape and are defined as follows: L1 (Initial locus), I (Initial range), D (Duration), F (Final interval), L2 (Final position). The values are divided in High, Medium, Low.
- The estimation of degrees of membership of the places of articulation is made when the manner of articulation of an utterance is known, because the fuzzy rules are structured in banks. The easier way to perform manner of articulation estimation is from the text of the input utterance. However, this is not a limit because normally the estimation of facial movements, which is the goal of our algorithm, is needed to move talking heads during text to speech translation.
- Since the degrees of membership of the places of articulation depend from the phonetic context, their optimal estimation should be performed in real time. This would require an excessive computational power. Our approximated but feasible approach is to estimate the degrees of membership once for a given phonetic context and a given speaker, stored in the knowledge base and used in any context and for any speaker.

As regards the fuzzy rules, consider for example the following consonants: /p/, /t/. According to IPA, they are described as voiceless, bilabial, plosive and voiceless, alveolar, plosive respectively. Calling p_0 and p_1 respectively the first and second phonemes, some of the fuzzy rules involved in generating the acoustic parameters for the generation of the transitions /pa/, /pi/, /pu/, /ta/, /ti/, /tu/, where /a/, /i/, /u/ are vowels, are the following:

```

if p0 is plosive and p1 is voiced then
L1(AH) is high;
I(AH) is medium;
D(AH) is low;
F(AH) is low;
if p0 is plosive and p1 is open then
L2(F1) is medium
if p0 is alveolar and p1 is voiced then
L1(F1) is medium;
L1(F2) is medium;
L1(F3) is medium;
if p0 is alveolar and p1 is open then
L1(F1) is low;
L1(F2) is low;
L2(F3) is low;

```

Clearly, the goal of these fuzzy rules is to generate the transitions of the acoustic parameters $AH, F1, F2, F3$ that generate the correct transition. The rule decoding process is performed by a de-fuzzification operation with the fuzzy centroid approach.

Another example is a transition of a vowel towards a vowel. In this case, the opening and the anteriority of the target phoneme determine the values of the first three formants. This knowledge can be formalized as follows:

```

if p0 is voiced and p1 is open then
L2(F1) is medium;
if p0 is voiced and p1 is anterior then
L2(F2) is medium high;
if p0 is voiced and p1 is notAnterior then
L2(F2) is low;
if p0 is voiced and p1 is Round then
L2(F3) is low;
if p0 is voiced and p1 is notRound then
L2(F3) is medium;

```

Such rules can be automatically determined with genetic programming for any language, provided that a good speaker in the selected language is available as tutor. We note, however, that automatic generation of fuzzy rules is not included in this paper, and thus we assume in this paper that the fuzzy rules are available in advance. On the other hand, the genetic optimization considered in this paper aims at computing the values of the degrees of membership for the articulatory features which minimize the distance from the input signal.

The fitness used in our genetic algorithm is based on the computation of an objective distance between original and artificial sentences [49]. The objective measure transforms the original and artificial voice sentences into a representation which is linked to the audio psycho-physical perception in the human auditory system. This measure predicts the subjective quality of the artificial signal. The measure requires alignment between the two signals, which is performed with DTW [50], where the slope described in Reference [51] is used. The accumulated distance D along the non linear mapping between the two signals is the distance between the original and synthetic utterances, we call in the following X and Y respectively. Clearly the genetic algorithm tries to minimize the distance $D(X, Y)$. Thus, the algorithm fitness function used for the genetic optimization of the Place of Articulation (PA) is:

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

The genetic optimization algorithm finds the membership values of the fuzzy variables that maximize the fitness, namely $PA = \operatorname{argmax}\{Fitness(PA)\}$, $PA = \cup PA_i, i = 1, \dots, 12 \cdot N$. The estimation problem can be viewed as an *inverse articulatory problem*, where articulatory configurations are estimated from speech. It is well known that this problem has many solutions. Hence, during optimization some constraints are imposed, as the impossibility that a plosive phoneme be simultaneously dental and velar, or that voiced consonants be completely voiced. The constraints are penalty factors added to the fitness. Calling P_j the penalty factors and N_c the number of constraints, the optimization problem can be formalized as described in Equation (2).

$$PA = \operatorname{argmax} \left\{ \frac{1}{D(X, Y)} + \sum_{j=1}^{N_c} P_j \right\}. \quad (2)$$

As regards facial movements, we have selected a set of muscles in a virtual face, as shown in Figure 8.

The muscles can move the upper and lower lips, the lateral facial movements and the mouth movements. We realize a link between articulatory positions and these muscle, automatic synchronization between facial movements and artificial speech is automatically obtained, since artificial speech is obtained with articulator synthesis.

Note that the facial movements are estimated by the speech signal acquired by the tutor. The corresponding facial movements are related to the physical emissions of the utterances and do not include any particular facial expressions or feeling.

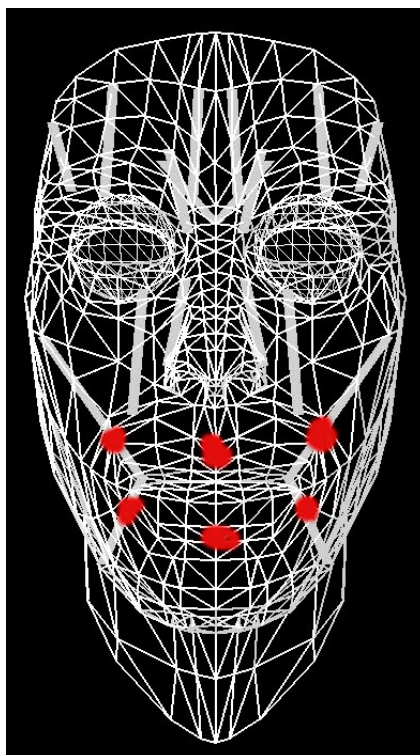


Figure 8. Muscles on the talking head.

6. An Innovative Automatic Learning Algorithm for Supporting Advanced Human-Machine Interfaces

In this section, we focus on the description of the human-machine interface of the algorithm proposed in this paper. These aspects are very relevant in practical applications and this algorithm provides a decisive contribution to this in advanced interface with respect to state of the art.

In addition to producing artificial voice from a written text, our algorithm also produces facial movements by estimating articulator parameters from the entry voice.

6.1. Training Phase

During an initial phase, the membership levels of articulator parameters for the speaker who trains the system are derived. Using the membership degrees along with the fuzzy rules, the acoustic parameters are generated to generate the artificial signal. Membership degrees are stored in the knowledge base. From an operational point of view, the initial phase is realized by making the human operator say a series of words that are presented to him.

The genetic optimization module estimates, using the fuzzy set of rules, the $[I(\cdot), D(\cdot), F(\cdot), L_1(\cdot), L_2(\cdot)]$ values of the acoustic parameters used to generate artificial speech by concatenation of the basic speech unit contained in each word. The list of words to pronounce cover the whole of the basic vocal units. These parameters are used to fill the basic vocal unit knowledge base shown in Figure 5. which describes the estimates degrees of membership of related articulator parameters used of the phonation organs on one side and the parameters for facial control on the other side.

6.2. Synthesis Phase

The input of the synthesis module is the text to be converted into an artificial voice by concatenating simple elementary vocal units. Using the stored membership degrees, the parameters $[I(\cdot), D(\cdot), F(\cdot), L_1(\cdot), L_2(\cdot)]$ are obtained from the fuzzy rules and the overall path of acoustic parameters related to the artificial phrase is constructed by concatenation. From this trajectory, artificial voice is

generated by synthesis for formants. To ensure greater voice naturalness, the excitation sequence is generated by adding a fricative noise to the model proposed in Reference [52], as shown in Figure 9.

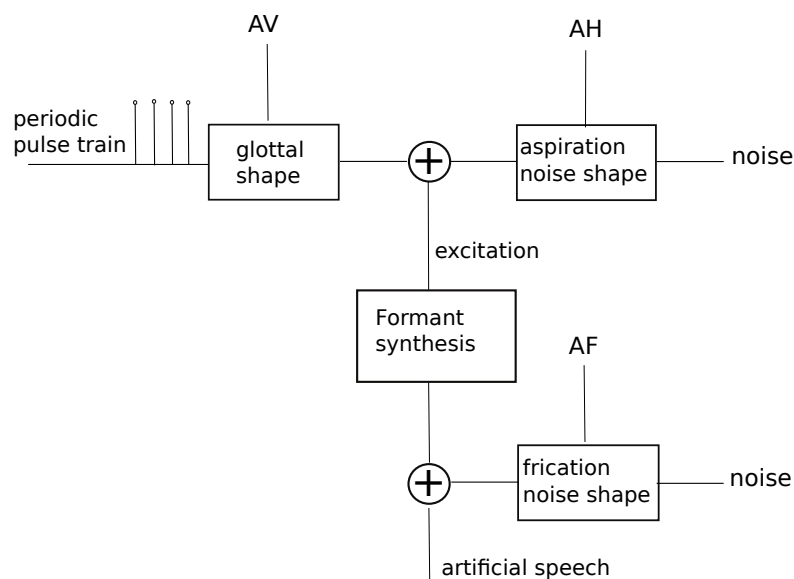


Figure 9. Mixed excitation sequence.

It is well known that one of the possibilities to improve the quality of the synthetic signal by diphone concatenation is to acquire a large database of basic units, from different phonetic contexts. During synthesis, it is possible to search for the best units, according to speech quality criteria. The goal of our algorithm is to estimate good facial movements corresponding to a synthetic speech through articulatory analysis. To this goal it is not important to achieve high speech quality, therefore we use the smallest possible dataset of speech units.

6.3. Articulatory to Facial Movements Mapping

It is known that facial movements during phonation are linked to articulation movements as shown by many authors, for example, Reference [24]. It is commonly believed that most of the correlation between facial movements and articulators can be expressed by linear transformation. Actual measurement of articulator movements and the corresponding articulation movements is possible but can be problematic; so we used the MOCHA-TIMIT [53] database containing articulator and facial parameters for English phonemes. The MOCHA-TIMIT database collects data using Electro Palathograph (EPG) and Electromagnetic Articulograph (EMA). The algorithm presented in this paper, at the time, uses the Italian language. Extending fuzzy rules to English is under development; for this reason, we used the available database beforehand. We have carefully selected a subset of MOCHA-TIMIT phonemes that are similar in both languages. This because, in our work, we have developed the fuzzy rules for Italian.

Because our fuzzy-genetic optimization estimates the articulator parameters of a spoken word, a mapping between articulator and facial parameters is required to obtain the corresponding facial movements. To make this mapping, unlike what is proposed in Reference [24], we trained a simple feedforward multilayer neural network. The network describes a nonlinear mapping (Figure 10) between articulatory and facial movements, as described in Reference [54].

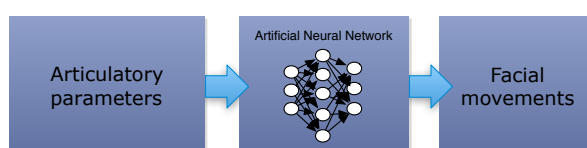


Figure 10. Articulatory to facial movements mapping.

For supervised training, we give as input to the network the articulatory parameters vector produced by the fuzzy-genetic module and as output the corresponding facial positions vector derived by the MOCHA-TIMIT database. With our fuzzy genetic algorithm we have obtained, for any given phoneme in the input sentence, several input vectors $g = ((rounded, open, front, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar))$ and from the MOCHA-TIMIT database the corresponding desired output vector $x = (UpperLipX, UpperLipY, LowerLipX, LowerLipY)$. Training is performed with Levenberg-Marquard algorithm with input/output parameters are extracted from 460 sentence.

The output vector is applied to the Geoface model implemented by Keith Waters [55] shown in Figure 11. The model employs a large number of polygons for modeling the whole face. The vector is applied to the coordinates of the corresponding textures of the model.

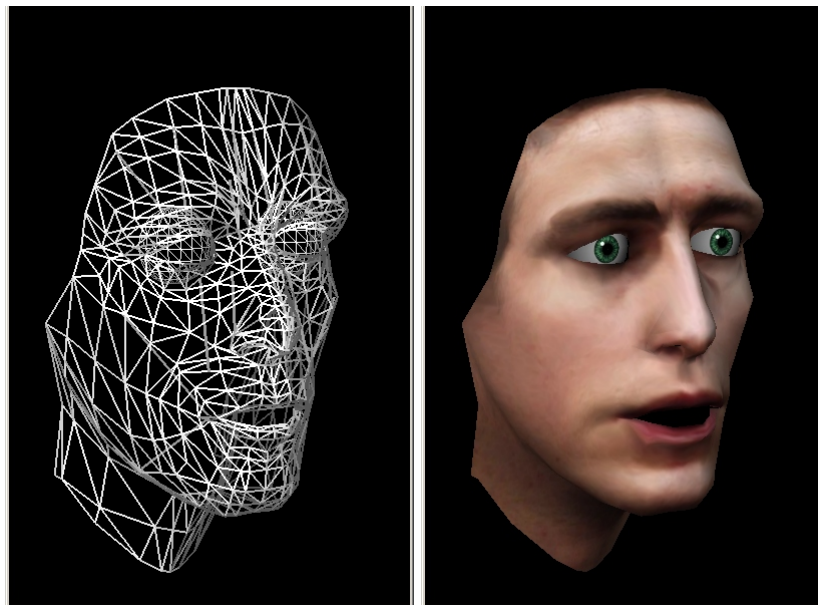


Figure 11. Geoface head model.

7. Genetic Optimization of Articulatory and Acoustic Parameters

In this Section, we discuss some optimizations that are oriented to magnify effectiveness and efficiency of our proposed algorithm.

Our Genetic optimization algorithm, described in Figure 6, 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.

7.1. Genetic Optimization Module

The optimal membership degrees of the articulatory places minimize the distance of artificial speech from the input speech; the text of input speech is given at input. From it, phonemes classification in terms of manner of articulation is performed.

Variables coding is a fundamental part of any genetic algorithm. We used five bits for coding the fractional part of the degrees of membership. A chromosome is built by concatenation of the binary coding of all the degrees to be optimized. In Figure 12 we report an example of the chromosome for the optimization of the degrees of membership in case of three phonemes.

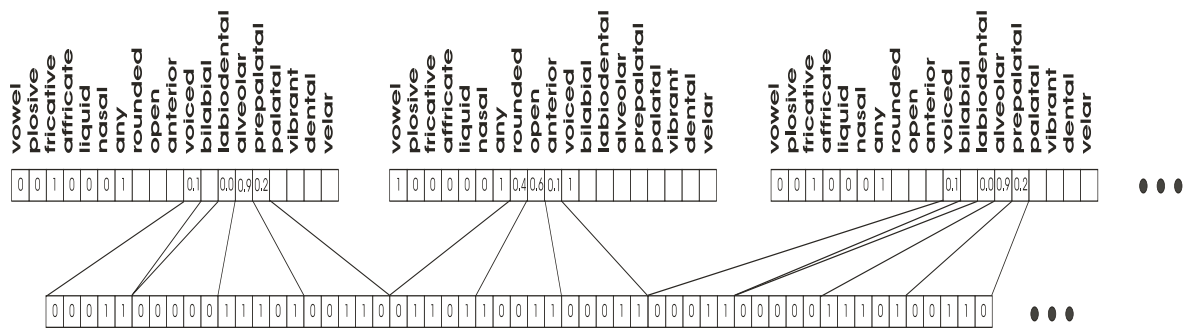


Figure 12. Example of the chromosome obtained by coding.

During genetic optimization this binary string is modified using mutations, that is, by randomly changing each bit of the chromosome with a constant mutation rate equal to 2%.

7.2. Fitness Computation and Articulatory Constraints

Another fundamental aspect of this algorithm is fitness calculation, which is indicated by the large circle in Figure 6. The similarity between the original and the artificial voice is the optimized fitness of the genetic algorithm. To represent this similarity, the spectral distortion based on the Bark scale (MBSD) derived from Reference [56] has been used. The MBSD measure is based on a psycho-acoustic term related to the intensity of the auditory sensation. In addition, we considered an estimate of the threshold of noise masking. This measure is used to estimate the similarity between the artificial voice signal generated by our algorithm with respect to the original signal.

More in detail, the original and artificial phrases are first aligned with DTW [50] and then divide into frames. Between the frames, the mean Euclidean distance between the spectral carriers obtained by Bark scale filters is then calculated.

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

$$AP = \operatorname{argmax} \left\{ \frac{1}{D(X, Y)} + \sum_{j=1}^{N_c} P_j \right\}, \quad (3)$$

where AP is the membership degrees of the articulatory parameters, N_c is the number of constraints and P_j are the penalties.

8. Possible Exploitation: Towards Control of Talking Heads

In this Section, we discuss some possible exploitation of the proposed algorithm.

The articulation trajectories, estimated by the algorithm, are transformed into facial parameter trajectories for controlling the movements of an animated head while vocalizing. As a preliminary step towards this goal, and to verify the correctness of the algorithm, we animated a mid-sagittal section of a human head using the descriptive procedure. Note that animation movements are automatically synchronized with the artificial vocalization produced..

In Figure 13 a sequence of four frames of the animation of a mid-sagittal section of a human head related to the phonemes /n/, /o/, /v/, /e/ is reported.



Figure 13. Sagittal sections of a talking head for the phonemes /n/, /o/, /v/, /e/.

9. Experimental Assessment and Analysis

This section contains the description of our complete experimental campaign together with the derived results. In particular, our experiments have been designed and developed to evaluate the true quality of synthetic discourse.

9.1. Experimental Method

It is worth recalling that the purpose of our algorithm is to generate facial movements during phonation for natural man-machine communication. To this end, we first estimate articulatory movements from spoken speech using our *Fuzzy – Genetic* optimization algorithm and then we transform the articulatory parameters into facial movements parameters. Many authors, for example Jintao Jiang et al. [57] have shown in fact that there exists a strong relation between articulatory and facial movements. Facial parameters are finally used to move the points corresponding to the involved facial muscles in a Geoface-articulated bone model. To measure the goodness of the generated facial movements we firstly test the quality of the synthesized speech in order to be sure that it is sufficient for supporting the subsequent optimization algorithm and then the quality of facial movements. As a matter of fact, the better the voice generated, the better the facial movements. The test of the generated speech is carried out with objective and subjective tests. We first test the objective quality of the generated formant transitions, which must be quite close to those extracted from the spoken voice. Then, the naturalness, quality and synchronization between facial movements and artificial vocalization are assessed with subjective experiments.

Clearly, the synthetic speech is not confused with naturally spoken speech, whose intelligibility mainly depends on its SNR. Many authors pointed out that since the perception of spoken speech is normally accompanied by face movement, the visual part of speech should be an important perception element [58]. For this reason, graphical talking heads synchronized with synthetic speech has been developed [59]. Subsequently, the perceptual benefits of the visual speech have been experimentally highlighted. We developed a talking head too, and in this section its performance over existing systems is reported.

The decrease of intelligibility or naturalness of synthetic speech can be modeled as a noise component added to the original speech, introducing the concept of *EquivalentNoiseLevel*. We use this concept to predict, when not available, results on the benefits of the audio-visual speech over only audio. In this paper, we assume that visual speech is obtained with the viseme-driven approach described in Reference [60] for all the considered speech synthesizers. In viseme-driven speech animation, each key pose is associated with a viseme, that is the position of the lips, jaw and tongue when producing a particular sound [61]. In Reference [60], the results of a subjective test of naturalness of the viseme-driven system are reported. In the test, 36 participants with normal hearing and vision tested individually the visual speech synthesis of a number of words with meaning. The average naturalness was about 0.84. This coefficient was used to weight the audio-visual subjective results of our tests.

9.2. Experiment 1: Analysis of the Transitions Generated by the Fuzzy-Genetic Model

The six phonetic classes considered in our system are the following: *vowels(VO)*, *plosive(PL)*, *fricatives(FR)*, *affricates(AF)*, *liquid(LI)*, *nasals(NA)*. Of course there are other classes, but we think that these six classes can be sufficient to cover most of the western languages. In theory, the six classes lead to thirty-six types of transitions between classes, like VO-PL or FR-VO; considering the transitions between phonemes, in theory we have 7560 transitions. However, many of them have been excluded, because or impossible or rarely used, so the resulting number is reduced to one hundred and ninety two (192). For the sake of completeness we report below the list of transitions used in our algorithm, where V stands for vowel and C for consonant.

- 70 elements /CV/, that is, such as /ba/ or /ka/;
- 70 elements /VC/, that is, such as /ap/ or /as/;
- 10 elements / \overline{CV} /, which are transitions between a, such as /tʃa/;
- 10 elements / \overline{VC} /, which are transitions between a vowel and an affricate;
- 8 elements /VV/, which are diphthongs;
- 14 elements /-C/, which represent transitions between silence and a consonant;
- 5 elements /V-/, which are transition between vowels and silence;
- 5 elements /-V/, which are transition between vowels and silence.

For each transition the human tutor utters a very short utterance, each containing only one transition. This operation is similar to the Knowledge Base Acquisition (KBA) discussed in Section 4. The difference from KBA is that in that case the tutor utters a suitable number of longer words containing all the one hundred ninety two transitions. The longer word is then automatically segmented into diphones, so all the diphones are extracted from a phonetic context, so that during synthesis the best diphone is selected.

In this experiment we just test the quality of the phonetic tracks, leaving to the KBA phase the task of acquiring the articulatory data which will be used for generating the articulatory movements. The analysis of the quality of the generated transitions is performed following these five steps:

1. Generation of the synthetic version of each transition using the *Fuzzy – Genetic* algorithm. In this way each transition, say, between the plosive /p/ and the vowel /a/, namely the /pa/ diphone, has a naturally spoken version /pa/ and a synthetic version / \overline{pa} /.
2. Phonetic speech alignment of all the segments with the same transition, for instance align /pa/ and / \overline{pa} / or /at/ with / \overline{at} / and so on using the DTW algorithm which is simple and efficient for this purpose [62]. This operation synchronizes the phonetic content of the natural and synthetic segments related to the same transition.
3. Application of the built-in formant tracker of the *Praat* tool [63], to each segment. The *Praat* tracker uses the *Burg* algorithm for formant extraction [64]
4. Generate the *GroundTruth* of the second formant tracks by manually inspecting the spectrograms of the one hundred ninety two spoken segments.
5. Compute the quality measure of the formant trajectories extracted from each diphone, synthetic and natural, by comparing them with the *GroundTruth*.

The quality measure is defined as follows. First we compute the three errors described in (4), (5) and (6).

$$\epsilon_{F_1}^{BuGT} = \frac{\sum_{n=0}^N (F_1^{Pr}(n) - F_1^{GT}(n))^2}{N} \quad (4)$$

$$\epsilon_{F_2}^{BuGT} = \frac{\sum_{n=0}^N (F_2^{Pr}(n) - F_2^{GT}(n))^2}{N} \quad (5)$$

$$\epsilon_{F_3}^{BuGT} = \frac{\sum_{n=0}^N (F_3^{Pr}(n) - F_3^{GT}(n))^2}{N}, \quad (6)$$

where *Pr* stands for *Praat* and *GT* for *GroundTruth*, *N* is the length of the alignment transitions, which is the number of tested frequency values, $F_1^{Pr}, F_2^{Pr}, F_3^{Pr}$ are the first, second and third formant obtained by *Praat* and $F_1^{GT}, F_2^{GT}, F_3^{GT}$ are the first, second and third formant of the *GroundTruth*. These three errors describe the mean squared deviations of the formant trajectories computed with the *Burg* algorithm from the *GroundTruth*. Likewise, other three error measures derive from the comparison of the formant trajectories obtained with the proposed *Fuzzy – Genetic* algorithm from the *GroundTruth*.

$$\epsilon_{F_1}^{Fg-GT} = \frac{\sum_{n=0}^N (F_1^{Fg}(n) - F_1^{GT}(n))^2}{N} \quad (7)$$

$$\epsilon_{F2}^{Fg-GT} = \frac{\sum_{n=0}^N (F_2^{Fg}(n) - F_2^{GT}(n))^2}{N} \tag{8}$$

$$\epsilon_{F3}^{Fg-GT} = \frac{\sum_{n=0}^N (F_3^{Fg}(n) - F_3^{GT}(n))^2}{N}, \tag{9}$$

where *Fg* stands for *FuzzyGenetic* and *GT* for *GroundTruth*.

Each set of three errors must be then combined to obtain the final Mean Squared Errors with weighted sums:

$$MSE^{BuGT} = \sqrt{w_{11}\epsilon_{F1} + w_{21}\epsilon_{F2} + w_{31}\epsilon_{F3}} \tag{10}$$

$$MSE^{Fg-GT} = \sqrt{w_{12}\epsilon_{F1} + w_{22}\epsilon_{F2} + w_{32}\epsilon_{F3}}. \tag{11}$$

In (10) and (11) the indexes $(\cdot)^{BuGT}$ and $(\cdot)^{Fg-GT}$, are omitted for simplicity. Of course it should be clear that the ϵ^{BuGT} must be used for computing MSE^{BuGT} and the ϵ^{Fg-GT} must be used for computing MSE^{Fg-GT} . The weights $w_{11}, w_{21}, w_{12}, w_{22}$ are computed such that the variances of the weighted sums is minimized [65]. More precisely, calling σ_{F1}^2 the variance of ϵ_{F1}^{BuGT} , σ_{F2}^2 the variance of ϵ_{F2}^{BuGT} and σ_{F3}^2 the variance of ϵ_{F3}^{BuGT} (note that we omit again the indexes $(\cdot)^{BuGT}$ for simplicity), and under the assumption of statistical independence among the three errors, the variance of MSE^{BuGT} is minimized with

$$w_{11} = \frac{\sigma_{F1}^{-2}}{\sum_{n=1}^3 \sigma_{Fn}^{-2}},$$

$$w_{21} = \frac{\sigma_{F2}^{-2}}{\sum_{n=1}^3 \sigma_{Fn}^{-2}}$$

and

$$w_{31} = \frac{\sigma_{F3}^{-2}}{\sum_{n=1}^3 \sigma_{Fn}^{-2}}.$$

The same considerations apply to the minum variance of MSE^{Fg-GT} .

The goal of this comparison is to measure the deviation of *FuzzyGenetic* algorithm from the *Burg* algorithm. This comparison is reported in Table 1. In this table, the pair of numbers corresponding to each transition represent the average deviation of *FuzzyGenetic* algorithm from the *Burg* algorithm respectively.

Table 1. Mean squared errors MSE^{FG-GT} of *Fuzzy – Genetic* and the mean squared error MSE^{Bu-GT} , of *Burg* where *Bu* stands for *Burg* and *FG* for *Fuzzy – Genetic* algorithms.

	VO	PL	FR	AF	LI	NA	Sil
	FG Bu	FG Bu	FG Bu	FG Bu	FG Bu	FG Bu	FG Bu
VO	630 844	504 610	432 645	605 505	744 634	560 450	334 234
PL	217 196		888 655	1032 670	976 546		
FR	249 150	945 754					
AF	855 745						
LI	640 780	1020 569	923 866				
NA	670 567	1200 754	234 378				
Sil	155 96	167 223				389 222	

From Table 1 we see the *Fuzzy – Genetic* algorithm give costantly very similar results to the *GroundTruth*.

In the following we report some graphical examples of *Burg*, *FuzzyGenetic* and *GroundTruth* PL – VO transitions, precisely the /ti/, /pi/, /pi/ and /pa/ diphones, with spectrograms and second formant tracks.

Another test is performed on the consistency of the C – V transitions with the *locus; theory* of formant transitions [66,67]. According to the locus theory, the second-formant trajectory of one consonant towards vowels, by extrapolation, points back in time to the same frequency, called the *locus* of that consonant. This is a perceptual invariant of each consonant’s transition toward the vowels, that must be necessary for correct perception of the consonants. A comparison of the *locus* estimated from the spectrograms and from the *Fuzzy – Genetic* algorithm for the plosive phonemes /p t q d b k g/ toward all the vowels results in a 100% accuracy.

An example is shown in Figure 14 related to the second-formant transitions of the phonemes /pa/, /pi/. As shown in Figure 14 the obtained curves are consistent with the trajectories derived from spectrograms. In particular, we can estimate from this Figure that the second frequency locus for the transitions, is approximately 500 Hz, a value perfectly consistent with the literature (e.g., Reference [57]).

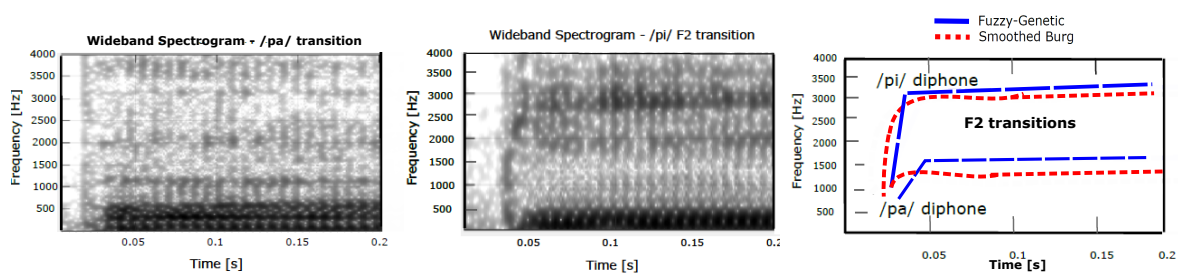


Figure 14. Second formant transitions of /pi/ /pa/ diphones.

As a further example, let’s consider the following transitions: /ti/, /ta/. The trend of the second formant determined by the algorithm is shown in Figure 15, and corresponds to the spectrograms. In this case the locus is about 2000 Hz, which is again consistent with the literature (e.g., Reference [57]).

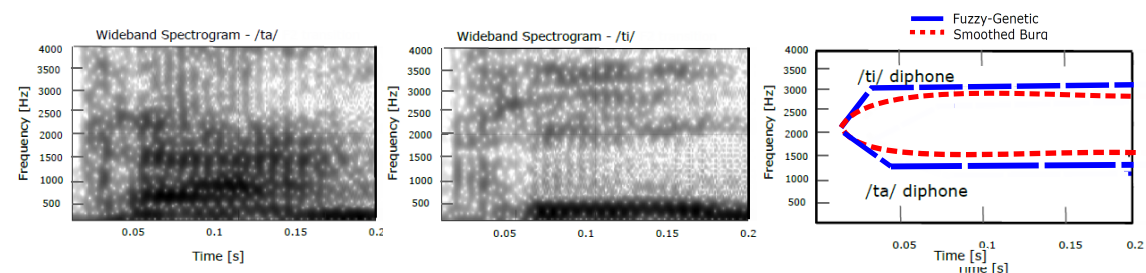


Figure 15. Second formant transitions of the /ti/, /ta/ diphones.

Using the /ti/, /ta/ diphones, we construct by concatenation the word /patita/. To verify the validity of the generation, we compare the spectrograms and the generated second formant (F2) trajectory of spoken and artificial words respectively. At the top of the Figure 16 we show the spectrogram of the word /patita/ spoken by the human tutor. At the bottom we show the spectrogram of the same word artificially generated.

The two spectrograms are different in timing but similar from a frequency point of view. This similarity confirm that the articulator parameters are generated correctly. We therefore can expect that facial movements are properly generated.

In conclusion, all the performed objective tests give very good results, and we now show some results of subjective tests.

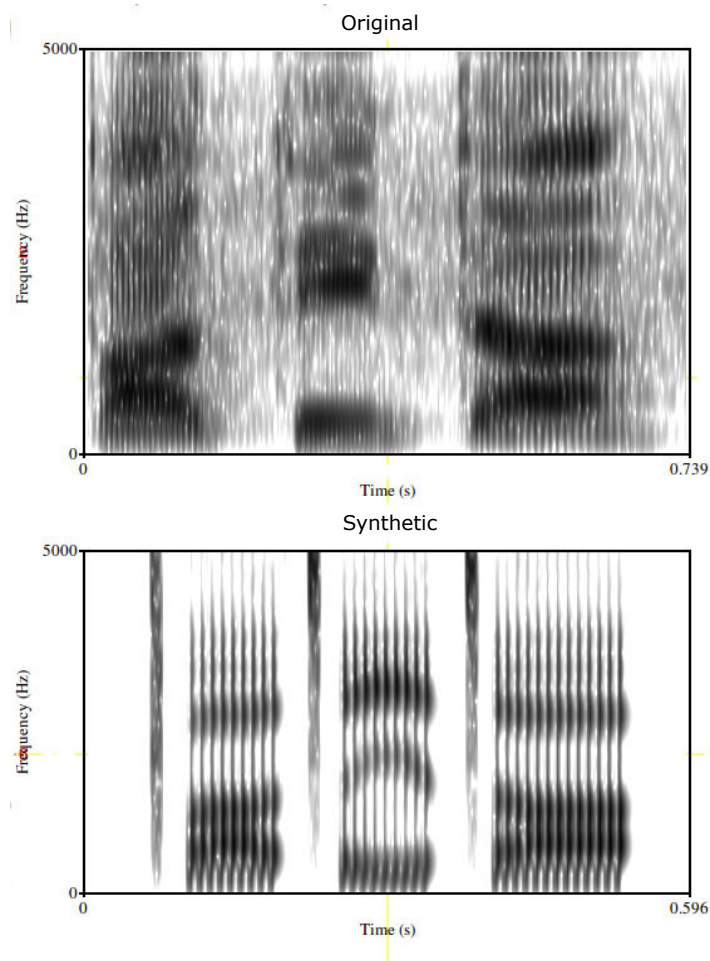


Figure 16. Spectrograms of the original (top) and artificial (bottom) word /p a t i t a/.

9.3. Experiment 2: Subjective Evaluations

A measure of the quality of speech cannot be given exclusively through objective measures as peculiar characteristics such as intelligibility, naturalness, fluidity and intonation can only be assessed from a human listener through subjective measures.

Naturalness describes how close synthesized speech is to human speech and is usually measured by a mean opinion score (MOS) test [68]. In a MOS test, subjects listen to speech and rate its relative perceived quality on some kind of a scale, for example, 5 = *excellent*, 4 = *good*, 3 = *fair*, 2 = *poor*, 1 = *bad*. Then the scores are averaged across subjects.

However, “naturalness” may be difficult to judge in any type of speech. More recent authors, like Nusbaum [69], suggest that *naturalness* be measured by presenting to a group of listener a couple of sentences, one spoken by humans and the other produced by an algorithm, in our case the *FuzzyGenetic*. It is important to note that human speech is low pass filtered in order to make it quite worse than the original. Each listener must then decide whether the speech is produced by a human or a synthesizer. We used a much simpler criterion, that is, to say whether the naturalness of the heard word is *acceptable* (POSITIVE) or not (NEGATIVE). Fifty listeners listened to a series of artificial words giving their degree of naturalness, comprehension and synchronization. The generated words include 120 meaningless words and 120 words with meaning. Moreover, all these words are also reproduced by a talking head animated with the facial movements obtained with the *Fuzzy – Genetic* algorithm. As a comparison, we considered the DECTalk speech synthesizer [70]. Naturalness data of DECTalk is taken from Reference [71]. Recall that we assume that the visual synthesis is performed with the algorithm described in Reference [60].

In the following we give a very short explanation on how the results on audio-visual synthesis are obtained. From Reference [71], the DECTalk subjective measure for audio nonsense words of 47% corresponds, using the data in Reference [58], to an Equivalent Noise Level of about -13 dB. Adding visual information to the synthetic speech the subjective measure increases to 95% but, considering the performance of the visual system, this means an expected average subjective score of 77.08%. Likewise, a subjective score of 57% for words with meaning synthesized with DECTalk leads to an expected improvement to 79.96% with the audio-visual system.

We show in the following the results of intelligibility test on the *Fuzzy – Genetic* synthetic speech with comparison with another system. Intelligibility of synthetic voices can be measured with one of the many available tests; if it is not possible to speak about standardized evaluation tests, it can only be said that some are more frequently used than others. The basic test is clearly to compare the synthetic word which was heard by the listeners participating to the test with the word actually synthesized to see if it is correct. However, many other more complex tests are available. Going back to the early works in this area, in Reference [72] the *Modified; RhymeTest* (MRT) was proposed. MRT is based on the use of a rhyming words list that differ for the initial or final consonant; subjects must identify the right word within the set used as a stimulus and its output is the rate of success. MRT is a *closed* test which allows to eliminate uncontrolled variations and it is very useful to test the intelligibility of particular aspect of the synthetic speech, as the initial or final CV transitions.

The Diagnostic Rhyme Test (DRT), described originally in Reference [73] is a test for the evaluation of intelligibility based on the choice between two words whose initial consonants differ in a single distinctive trait, namely just a phoneme. During the test the subjects must determine which of the two words were pronounced. The subjective results are then given on a 1 to 5 scale.

However, as for the naturalness, we adopted a much simpler criterion, which is to establish whether the Comprehension of the word is *acceptable* or it is too difficult. Comparison is made with the Votrax speech synthesizer [74] and we again assume that visual synthesis is performed with the algorithm described in Reference [60].

A further subjective test concerns the synchronization between synthetic speech and facial movements. While we tested synchronization of face movements with speech synthesized with *Fuzzy – Genetic* algorithm, comparison was made with spoken speech degraded by speech-shaped noise added to the spoken speech. The SNR of noise was -18 and -20 dB as described in Reference [60]. However, while audiovisual results reported in Reference [60] are related to the Viseme-Driven algorithm our results, they obtain *Naturalness; Rating* and not synchronization as we do. However, we assume that synchronization quality can be also expressed in terms of naturalness of the audiovisual information delivered to the human. The subjective results are described later. Note, however, that in Reference [60] naturalness results with nonsense words are not available, and therefore we indicate NA (Not Available) for them.

In conclusion of this section, the tests carried out concern the measurement of the fundamental subjective characteristics of the artificial voice generated by the *Fuzzy – Genetic* algorithm such as: naturalness and comprehension of artificial words. Moreover, above all, the quality of facial movements, such as the perceived quality of the synchronization between artificial voice and facial movements. Naturalness and understanding of artificial voice have been measured with and without facial movements. Naturalness is reported in Table 2. Comprehension is shown in Table 3; of course, it emerges that comprehension of the artificial words is greater for words with meaning. It also emerges from the measurements that the facial movements helps in the understanding of artificial words. The evaluation of the synchronization of facial movements with the artificial voice is shown in Table 4. From this evaluation it emerges that synchronization is best perceived in meaningless words than in meaningful ones, probably because facial movements are appreciated as an aid in trying to understand the words. In any case, all the objective and subjective results are very close to that obtained with existing systems and thus we are quite sure that the algorithm produces audiovisual information of high quality.

Table 2. Naturalness results of synthetic utterances for nonsense words and words with meaning.

	Nonsense Words				Words with Meaning			
	Only Audio		Audio + Video		Only Audio		Audio + Video	
	FG DECTalk	FG DECTalk	FG DECTalk	FG DECTalk	FG DECTalk	FG DECTalk	FG DECTalk	FG DECTalk
POSITIVE	49.0%	47.0%	61.5%	77.08%	71.5%	57.8%	80.0%	79.96%
NEGATIVE	51.0%	53.0%	38.5%	29.7%	28.5%	42.2	20.0%	20.04%

Table 3. Comprehension results: nonsense words vs. words with meaning.

	Nonsense Words				Words with Meaning			
	Only Audio		Audio + Video		Only Audio		Audio + Video	
	FG Votrax	FG Votrax	FG Votrax	FG Votrax	FG Votrax	FG Votrax	FG Votrax	FG Votrax
POSITIVE	58.0%	57.2%	66.0%	73%	88.0%	83.8%	90.5%	89.16%
NEGATIVE	42.0%	42.8%	34.0%		12.0%	17.2%	9.5%	10.84%

Table 4. Speech/facial movements synchronization.

	Nonsense Words	Words With Meaning
POSITIVE	91% NA	83.16% 84%
NEGATIVE	8.5% NA	16.84% 16.0%

9.4. Experiment 3: Basic Speech Unit Extraction

Some examples of extraction of Basic Speech Units (BSU) are shown below. Suppose the human operator pronounces the training words / atʃa /, / ala /. The training words are divided into the following Basic Speech Units: / atʃa / = [-a] [atʃ] [tʃa] [a-]; / ala / = [-a] [al] [la] [a-]. Therefore, the articulatory characteristics of the following six BSU are stored in the knowledge base: [-a], [a-], [atʃ], [tʃa], [al], [la].

The values of the acoustic parameters ($AV, AF, F1, F2, F3, B1, B2, B3$), described in terms of $[I(\cdot), D(\cdot), F(\cdot), L_1(\cdot), L_2(\cdot)]$, obtained from the fuzzy/genetic algorithm for the BSU / atʃa /, are shown in Table 5. The same table shows the acoustic parameters of the following BSUs: [-a], [atʃ], [tʃa], [a-].

The values of degrees of membership of the articulator parameters estimated by the algorithm for the same BSUs are reported in Table 6.

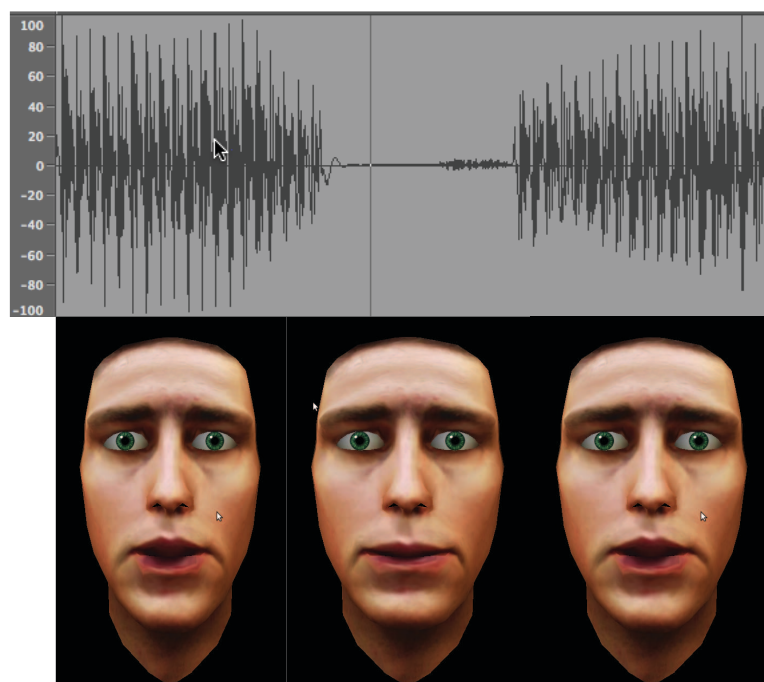
Table 5. Estimated acoustic parameters per transition.

	AV	AH	AF	F1	F2	F3	B1	B2	B3
[-a]	D			65.00	65.00	65.00	65.00	65.00	65.00
	L	52.00		695.47	1197.87	2609.58	142.50	152.50	147.50
	I	65.00							
	F	130.00							
[atʃ]	D			45.00	45.00	45.00			
	L	-4.67	40.00	246.89	1775.00	1892.50			
	I	25.00	65.00						
	F			65.00	65.00	65.00			
[tʃa]	D			25.00	25.00	25.00	65.00	65.00	65.00
	L	52.00	-4.67	695.47	1197.87	2609.58	142.50	152.50	147.50
	I								
	F			45.0	45.0	45.0			
[a-]	D			65.00	65.00	65.00	65.00	65.00	65.00
	L	-4.67		499.29	1541.19	2445.85	142.50	152.50	147.50
	I	65.00		65.00	65.00	65.00	65.00	65.00	65.00
	F	65.00							

Table 6. Estimated articulatory parameters per transition.

	QUA	TON	APE	ANT	SON	BIL	LAB	ALV	PRE	PAL	VIB	D-V	DEN	VEL
initial	0.50	0.40	0.45	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[-a]	0.178	0.990	0.022	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
[atʃ]	0.000	0.000	0.000	0.910	0.000	0.000	0.971	0.009	0.000	0.000	0.000	0.000	0.000	0.000
[tʃa]	0.276	0.855	0.098	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
[a-]	0.50	0.40	0.45	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

As a further example, in Figure 17 we provide three snapshots of the facial movements generated by the Italian word /acia/, (atʃa /), generated artificially. To highlight the synchronization of the facial movements with the artificial word, the signal waveform is displayed in the panel above the facial movements.

**Figure 17.** Synthetic speech: /atʃa/.

9.5. Experiment 4: Some Examples Produced for the Italian Language

In this Section, some experimental analysis of the Italian word /nove/ (nine) are reported as an example. The analysis is performed with the fuzzy-genetic algorithm with the following parameters: population size of 200 elements; mutation rate equal to 0.02. Several algorithms for computing the parameters of the Klatt synthesizer have been proposed in the past. Many of them are presented as solutions to the ‘copy synthesis’ problem, which is the problem to estimate the input parameters to reconstruct a speech signal using a speech synthesizer. Copy synthesis is a difficult inverse problem because the mapping is non-linear and often is a ‘from many to one’ problem. One of them is that proposed by Kasparaitis [75], who proposed an iterative algorithm for the automatic estimation of the factors for the Klatt model using the corpus of an annotated audio record of the speaker. Another is that proposed by Laprie, [76], who describe an approach to track formant trajectories first, and to compute the amplitudes of the resonators by an algorithm derived from cepstral smoothing they called “true envelope”.

In Reference [77], a framework for automatically extracting the input parameters of a class of formant-based synthesizers is described. The framework is based on a genetic algorithm. Also Borges et al., in Reference [78], describe a system based on GA optimization for automatically computing the

parameters for the Klatt synthesizer. In Reference [79] it is presented *KlattWork* which relies on the 1980 version of the Klatt synthesizer. Rather than a copy synthesis system, *KlattWorks* is a “manager”, and is designed to allow the user to rapidly develop new synthetic speech for experiments using existing tools.

Weenink developed a class called *KlattGrid*, which implements a Klatt-type synthesizer. Parameters are automatically computed with the standard speech signal processing routines provided by the Praat program [80]. For simplicity, the algorithm *KlattGrid* provided by Praat for computing the Klatt synthesis system has been used for comparison with the *Fuzzy – Genetic* algorithm,

In Figure 18 we report the trajectories of the first three formant frequencies and of the *AF* and *AV* amplitudes obtained with the algorithm from the word *nove* (nine) in upper and lower panels respectively. The centers of the three phonemes of the word, that is */n/*, */o/*, */v/*, */e/*, are indicated by the three vertical lines.

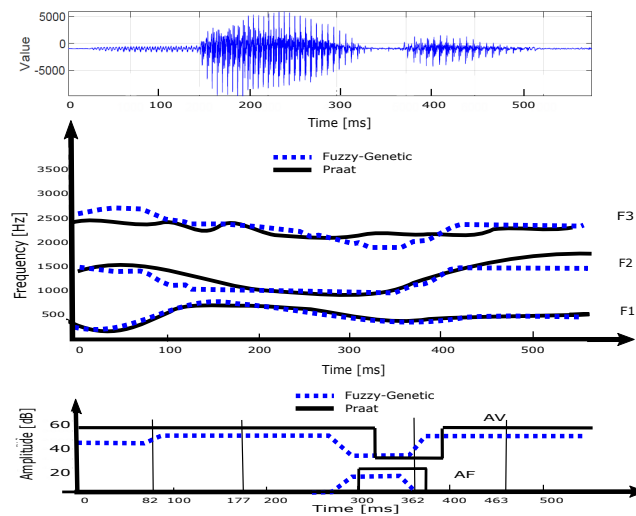


Figure 18. Acoustic analysis of the Italian word “nove”.

In Figure 19 the dynamic of the membership degrees of the articulatory places of articulation is reported.

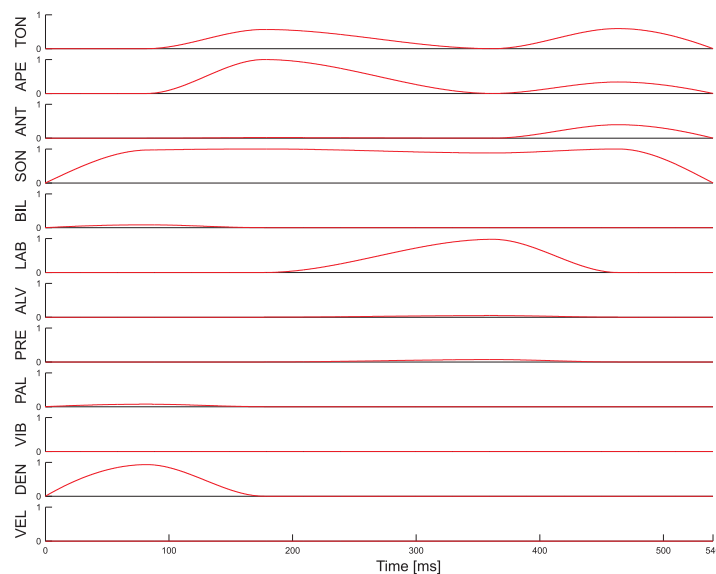


Figure 19. Articulatory places of articulation of the Italian word “nove”.

Some subjective evaluation results of a phonetic listening test are reported in Table 7. In this test, subjects must identify the phonetic category they perceive. It comes out that the phonetic categories are quite well perceived.

Table 7. Subjective evaluation results.

Phonetic Categories	Number of Signals	Exact Recognitions [%]
plosive	193	96
fricative	105	86
affricate	47	87
liquid	83	88
Total	428	89

10. Final Remarks and Conclusions

In this paper we describe an algorithm for automatic acquisition of human vocalization. The article is particular relevant in the context of advanced human-machine interfaces, with particular emphasis on the emerging big data trend. The algorithm is divided in two part: training phase and synthesis phase. In the training phase, the algorithm automatically gathers the articulatory characteristics of a voice uttered by a human tutor and given at input. In the synthesis phase, the articular knowledge learned in the training phase is used to describe an arbitrary input text from an articulatory point of view. It thus produces the artificial voice corresponding to the input text by articulatory synthesis on the one hand and, the facial movements synchronized with the artificial voice on the other hand. We use facial movements to animate a virtual talking head, drawn on a computer screen. Our approach allows to apply these facial movements to animate virtual speaking heads (Avatar) with the goal of achieving high-quality human-machine interfaces.

Our algorithm is based on a genetic optimization algorithm and a set of fuzzy rules to determine the degrees of membership of the places of articulation. An interesting feature of fuzzy rules, as it is well known, is that they can be easily edited and fine-tuned. The main features of the algorithm are the description of any input text in articulatory form. It is then possible to generate artificial speech by articulatory synthesis (Text to Speech Synthesis). In addition, by linear/nonlinear mapping between articulator and facial parameters, the facial configurations of an human speaker who reads the input text.

Many experimental results obtained with the algorithm are reported in the paper to highlight the quality of our work. Indeed, experimental results show that high quality facial movements synchronized with artificial speech are obtained. voice have a good degree of acceptance in subjective tests.

The most important future developments of our research are double-fold: from a side, we aim at extending our algorithm to other languages; from another side, we aim at further specializing the overall framework to emerging big data features (e.g., References [81–90]).

Author Contributions: Conceptualization, A.C. and E.M.; methodology, A.C.; software, E.M.; validation, A.C. and G.M.G.; formal analysis, E.M.; investigation, A.C.; resources, G.M.G.; data curation, E.M.; writing—original draft preparation, E.M.; writing—review and editing, A.C.; visualization, G.M.G.; supervision, A.C.; project administration, E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zikopoulos, P.; Eaton, C. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*; McGraw-Hill Osborne Media: New York, NY, USA, 2011.
2. McAfee, A.; Brynjolfsson, E. *Big Data: The Management Revolution*; Harvard Business Review: Boston, MA, USA, 2012.
3. Cuzzocrea, A.; Song, I.-Y.; Davis, K.C. Analytics over large-scale multidimensional data: The big data revolution! In Proceedings of the 2011 International Workshop on Data Warehousing and OLAP, Glasgow, UK, 28 October 2011; pp. 101–104.
4. Cuzzocrea, A.; Saccá, D.; Ullman, J.D. Big data: A research agenda. In Proceedings of the 17th International Database Engineering & Applications Symposium, Barcelona, Spain, 9–13 October 2013; pp. 198–203.
5. Cuzzocrea, A.; Song, I.-Y.; Bellatreche, L. Data warehousing and OLAP over big data: Current challenges and future research directions. In Proceedings of the ACM 16th International Workshop on Data Warehousing and Online Analytical Processing (DOLAP), San Francisco, CA, USA, 28 October 2013; pp. 67–70.
6. San Ang, P.; Fan, L.Y.; Tham, M.Y.; Tan, S.H.; Soh, S.B.; Foo, B.P.; Loke, C.W.; Hu, S.; Sung, C. Towards Human-Machine Collaboration in Creating an Evaluation Corpus for Adverse Drug Events in Discharge Summaries of Electronic Medical Records. *Big Data Res.* **2016**, *4*, 37–43. [[CrossRef](#)]
7. Ofli, F.; Meier, P.; Imran, M.; Castillo, C.; Tuia, D.; Rey, N.; Briant, J.; Millet, P.; Reinhard, F.; Parkan, M.; et al. Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response. *Big Data* **2016**, *4*, 47–59. [[CrossRef](#)] [[PubMed](#)]
8. Weber, J. A Multi-user-collaboration Platform Concept for Managing Simulation-Based Optimization of Virtual Tooling as Big Data Exchange Service—An Implementation as Proof of Concept Based on Different Human-Machine-Interfaces. In Proceedings of the 7th EAI International Conference on Big Data Technologies and Applications, Seoul, Korea, 17–18 November 2016; pp. 144–153.
9. Cuzzocrea, A.; Song, I.-Y. Big Graph Analytics: The State of the Art and Future Research Agenda. In Proceedings of the 17th International Workshop on Data Warehousing and Online Analytical Processing (DOLAP), Shanghai, China, 3–7 November 2014; pp. 99–101.
10. Cuzzocrea, A. Aggregation and multidimensional analysis of big data for large-scale scientific applications: models, issues, analytics, and beyond. In Proceedings of the 27th International Conference on Scientific and Statistical Database Management, La Jolla, CA, USA, 29 June–1 July 2015; pp. 23–28.
11. Russom, P. *Big Data Analytics*; TDWI Best Practices Report; 4th Quarter: Renton, WA, USA, 2011.
12. Lavalley, S.; Lesser, E.; Shockley, R.; Hopkins, M.; Kruschwitz, N. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Manag. Rev.* **2011**, *52*, 21–32.
13. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
14. Zeng, Q.; Jiang, B.; Duan, Q. Integrated evaluation of hardware and software interfaces for automotive human-machine interaction. *IET Cyber. Phys. Syst. Theory Appl.* **2019**, *4*, 214–220. [[CrossRef](#)]
15. Kim, M.; Cho, J.; Lee, S.; Jung, Y. IMU Sensor-Based Hand Gesture Recognition for Human-Machine Interfaces. *Sensors* **2019**, *19*, 3827. [[CrossRef](#)] [[PubMed](#)]
16. Lim, Y.; Ramasamy, S.; Gardi, A.; Kistan, T.; Sabatini, R. Cognitive Human-Machine Interfaces and Interactions for Unmanned Aircraft. *J. Intell. Robot. Syst.* **2018**, *91*, 755–774. [[CrossRef](#)]
17. Estrany, B.; Marin, C.; Mascaró, M.; Bibiloni, A.; Luo, Y. Multimodal human-machine interface devices in the cloud. *J. Multimod. User Interfaces* **2018**, *12*, 125–143. [[CrossRef](#)]
18. Brooks, R.A.; Breazeal, C.; Marjanović, M.; Scassellati, B.; Williamson, M.M. The Cog Project: Building a Humanoid Robot. In *Computation for Metaphors, Analogy, and Agents*; Nehaniv, C.L., Ed.; Springer: Berlin/Heidelberg, Germany, 1999; pp. 52–87.
19. Choi, C.; Kong, D.; Kim, J.; Bang, S. Speech Enhancement and Recognition Using Circular Microphone Array For Service Robotics. In Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453), Las Vegas, NV, USA, 27 October–1 November 2003; pp. 3516–3521.

20. Pelachaud, C.; Badler, N.I.; Steedman, M. Generating Facial Expressions for Speech. *Cognit. Sci.* **1996**, *20*, 1–46. [[CrossRef](#)]
21. Garg, S.; Hamarneh, G.; Jongman, A.; Sereno, J.A.; Wang, Y. Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Commun.* **2019**, *113*, 47–62. [[CrossRef](#)]
22. Meng, Z.; Han, S.; Liu, P.; Tong, Y. Improving Speech Related Facial Action Unit Recognition by Audiovisual Information Fusion. *IEEE Trans. Cybern.* **2019**, *49*, 3293–3306. [[CrossRef](#)]
23. Nishikawa, K.; Takanobu, H.; Mochida, T.; Honda, M.; Takanishi, A. Modeling and Analysis of Elastic Tongue Mechanism of Talking Robot for Acoustic Simulation. In Proceedings of the 2003 IEEE International Conference on Robotics and Automation, ICRA 2003, Taipei, Taiwan, 14–19 September 2003; pp. 2107–2114.
24. Yehia, H.; Rubin, P.; Vatikiotis-Bateson, E. Quantitative association of vocal-tract and facial behavior. *Speech Commun.* **1998**, *26*, 23–43. [[CrossRef](#)]
25. Vatikiotis-Bateson, E.; Kroos, C.; Munhall, K.G.; Pitermann, M. Task Constraints on Robot Realism: The Case of Talking Heads. In Proceedings of the 9th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2000, Osaka, Japan, 27–29 September 2000; pp. 352–357.
26. Nishikawa, K.; Takanobu, H.; Mochida, T.; Honda, M.; Takanishi, A. Speech Production of an Advanced Talking Robot based on Human Acoustic Theory. In Proceedings of the 2004 IEEE International Conference on Robotics and Automation—IEEE ICRA, New Orleans, LA, USA, 26 April–1 May 2004; pp. 3213–3219.
27. Lotto, A.J.; Hickok, G.S.; Holt, L.L. Reflections on mirror neurons and speech perception. *Trends Cognit. Sci.* **2009**, *13*, 110–114. [[CrossRef](#)] [[PubMed](#)]
28. Imada, T.; Zhang, Y.; Cheour, M.; Taulu, S.; Ahonen, A.; Kuhl, P.K. Infant speech perception activates Broca’s area: A developmental magnetoencephalography study. *Neuroreport* **2006**, *17*, 957–962. [[CrossRef](#)] [[PubMed](#)]
29. Fukui, K.; Ishikawa, Y.; Ohno, K.; Sakakibara, N.; Honda, M.; Takanishi, A. Three dimensional tongue with liquid sealing mechanism for improving resonance on an anthropomorphic talking robot. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 11–15 October 2009; pp. 5456–5462.
30. Fogassi, L.; Ferrari, P.F. Mirror Neurons and the Evolution of Embodied Language. *Curr. Dir. Psycholog. Sci.* **2007**, *16*, 136–142. [[CrossRef](#)]
31. Mumolo, M.; Abbattista, G. High Quality Real-Time Text-to-Speech System for Italian Language. In Proceedings of the 1990 VERBA, Rome, Italy, 8–9 November 1990, pp. 50–59.
32. Mumolo, E.; Nolich, M. Towards articulatory Control of Talking Heads in Humanoid Robotics Using a Genetic-Fuzzy Imitation Learning Algorithm. *Int. J. Human. Robot.* **2007**, *4*, 151–179. [[CrossRef](#)]
33. Jiang, J.; Alwan, A.; Bernstein, L.E.; Keating, P.A.; Auer, E.T. On the correlation between facial movements, tongue movements and speech acoustics. In Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000; pp. 42–45.
34. Lyakh, G.S. Imitation of articulatory movements and of sound production in early infancy. *Neurosci. Trans.* **1968**, *2*, 913–917. [[CrossRef](#)]
35. Shiomi, M.; Kanda, T.; Miralles, N.; Miyashita, T.; Fasel, I.; Movellan, J.; Ishiguro, H. Face-to-face interactive humanoid robot. In Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sendai, Japan, 28 September–2 October 2004; pp. 1340–1346.
36. Nishikawa, K.; Imai, A.; Ogawara, T.; Takanobu, H.; Mochida, T.; Takanishi, A. Speech Planning of an Anthropomorphic Talking Robot for Consonant Sounds Production. In Proceedings of the 2002 IEEE International Conference on Robotics and Automation, Washington, DC, USA, 11–15 May 2002; pp. 1830–1835.
37. Nishikawa, K.; Takanobu, H.; Mochida, T.; Honda, M.; Takanishi, A. Development of a New Human-like Talking Robot Having Advanced Vocal Tract Mechanisms. In Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), Las Vegas, NV, USA, 27 October–1 November 2003; pp. 1907–1013.
38. Higashimoto, T.; Sawada, H. 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, Washington, DC, USA, 11–15 May 2002; pp. 3858–3863.

39. Kanda, H.; Ogata, T.; Takahashi, T.; Komatani, K.; Okuno, H.G. Phoneme acquisition model based on vowel imitation using Recurrent Neural Network. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, MO, USA, 11–15 October 2009; pp. 5388–5393.
40. Sargin, M.E.; Erzin, E.; Yemez, Y.; Tekalp, A.M.; Erdem, A.T.; Erdem, C.E.; Özkan, M.E. Prosody-Driven Head-Gesture Animation. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, USA, 15–20 April 2007; pp. 677–680.
41. Albrecht, I.; Haber, J.; Seidel, H. Automatic Generation of Non-Verbal Facial Expressions from Speech. In *Advances in Modelling, Animation and Rendering*; Vince, J., Earnshaw, R., Eds.; Springer: London, UK, 2002; pp. 283–293.
42. Salvi, G.; Beskow, J.; Al Moubayed, S.; Granström, B. SynFace–Speech-Driven Facial Animation for Virtual Speech-Reading Support. *EURASIP J. Audio Speech Music Process.* **2009**, *1*, 177:1–177:10. [[CrossRef](#)]
43. Zoric, G.; Smid, S.; Pandzic, I.S. Towards Facial Gestures Generation by Speech Signal Analysis Using HUGE Architecture. In *Multimodal Signals: Cognitive and Algorithmic Issues*; Esposito A., Hussain A., Marinaro, M., Martone, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 112–120.
44. International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*; Cambridge University Press: Cambridge, UK, 1999.
45. Mumolo, E.; Nolich, M.; Menegatti, E. A genetic-fuzzy algorithm for the articulatory imitation of facial movements during vocalization of a humanoid robot. In Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2005, Tsukuba, Japan, 5–7 December 2005; pp. 436–441.
46. Allen, J.; Sharon Hunnicutt, M.; Klatt, D. *From Text to Speech: The MITalk System*; Cambridge University Press: Cambridge, UK, 1987.
47. Stella, M.; Charpentier, F. Diphone synthesis using multipulse coding and a phase vocoder. In Proceedings of the 1985 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Tampa, FL, USA, 26–29 March 1985; pp. 740–743.
48. Gussenhoven, C.; Jacobs, H. *Understanding Phonology (Understanding Language)*, 3rd ed.; Hodder Education Publishers: London, UK, 2011.
49. Souček, P.; Slavata, O.; Holub, J. New approach in subjective and objective speech transmission quality measurement in TCP/IP networks. *J. Phys. Conf. Ser.* **2015**, *588*, 12–20. [[CrossRef](#)]
50. Sakoe, H.; Chiba, S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. In *Readings in Speech Recognition*; Waibel, A., Lee, K.-F., Eds.; Morgan Kaufmann Publisher: Burlington, MA, USA, 1990; pp. 159–165.
51. Rabiner, L.R.; Juang, B. *Fundamentals of Speech Recognition*; Prentice Hall: Upper Saddle River, NJ, USA, 1993.
52. McCree, A.; Barnwell, T.P. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 242–250. [[CrossRef](#)]
53. Wrench, A. The MOCHA-TIMIT Articulatory Database. Available online: <http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html> (accessed on 15 September 2019).
54. Moro, A.; Mumolo, E.; Nolich, M. Automatic 3D Virtual Cloning of a Speaking Human Face. In Proceedings of the 2010 ACM Symposium on Applied Computing, Florence, Italy, 25–29 October 2010; pp. 45–50.
55. Parke, F.I.; Waters, K. *Computer Facial Animation*; AK Peters/CRC Press: Boca Raton, FL, USA, 2008.
56. Yang, W.; Dixon, M.; Yantorno, R. A modified bark spectral distortion measure which uses noise masking threshold. In Proceedings of the 1997 IEEE Workshop on Speech Coding for Telecommunications, Pocono Manor, PA, USA, 7–10 September 1997; pp. 55–56.
57. Jintao, J.; Abeer, A.; Keating, P.A.; Auer, E.T., Jr.; Bernstein, L.E. On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics. *EURASIP J. Appl. Signal Process.* **2002**, *11*, 1174–1118.
58. Sumbly, W.H.; Pollack, I. Visual Contribution to Speech Intelligibility in Noise. *J. Acoust. Soc. Am.* **1954**, *26*, 212–215. [[CrossRef](#)]
59. Mattheyses, W.; Verhelst, W. Audio-visual speech synthesis: An overview of the state of the art. *Speech Commun.* **2015**, *66*, 182–217. [[CrossRef](#)]
60. Dey, P.; Maddock, S.C.; Nicolson, R. Evaluation of A Viseme-Driven Talking Head. In Proceedings of the EG UK Theory and Practice of Computer Graphics 2010, Sheffield, UK, 6–8 September 2010; pp. 139–442.
61. Lewis, J.P.; Parke, F.I. Automated lip-synch and speech synthesis for character animation. *SIGCHI Bull.* **1987**, *17*, 143–147. [[CrossRef](#)]

62. Sergio P.; Oliveira, L.C. DTW-based Phonetic Alignment Using Multiple Acoustic Features. In Proceedings of the 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003; pp. 309–312.
63. Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer—Version 6.0.23. Available online: <http://www.praat.org/> (accessed on 15 September 2019).
64. Childers, D.G. *Modern Spectrum Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1978.
65. Shahar, D.J. Minimizing the Variance of a Weighted Average. *Open J. Stat.* **2017**, *7*, 216–224. [[CrossRef](#)]
66. Harvey, M.S.; Bessell, N.; Dalston, E.; Majors, T. An investigation of stop place of articulation as a function of syllable position: A locus equation perspective. *J. Acoust. Soc. Am.* **1997**, *101*, 2826–2838.
67. Sussman, H.; Hoemeke, K.A.; Ahmed, F. A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *J. Acoust. Soc. Am.* **1993**, *94*, 1256–1268. [[CrossRef](#)]
68. International Telecommunications Union—ITU-T Recommendation P.85 1994. Telephone Transmission Quality Subjective Opinion Tests—A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. Available online: <http://www.itu.int/rec/T-REC-P.85-199406-1/en> (accessed on 15 September 2019).
69. Nusbaum, H.C.; Francis, A.L.; Henly, A.S. Measuring the naturalness of synthetic speech. *Int. J. Speech Technol.* **1995**, *2*, 7–19. [[CrossRef](#)]
70. Klatt, D. How Klattalk became DECTalk: An Academic’s Experiences in the Business World. In Proceedings of the Official Proceedings of Speech Tech’87 : Voice Input/Output Applications Show and Conference, New York, NY, USA, 28–30 April 1987; pp. 293–294.
71. Thomas, P.C.R.; Gilson, M.H.; Kincaid, R.D.; Peter, J. Linguistic cues and memory for synthetic and natural speech. *Hum. Fact.* **2000**, *42*, 421–431.
72. Fairbanks, G. Test of Phonemic Differentiation: The Rhyme Test. *J. Acoust. Soc. Am.* **1958**, *30*, 596–600. [[CrossRef](#)]
73. House, A.S.; Williams, C.E.; Hecker, M.H.; Kryter, K.D. Articulation Testing Methods: Consonant Differentiation with a Closed Response Set. *J. Acoust. Soc. Am.* **1965**, *37*, 158–166. [[CrossRef](#)]
74. Lee, D. A voice response system for an office information system. In Proceedings of the SIGOA Conference on Office Information Systems 1982, Philadelphia, PA, USA, 21–23 June 1982; pp. 113–121.
75. Kasparaitis, P. Automatic Parameters Estimation of the D.Klatt Phoneme Duration Model. *Inf. Lith. Acad. Sci.* **2016**, *27*, 573–586.
76. Laprie, Y.; Bonneau, A. A copy synthesis method to pilot the Klatt synthesiser. In Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002; p. 4.
77. Figueiredo, A.; Imbiriba, T.; Bruckert, E.; Klautau, A. Automatically Estimating the Input Parameters of Formant-Based Speech Synthesizers. In Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006—4th Workshop in Information and Human Language Technology (TIL’2006), Ribeirão Preto, Brazil, 23–28 October 2006; pp. 1–10.
78. Borges, J.; Couto, I.; Oliveira, F.; Imbiriba, T.; Klautau, A. GASpeech: A Framework for Automatically Estimating Input Parameters of Klatt’s Speech Synthesizer. In Proceedings of the 2008 10th Brazilian Symposium on Neural Networks, Salvador, Bahia, Brazil, 26–30 October 2008; pp. 81–86.
79. McMurray, B. *KlattWork—Version 1.6*; Department of Brain and Cognitive Sciences, University of Rochester: Rochester, NY, USA, 2009.
80. Weenink, D. *The KlattGrid Speech Synthesizer*; Institute of Phonetic Sciences, University of Amsterdam: Amsterdam, The Netherlands, 2009.
81. Cuzzocrea, A.; Russo, V. Privacy Preserving OLAP and OLAP Security. In *Encyclopedia of Data Warehousing and Mining*; IGI Global: Pennsylvania, PA, USA, 2009; pp. 1575–1581.
82. Cuzzocrea, A.; Bertino, E. Privacy Preserving OLAP over Distributed XML Data: A Theoretically-Sound Secure-Multiparty-Computation Approach. *J. Comput. Syst. Sci.* **2011**, *77*, 965–987. [[CrossRef](#)]
83. Cuzzocrea, A. Combining multidimensional user models and knowledge representation and management techniques for making web services knowledge-aware. *Web Intell. Agent Syst.* **2006**, *4*, 289–312.
84. Chatzimilioudis, G.; Cuzzocrea, A.; Gunopulos, D.; Mamoulis, N. A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement. *J. Comput. Syst. Sci.* **2013**, *79*, 349–368. [[CrossRef](#)]

85. Bonifati, A.; Cuzzocrea, A. Storing and retrieving XPath fragments in structured P2P networks. *Data Knowl. Eng.* **2006**, *59*, 247–269. [[CrossRef](#)]
86. Cuzzocrea, A.; De Maio, C.; Fenza, G.; Loia, V.; Parente, M. OLAP analysis of multidimensional tweet streams for supporting advanced analytics. In Proceedings of the SAC 2016—23rd International Conference, Pisa, Italy, 4–8 April 2016; pp. 992–999.
87. Cuzzocrea, A.; Moussa, R.; Xu, G. OLAP*: Effectively and Efficiently Supporting Parallel OLAP over Big Data. In Proceedings of the International Conference on Model and Data Engineering, Amantea, Cosenza, Italy, 25–27 September 2013; pp. 38–49.
88. Cuzzocrea, A.; Wang, W. Approximate range-sum query answering on data cubes with probabilistic guarantees. *J. Intell. Inf. Syst.* **2007**, *28*, 161–197. [[CrossRef](#)]
89. Schuller, B.W. Speech Analysis in the Big Data Era. In Proceedings of the TSD 2015: 18th International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic, 14–17 September 2015; pp. 3–11.
90. Huang, X. Big Data for Speech and Language Processing. In Proceedings of the 2018 IEEE International Conference on Big Dat, Seattle, WA, USA, 10–13 December 2018; p. 2.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).