Improved humanoid vocalization acquisition from a human tutor

Enzo Mumolo and Massimiliano Nolich

Abstract—This paper describes an approach to automatically acquire the vocalization for an humanoid robot by learning from a human tutor. The learned vocalization can be used to multimodal reproduction of speech, based on articulatory and acoustic parameters that compose the vocalization database. The proposed algorithm can synthesize speech utterances from unrestricted text and generate articulatory and facial movements of the humanoid talking face synchronized with the generated speech. The fuzzy articulatory rules are derived from the International Phonetic Alphabet (IPA) to allow simpler adaptation to different languages. Experimental results show a good subjective acceptance of the acquired vocalization in terms of quality, naturalness and synchronization. Although the algorithm has been implemented on a virtual talking face, it could eventually be used also in mechanical vocalization systems.

I. INTRODUCTION

The interest in humanoid robotics is increasing fast. A humanoid is a robot designed to work with humans as well as for them. Humanoid robots have been designed to give better services to the human beings.

Therefore, a fundamental issue in humanoid robotics is the interaction with humans. At MIT the Cog project [1] was developed under the hypothesis that humanoid intelligence requires humanoid interactions with the world.

The simpler way of communication between humans and humanoids is speech. In fact, speech is a natural way of communication between humans in real world. Therefore, it should be also a natural way for communication between humans and humanoids.

Human-humanoid interaction by voice requires that on one hand the humanoid should recognize the human utterances and on the other hand that it should respond to the human using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization. By artificial vocalization we mean that the message delivered to the human is made by using artificial vocalization.

Let us look at our model of verbal/facial communication reported in Fig. 1. Adapting Fig. 1 to an humanoid, the concept formation and message generation modules may be located at the cognitive level of the robot, while the control and phonatory modules are located at the vocalization level.

Some researchers, for example [2][3], faced the vocalization problem from a mechanical point of view. They have developed a mechanical replica of the human vocal tract, of the larynx and of the tongue and are trying to control their movements for producing mechanically generated artificial speech using various open and closed loop strategies. In the closed loop strategy, for example, the extraction of control parameters from speech is optimized by minimizing the distance between original and generated speech. Our work is inspired by the work described in [2][3] in the sense that we extract from the input utterances some parameters that could be used to generate an artificial replica of the real utterances on one hand and to control phonatory organs in order to produce that utterance from the other hand. To this end, we extract articulatory parameters using a fuzzy model of the vocal tract. The related algorithms have been introduced in [4][5]. In these works, the mechanical generation of speech is substituted by a virtual model due to the unavailability of a mechanical system. The virtual system, however, leads to important results in the generation of artificial vocalization both in terms of speech and facial movements. The actual paper make a breakthrough over the algorithms described in [4][5]. The major points of improvement, which are the contributions of this paper, are the following:

- while the rules adopted in [4][5] were tailored on the Italian language, the rules developed in this paper use the International Phonetic Alphabet (IPA) [6], thus leading to the possibility to extend the system to other languages;
- while in [4][5] we used single words and phrases, in this paper we describe a system able to produce vocalization from unrestricted text;
- the algorithm uses the average pitch extracted from the tutor as a base of the artificial prosody;
- we include in this paper extensive subjective evaluations of the proposed system that were impossible to perform previously.

The unrestricted text vocalization is performed by acquiring, in an initial training phase, a data base of small speech units suitably defined. The acquisition is performed automatically using the speech of a human tutor. The tutor is asked to pronounce a number of given utterances which are analyzed and automatically segmented in the defined small units.

Our algorithm, inspired by the human process of language
acquisition, is depicted by Fig. 2.

The proposed algorithm is divided into a training phase and a synthesis phase. In the training phase the algorithm automatically learns from the human voice of the tutor. At the end of this phase, a knowledge base of vocalizations is produced. In the synthesis phase the algorithm uses the knowledge base to generate speech and facial movements.

The current version of the virtual talking face speaks Italian. However, the extension to other languages is simplified because we use IPA to build the fuzzy rules.

As explained previously, the knowledge base of vocalization contains speech units that are segments of utterances extracted from different words. Each speech unit is stored together with its contextual information, i.e. previous and following phonemes. This information is used in the concatenation phase to choose the segment that best fit the actual phonetic environment of the unit.

II. RELATED WORK

It is worth noting that even if there are many existing solutions that generate facial movements together with artificial speech, very few of them go through the articulatory movements of the phonatory organs. As explained before, we considered the articulatory level as a fundamental one, because the related commands could be used eventually to control mechanical talking face, such as that developed in [2][3].

Besides driving the mechanical phonatory organ of a mechanical talking head, the knowledge of articulatory movements is fundamental to deliver the facial movements, since an important link between the movements of the inner organs of the vocal tract and the facial movements has been recognized by many researchers [7], [8].

Similarly to the approach proposed in this paper, Nishikawa et al. [2] [3] developed an algorithm which estimates the articulatory parameters using Newton optimization. In this work we instead use more general genetic optimization by which we can generate continuous utterances.

In [9], a phoneme-acquisition system was developed using a computational model that explains the developmental process of human infants in the early period of acquiring language. It is based on recurrent neural networks.

III. UNRESTRICTED TEXT GENERATION OF ARTIFICIAL VOCALIZATION

The speaking humanoid produces artificial speech by concatenation of small speech units. Inspired to the early developments in speech synthesis we use the following set of speech units which represent transitions between phonemes:

- 70 elements /CV/, i.e. consonant-vowel transitions such as /ba/ or /ka/,
- 10 elements /C V/ which are transitions between affricate and vowels, such as /tʃa/,
- 8 elements /NV/ which are Diphthong
- 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

Thus a total of 112 speech units are contained in this basic speech units list.

Each speech unit is represented by a transition of acoustical parameters. An example related to a formant trajectory is reported in Fig. 3.

As described in the figure, each transition is represented by five parameters: the initial and final value of the trajectory locus called L1 and L2 respectively, and the initial, duration and final values called I, D and F respectively. These five parameters are used to generate a transition. More transitions are concatenated to form complex trajectories of multiples units as shown in Fig. 4, where the F2 trajectory resulting from the concatenation of the segments /pa/ - /ti/ - /ta/ is reported. Together with the formant trajectory parameters, which are used to synthesize the signal, we store in the knowledge base also the configuration of the articulatory organs which would be used to generate that signal. The knowledge base is reported in Fig. 5.
A. Fuzzy model of the vocal tract

Our algorithm uses, as imitation learning approach, the model proposed in [4] for the Italian language. In this model, articulatory parameters are estimated from the tutor speech using a genetic optimization scheme.

The genetic optimization is based on the following consideration. Phonemes can be classified in terms of manner and place of articulation. The manner of articulation is related to the degree of constriction imposed by the vocal tract on the airflow; the place of articulation refers to the location of the most narrow constriction in the vocal tract. Under the considered model, the following six categories of the manner of articulation have been considered: vowel, in which air flows throw the vocal tract without constrictions; liquid, similar to the vowels but that use the tongue as an obstruction; nasal, which is characterized by a lowering of the velum, allowing airflow out of the nostril; fricative, which employ a narrow constriction in the vocal tract which introduces turbulence in the air flow; plosive, involving a complete closure and subsequent release of a vocal obstruction; affricate, which is a plosive followed by a fricative.

Using manner and place of articulation, any phoneme can be characterized in fuzzy form. In this way, the phonetic description appears as an extension of the classical binary definition described for instance by Fant in [10], and a certain vagueness in the definition of the place of articulation of the phonemes is introduced.

Representing the array of features as (rounded, open, anterior, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar), the /a/ phoneme, for example, can be represented by the array:

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

indicating that /a/ is a vowel, with a degree of opening of 0.9, of rounding of 0.32, and it is anterior at a 0.12 degree. Such arrays, defined for each phoneme, are the membership values of the fuzzy articulatory features of the phonemes.

As an example of fuzzy rules, let us consider the following consonant: /p/ and /t/. These consonants are described in IPA form as reported in Tab. I. In other words, they are defined as voiceless bilabial plosive and voiceless alveolar plosive respectively.

| TABLE I  |
|---|---|---|
| PLOSIVE CONSONANT TRANSITION. |  |  |
| plosive | /p/ | /t/ |

All the fuzzy sets for the acoustic parameters have trapezoidal membership functions and have been defined as follows: Initial Locus L1(p), Initial Interval I(p), Duration D(p), Final Interval F(p), Final Locus L2(p). The fuzzy values of this variable depend on the actual variable to be controlled. Calling p0 and p1 the first and second phonemes of the transition respectively, some of the fuzzy rules involved in the generation of the transitions /pa/, /pi/ and /pu/ and /ta/, /ti/ and /tu/ 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;

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 two 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 $p_0$ is voiced and $p_1$ is Round then
If $L_2(F_3)$ is low;
if $p_0$ is voiced and $p_1$ is not Round then
$L_2(F_3)$ is medium;

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

Moreover, in general the fuzzy expressions involve the fuzzy operators AND, NOT and OR. The rule decoding process is completed by the defuzzification operation, which is performed with the fuzzy centroid approach.

The output speech is synthesized using a formant synthesizer [11]. The output speech is compared to the input speech and the measure of their difference is used as fitness function of the genetic optimization phase. The genetic optimization aims at computing the optimum values of the degrees of membership for the articulatory features used to generate an artificial replica of the input signal.

The optimal degrees of membership of the articulatory places minimize the distance from the uttered signal. The inputs are the number of phonemes of the signal and their classification in terms of manner of articulation.

The fitness, which is the distance measure between original and artificial utterances and is optimized by the genetic algorithm, is an objective measure that reflects the subjective quality of the artificially generated signal. Such measure is based on the computation of the pitch loudness, which is a psycho-acoustical term defined as the magnitude of the auditory sensation. This measure is used to compare the artificial signal generated by the fuzzy module and the speech generation module against the original input signal.

The alignment between the original and artificial utterances is performed by using dynamic programming [12], with slope weighting as described in [13]. Using the mapping curve between the two signals obtained with dynamic programming, the distance $D$ between original and artificial utterances is computed. Thus, the fitness function of the Place of Articulation $PA$ is:

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

The goal of the genetic optimization is to find the membership values that lead to a maximization of the fitness, i.e., the minimization of the distance $D(X,Y)$, namely $PA = \text{argmax}\{Fitness(PA)\}, \ PA = \bigcup PA_i, \ i = 1, \ldots, 12 \cdot N,$ where $PA_i$ is the degree of membership of the $i$-th place of articulation, $N$ is the number of phonemes of the input signal.

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

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

where $P_j$ is the penalty function and $N_c$ is the number of constraints. The system is connected to the virtual talker face that should speak synchronizing audio and video.

B. Methodology

The speaking humanoid robot produces artificial vocalization from unrestricted text. The vocalization module is composed by a training phase and a synthesis phase.

1) Training phase: In the training phase the knowledge base of the speech units is filled using the genetic fuzzy articulatory module. The overall architecture of this phase is the following: to the human operator who train the system is presented a list of words to utter. The genetic-fuzzy articulatory system produces the $(I, D, F, L_1, L_2)$ parameters on one side to describe the formant trajectory and the articulatory parameters on the other side to describe what are the movements of the articulatory organs in a human being that would produce a similar signal.

2) Synthesis phase: In the synthesis phase, we gather the text to be converted in artificial speech and facial movements from the cognitive module. The text is divided into speech units. The units are concatenated on the basis of the $(I, D, L, F_1, F_2)$ parameters to form the global formant trajectory of the artificial sentence. Then, artificial speech is obtained using formant synthesis. The excitation sequence is generated by adding a fricative noise to the excitation model proposed in [14] as reported in Fig. 6.

C. Unrestricted text synchronized facial movements

Facial movements are linked with articulatory movements as demonstrated by many authors. Many of them show that most of the correlation between facial and vocal tract (articulatory) movements (slightly more than 90%) can be expressed by a linear transform [15]. Using a linear correlation between facial and articulatory movements, we have selected a set of muscles in a virtual face to demonstrate the automatic synchronization between facial movements and uttered speech. The muscles selected are one that can move the upper lip, one that can move lower lip, two for lateral facial movements and two for mouth movements (Fig. 7).

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 ut-
Fig. 7. Muscles on the talker face.

Fig. 8. Analysis of transitions: /pa/ /pi/ /pu/.

Fig. 9. Analysis of transitions: /ti/ /ta/ /tu/.

IV. EXPERIMENTAL RESULTS

In order to verify the quality of the speech synthesized and the synchronization between movements and speech uttered, two sorts of experiments have been worked out. First, we have considered single words composed by similar sets of transitions, to show the behavior of the first formants of speech. Second, we have performed a set of subjective tests to evaluate the naturalness and the quality of the synthetic speech and the quality of synchronization between the uttered speech and the corresponding facial movements computed with the proposed model.

A. Analysis of single transitions

For producing the database of vocalization from a tutor, a set of words in Italian has been selected. This set of words contain a full range of transitions consonant-vowel of the Italian language, and are coded using IPA phonetic symbols. Using the genetic-fuzzy optimization on these words pronounced by the tutor, the learning phase collects a database of vocalization for these transitions.

To test the quality of the vocalization obtained, we have plotted the behavior of the second formant for the transitions /pa/, /pi/, /pu/. As shown in Fig. 8 the trends of these curves are coherent with the expected phonetic behavior of these transitions. More precisely, from Fig. 8 we can estimate the locus for the second formant, that correspond to the frequency of the transition /pu/, at around 1000 Hz.

We have also considered another set of transitions, namely /ti/, /ta/, /tu/, and the behavior of the second formant for these transition, shown in Fig. 9, is as expected, and the locus of the second formants is around 1500 Hz.

Using the above shown transition, we can obtain the word /patita/ using concatenation of transitions. In Fig. 10 the spectrogram of the word /patita/ is depicted: on the top uttered by the human tutor, on the bottom the corresponding synthetic speech.

B. Subjective evaluations

To test the quality of the synthetic speech a set of subjective evaluation tests of words generated with the proposed approach have been conducted. A set of 240 words, 120 no sense word and 120 existent word of the Italian language, have been produced starting from the database of vocalizations created by learning from a human tutor. All the words selected contains only transitions consonant-vowel; different transitions has not been tested. These words have been reproduced by a talking face driven by the facial movements connected to the articulatory parameter as described in Sec. III-C.

We have tested three main features: naturalness of the uttered speech, the comprehension of the word, and the quality of synchronization between uttered speech and facial movements. Tab. II shows that the synthesized speech has a good level of naturalness, but it can be improved as future development. Tab. III shows that the comprehension of the uttered words is greater for known words, as expected, and that the visual representation of the face help to understand
Fig. 10. Spectrogram of the word /pa/ti/ta/: human utterance on the top, synthetic utterance on the bottom.

TABLE II

|
| Naturalness Results for No Sense and Existing Words Synthetic Utterances |
|-----------------------------|-----------------------------|-----------------------------|
|                             | Only audio | Audio+video | Only audio | Audio+video |
| No sense words              |             |              |             |              |
| POSITIVE                    | 43.0%      | 55.5%        | 65.5%      | 74.0%        |
| NEGATIVE                    | 57.0%      | 44.5%        | 34.5%      | 26.0%        |
| Existing words              |             |              |             |              |
| POSITIVE                    | 52.0%      | 60.0%        | 82.0%      | 84.5%        |
| NEGATIVE                    | 48.0%      | 40.0%        | 18.0%      | 15.5%        |

the utterance. In Tab. IV the synchronization between speech and facial movements has been tested. The model provides good synchronization for both no-sense and existing word of the Italian language.

V. CONCLUSIONS

In this paper an algorithms which automatically acquires vocalization capability from a human tutor and that can produce artificial vocalization from unrestricted text is presented. The main features of the proposed methods concern the possibility of produce articulatory movements, facial movements and speech at the same time, to allow to use the same algorithm on both artificial humanoid head that has mechanical organs for utterance emission and on virtual talking faces. The experimental results show that the technique is effective, as the speech produced present a phonetic behavior similar to human utterances, and speech and the synchronized facial movements have a good degree of acceptance in subjective tests.

TABLE III

<table>
<thead>
<tr>
<th></th>
<th>Only audio</th>
<th>Audio+video</th>
<th>Only audio</th>
<th>Audio+video</th>
</tr>
</thead>
<tbody>
<tr>
<td>No sense words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSITIVE</td>
<td>52.0%</td>
<td>60.0%</td>
<td>82.0%</td>
<td>84.5%</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>48.0%</td>
<td>40.0%</td>
<td>18.0%</td>
<td>15.5%</td>
</tr>
<tr>
<td>Existing words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSITIVE</td>
<td>57.0%</td>
<td>55.5%</td>
<td>84.5%</td>
<td>87.8%</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>44.5%</td>
<td>34.5%</td>
<td>16.0%</td>
<td>12.2%</td>
</tr>
</tbody>
</table>

TABLE IV

<table>
<thead>
<tr>
<th></th>
<th>Only audio</th>
<th>Audio+video</th>
<th>Only audio</th>
<th>Audio+video</th>
</tr>
</thead>
<tbody>
<tr>
<td>No sense words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSITIVE</td>
<td>91.5%</td>
<td></td>
<td>84.0%</td>
<td></td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>8.5%</td>
<td></td>
<td>16.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>90.0%</td>
<td></td>
<td>84.0%</td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES