Text-driven avatars based on artificial neural networks and fuzzy logic

Mario Malcangi

***Abstract*—**We discuss a new approach for driving avatars using synthetic speech generated from pure text. Lip and face muscles are controlled by the information embedded in the utterance and its related expressiveness. Rule-based, text-to-speech synthesis is used to generate phonetic and expression transcriptions of the text to be uttered by the avatar. Two artificial neural networks, one for text-to-phone transcription and the other for phone-to-viseme mapping have been trained from phonetic transcription data. Two fuzzy-logic engines were tuned for smoothed control of lip and face movement. Simulations have been run to test neural-fuzzy controls using a parametric speech synthesizer to generate voices and a face synthesizer to generate facial movement. Experimental results show that soft computing affords a good solution for the smoothed control of avatars during the expressive utterance of text.

***Keywords*—**Speech-driven avatar, phone-to-viseme conversion, text-to-speech synthesis, artificial neural network, fuzzy logic.

I. Foreword

Over the last ten years, interest in human-like conversational interfaces [1] has grown rapidly, driven by applications where the human-machine interface requires natural interaction. Because reading the human being's mental state during natural conversation improves the understanding of speech, synthesizing facial expression is a primary requirement for developing expressive talking heads [2]. Avatars able to speak expressively have become an important requirement for applications such as interactive games, interactive web pages, communicating with the hearing impaired, and so forth.

Facial gesture is a combination of verbal e non-verbal communication made with movements of the head and face. Head and eyebrow movements (blinking, frowning, rising, lowering, gaze direction, etc.) can be automatically generated from the speech [28] signal. Lip synching systems [10], [3] and automatic systems for full facial animation driven by speech signal [15] are implemented using softcomputing methods (neural networks and genetic algorithms).

In a recent work [20], Zoric et al. develop an automatic system for full facial animation driven by speech, using universal architecture for statistically based human gesturing. Such work demonstrated that also a small number of speech features can improve facial animation in terms of naturalness and communication capabilities. Using more speech parameters (pitch and its related dynamic) a very realistic facial animation process can be gained.

Speech can be considered a single communication medium in which information is represented multimodally. Information conveyed by speech is not only semantic or syntactic but also emotional, expressive, gestural and intonational.

Synthesis by analysis [25] is an optimal approach to drive a talking head. The audiovisual analysis creates the face model of the avatar with associated a large database of mouth images. The phonetic transcript of text is then used to drive the mouth selecting and concatenating appropriately mouth images from the database.

In applications for lip-synching [24] that directly synchronize uttered speech with lip and face movement, the information embedded in speech is often lost because it is too difficult to extract information like emotion or gesture. Only a few general speech parameters, such as amplitude and pitch variability, can be measured and tracked. However, these low-level measurements fall far short of those needed to drive an avatar with the full information content of uttered speech.. Tracking these variables leads to very good results for lip synchronization [3], but the avatar is driven with greatly impoverished expression, resulting in very limited naturalness. To overcame this problem, synthetic speech can be used instead of natural speech to drive the avatar [4].

Text-to-viseme may by the right approach to control an avatar for natural utterance [26]. The text-to-viseme process can translate text into the appropriate viseme and supplement this basic information with other related information such as emotion or gesture [5] , [6], [7].

Rule-based, text-to-viseme synthesis has been successfully implemented by considering emotion an additional item of information [8] and used for direct visual speech synthesis [9]. Such approaches separate the tasks of speech synthesis and face-control synthesis, despite the fact that they are part of a single, integrated task in human utterance behavior.

Artificial, neural-network-based, text-to-viseme synthesis has been also explored [10], [11], demonstrating that greater naturalness can be achieved with a soft-computing rather than a hard-computing approach. Fuzzy logic has proven highly effective in smoothing the action of the logical control rules that move an avatar's face muscles during emotional behavior [12].

This research employs both artificial neural networks and

Manuscript received March 9, 2010: Revised version received mm dd, yyyy.

M. Malcangi is professor and researcher at Università degli Studi di Milano – DICo (Informatic and Communication Department), Via Comelico 39, 20135 Milano, Italy (Tel: +39.02.50314003, Fax: +39.02.50316373, e-mail: malcangi@dico.unimi.it).

fuzzy logic to generate phoneme and viseme information that drives face movements during the utterance of a text, as human do. Its goal is to use pure text to feed the whole process, as a human being does when reading a text. The research tries also to solve the problem of reading the words of pure text aloud by generating both speech and the related whole-avatar face motion.

Reading text aloud consists of a set of increasingly complex tasks. The least complex is correctly uttering an isolated word. Of intermediate complexity is correctly uttering of a word in a sentence. The most complex task is uttering a word according to the semantics of the sentence in which the word happens to be located in that particular instance.

Correct utterance of any given word as a single isolated word is accomplished using a set of pronunciation rules that can be inferred from orthographic notation, although they are not explicitly visible. Such rules concern transcribing letters to phones, correctly positioning the tonic syllable on each word, how this stress applies, and controlling the duration of phones and syllables.

Uttering each word in a sentence with sentence-level expressive requires a set of prosodic rules, primarily related to how punctuation marks are distributed. This information can also be encoded into the pronunciation rules, thus enabling coherent control of pitch and duration during the course of the utterance. This expressive uttering of the words in a sentence is the most complex task, because it is not strictly related to the text of the sentence, but above all to the context to which the sentence belongs. Some basic expression rules can be classified and embedded in the pronunciation rules. These can be set up to trigger when certain key words or punctuation marks occur in the sentence (or at the end of it, or even at the end of the preceding sentence).

Information for face-movement control can be extracted automatically from the text according to rules extended from these pronunciation signals.

II. SYSTEM ARCHITECTURE

To design the expressive synchronized-speech and face-synthesis system, a two-phase process framework was build. The whole process can be considered a general-purpose model for designing an integrated system of expressive, avatar-based speech communication in human-computer interfaces.

The first phase involves training and tuning two artificial neural networks (ANNs) for text-to-phones and for phones-to-viseme synthesis, respectively. Two fuzzy-logic engines are also used to smooth speech and face-muscle control.

Fig. 1 shows how a rule-based, text-to-phone and text-to-expression transcriber trains the ANN-based, text-to-phone generator and the ANN-based text-to-viseme generator. Using such a transcriber, only pure ASCII text is used to train the ANNs. Ancillary data for speech and facial expressiveness is automatically extracted from the text by means of regular-expression-based description rules. The two fuzzy-logic engines are manually tuned using a fuzzy logic development environment. This allows for editing of the fuzzy rules and the membership functions according to expert experience. (The tuning task can be also performed by a genetic algorithm).

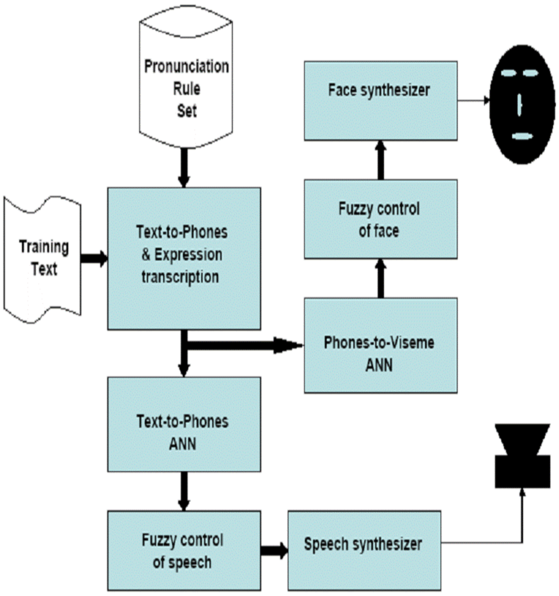

Fig. 1 Training and tuning process of the ANNs and the fuzzy logic engines.

The second phase consists of testing speech synthesis when executed synchronously with face motion, as shown in Fig. 2.

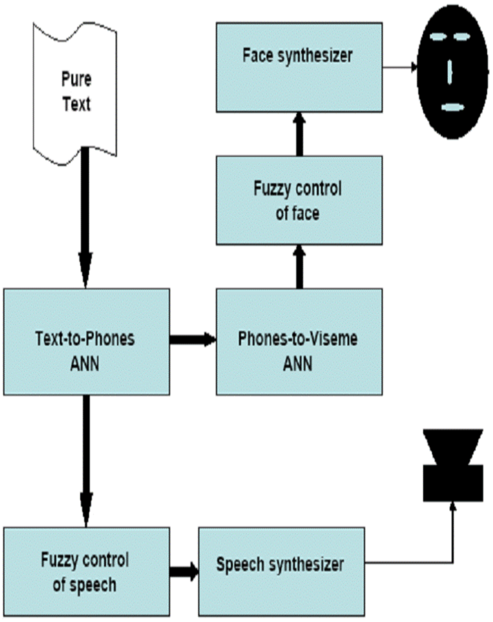

Fig. 2 Testing process for expressive speech and face-motion control.

The additional components of the test process consist of a formant-based speech synthesizer and a viseme generator. The formant-based synthesizer allows full control of speech parameters, so any modulation of speech can be achieved. The viseme generator allows control of facial movement and expression during utterance.

III. Rules for Transcribing into Phones and Expressions

Transcribing text into phonemes and expressions [17] consists of a series of processing steps that are applied to the text. The text is first preprocessed to convert non-alphabetical elements such as numbers, letters pronounced singly as in words spelled aloud, abbreviations, and special ASCII symbols into the corresponding expanded text. Punctuation and word boundaries are processed by a set of rules that encodes the expression. Each word in the text is converted into phone-and-expression stream using a language-specific set of rules.

The rules have the following format:

$$C(A)D = B \tag{1}$$

where:

where A is the the text transformed into the phonetic utterance and facial expression B if the text to which it belongs matches A in the sequence CAD. C is a pre-context string and D is a post-context string.

To compile the rules, the following classes of elements were defined:

```
[!]  |  (^) | ($)
[#]  |  ([AEIOUY]+)
[:]  |  ([^AEIOUY]*)
[+]  |  ([EIY])
[$]  |  ([^AEIOUY])
[.]  |  ([BDGJMNRVWZ])
[^]  |  ([NR])                (2)
```

[!] any non-alphabetical character

[#] single or multiple vowels

[:] zero or more than one consonant

[+] one front vowel

[$] one consonant

[.] one voiced consonant

[^] N or R consonant

For each class, a regular expression is used to encode the rules in compact fashion.

The salient rules for encoding {p} are as follows (note that the dash in a phone slot represents its boundaries and that numbers preceding the phone indicate duration):

```
!(P)!=/1p/1i/
(P)!=/1p/4h/4-/
(PA)STE=/1p/1eI/1j/
!(PHOTO)=/1f/1o/1w/1t/2o/2w/
!(PHYS)=/1f/2I/1z/
(PH)=/1f/
(PPH)=/1f/
(PEOP)=/1p/1i/1p/
!(POE)T=/1p/1o/1E/
(POUR)=/1p/1o/13/
(POW)=/1p/1O/1u/                    (3)
(PP)=/1p/
!(PRETT)=/1p/1r\/2I/1t/
(PRO)VE=/1p/1r\/1u/
(PROO)F=/1p/1r\/1u/
(PRO)=/1p/1r\/1o/
(PSEUDO)=/1s/2u:/2d/3o/3w/
(PSYCH)=/1s/2a/2a/3j/1k/
!(PS)=/1s/
!(PT)=/1t/
CEI(PT)=/1t/
(PUT)!=/1p/1U/1t/4-/
!(P)=/1p/2H_f/
(P)=/1p/
```

We have based the transcription standard on X-SAMPA [29] (right context). The phonetic information is combined with other voice data to control a series of parameters such as duration, stress, etc. The following is the set of symbols used to encode the phonetic information:

X Sampa-like	*Sample word*
@	(a)bout
{	m(a)p
{:	m(a)d
e	sc(e)lte
E	b(e)lla
E_r	b(e)nché
i	v(i)sti
a	(o)dd
u	p(u)nto
o	p(o)ngo
O	(o)rto
O_r	c(o)priletto
I	h(i)t
A	f(a)ther
p	(p)ongo
pp	ca(pp)otto
p:p	stra(pp)o
_h	p()ail
h	(h)eart
H_f	chic(k)en
H_n	uh-(h)uh
H_v	ba(h)
H_c	d()oes
?	Clin(t)on
4	ie(r)i
r	(r)aro
r:4	ca(rr)o
4r	co(rr)esse
44	dà (r)agione
r\	(r)ed
R	(r)oi
R\	D(r)ang
J	a(ñ)o
F	go(n)fio
9	n(eu)f
oU	b(oa)t
V	c(u)t
U	f(oo)t
Q	cl(o)ck
y	t(u)

(4)

IV. Training the ANNs to transcribe

The two ANNs used to transcribe text into phones and expressions and to convert phones and expressions into visemes both have three-layer, feed-forward, backpropagation architectures (FFBP-ANN).

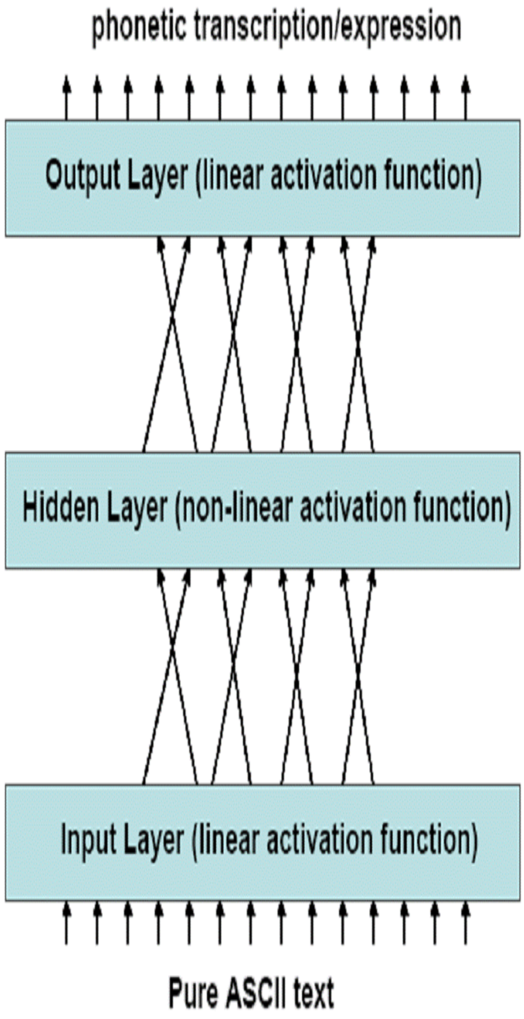
phonetic transcription/expression
Output Layer (linear activation function)
Hidden Layer (non-linear activation function)
Input Layer (linear activation function)
Pure ASCII text

Fig. 3 Architecture of the FFBP-ANN

The first ANN [16] transcribes phones and expressions from text input. Its output is the input for the second ANN, whose output is viseme encoding.

A linear activation function controls connections at the input and the hidden layer nodes. A non-linear (sigmoid) activation function connects hidden-layer nodes to the output-layer. The non-linear activation function is:

$$s_i = \frac{1}{1+e^{-I_i}} \qquad (5)$$
$$I_i = \sum_j w_{ij} s_j$$

where:

s is the output of the *i*-th unit
E_i is the total input
w_{ij} is the weight from the *j*-th to *i*-th unit

The first ANN's input is a text window of nine consecutive characters. This window slides from right to left. Current output encodes the phone and the expression that correspond to the middle character in the input-layer string, taking into account the pre-context and post-context of the current input character.

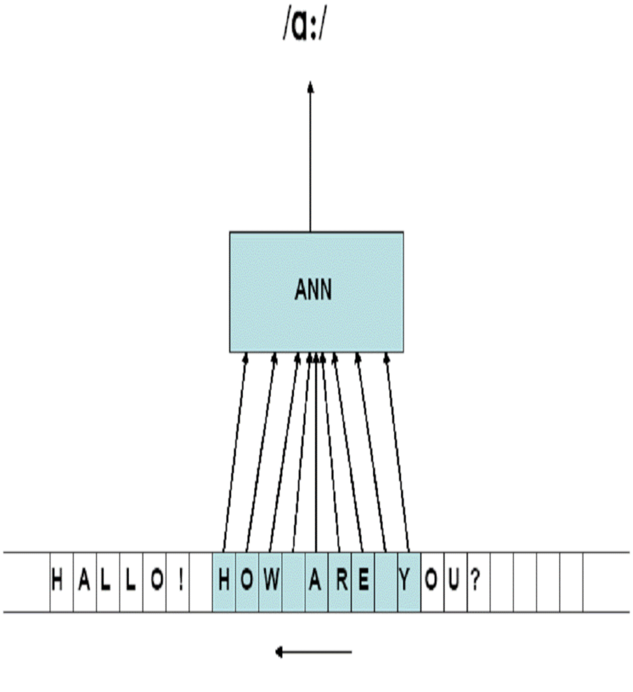

Fig. 4 Sliding window

The text-to-phone/expression transcription system is used to train the ANN to transcribe text to phones and expression. This generates ANN input-output training patterns for a large variety of texts. The ANN thus learns to read unknown text with expression. The second ANN is trained in analogous fashion, but only after the first ANN has been fully trained. The first ANN's output is used as input for the second ANN, employing the same sliding-window strategy. A basic viseme set is used as reference to train the ANN during the error back-propagation process.

V. Smoothing Speech and Movement with Fuzzy Logic

The two trained ANNs can drive the speech synthesizer and the avatar face. However, in order to utter speech and move the face more naturally, the ANNs' output needs smoothing before it is applied to the speech synthesizer and to the avatar's face controller. This smoothing is accomplished with fuzzy logic in two steps. Two separate fuzzy subsystems convert the ANN-output expression state into control levels, first for speech dynamics and then for face muscles. Crisp information (intensity, level, etc.) about expression was transformed into fuzzy rules. The resulting crisp control level comes from an appropriate defuzzifying process.

The two fuzzy subsystems are identically structured. They differ only in their settings (i.e., the knowledge base). Each consists of a fuzzifying front end, a rule-based inference engine, and a defuzzifying back end.

The first step in the fuzzy-engine tuning process consists of modeling crisp intensity and level information into fuzzy measurements. This is done by modeling seven fuzzy sets: imperceptibly low, very low, moderately low, medium, moderately high, very high, and extremely high.

Triangular and trapezoidal membership functions are used to implement these fuzzy sets. The shape and relations among these are reported qualitatively in Figure 5. Tuning is accomplished by an expert who uses a fuzzy-logic development environment to simulate and evaluate the resulting membership degrees for each crisp input.

The second step consists of editing and tuning a set of inference rules such as:

$$\text{IF x AND y THEN z} \qquad (6)$$

where x and y are membership grades for the intensity and level of speech and facial expression we intend to smooth before they are applied as controls. z is the degree of control to be applied.

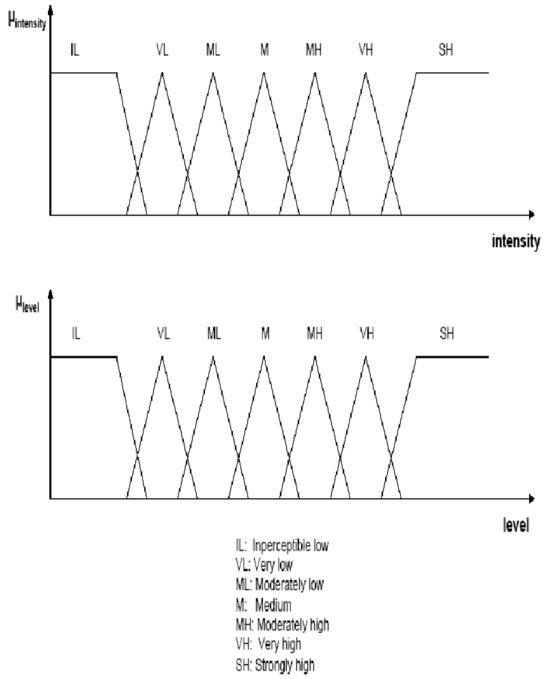

Fig. 5 Fuzzy modeling of speech synthesis and facial control inputs.

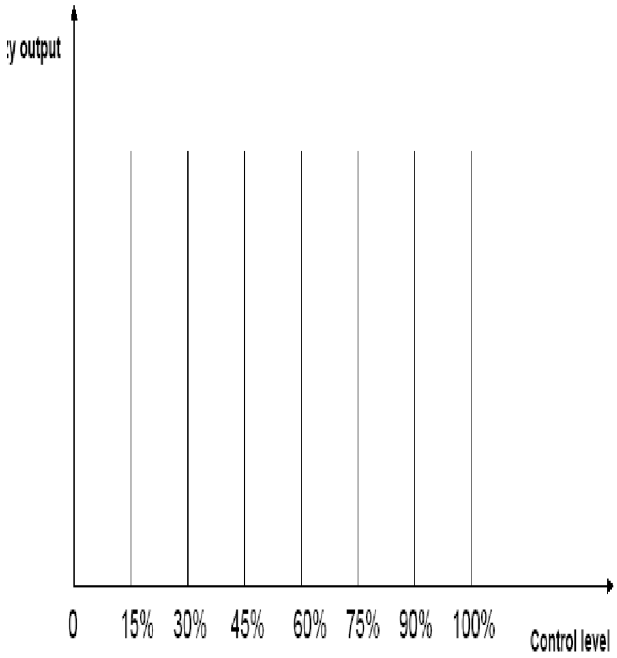

Fig. 6 Singleton membership functions to defuzzify the output control level.

The third step consists of defuzzifying output control. This is done by converting degrees of control into crisp control with set of singleton membership functions (fig. 6) and a weighted-average (center of gravity):

$$Control = (AxB)/(A+B) \qquad (7)$$

VI. Speech synthesis model

Several solutions can be used to synthesize speech when text-to-speech (TTS) transcription is available. Concatenative synthesis [13] is currently used for TTS applications, but formant-based synthesis is more interesting for embedded applications because it can potentially synthesize any voice and can be modulated according to emotional information.

The speech-synthesizer model we refer to emulates the human vocal tract. The reason for this choice is that unlimited utterances need to be generated with extensive audio such as singing, yawning, cough, laugh, etc.

This speech-synthesis model achieves naturalness by producing speech with dynamically controlled processing elements: filters, generators, and modulators. Coarticulation, phonetic articulation-rate, and inflection (pitch) can all be controlled, statically or dynamically. Speech type (male, female, child, etc.) and alterations (bass, baritone) are also controllable.

Formant-based speech synthesis [22] is an extremely accurate model for human speech modelling because it models the sound source and the formant frequencies of each speech sound that the avatar needs to utter. This model essentially consists of a set of virtual synthetic human vocal cords (a glottis) and of the articulators that go with it.

Both these components are parameter-controllable, so that no speech sounds need to be pre-recorded to accomplish the synthesis. Therefore such a speech synthesizer represents the avatar’s embedded phonetic apparatus.

High quality speech can be generated if appropriate input parameters are supplied to the speech synthesizer. Natural utterance is obtained with parallel formant synthesis controlling appropriately the first three formants (F1, F2, and F3) in terms of bandwidth and excitation sources.

Such control is generated by the fuzzy logic engine that produce control trajectories for natural coarticulation of phones sequences generated by the text-to-phone ANN engine.

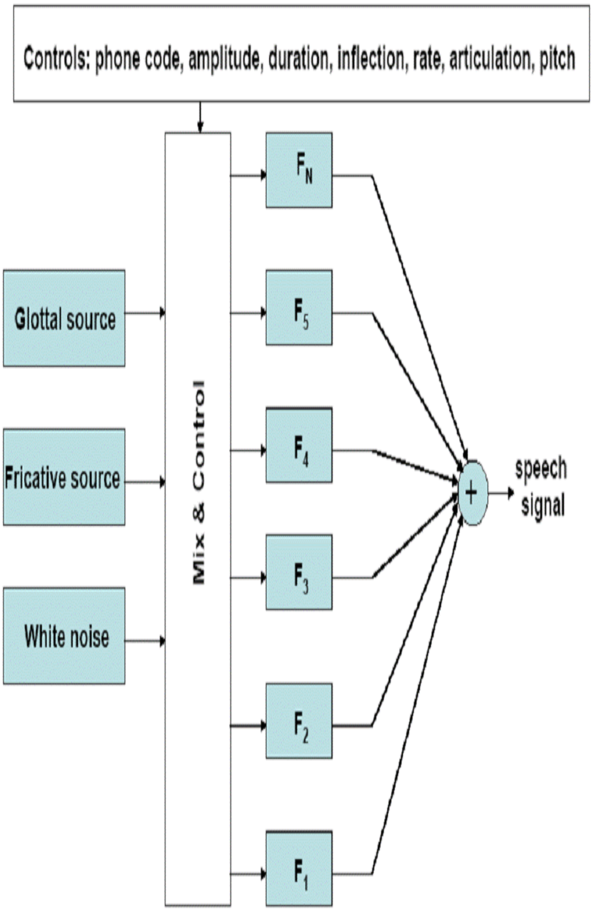

Fig. 7 Formant-based speech synthesizer.

The speech synthesizer (fig. 7) consists mainly of two primary sound-source generators, one for voiced sounds and one for unvoiced sounds. Multiple amplitude control is also implemented because different sound sources (voice, aspiration, frication, etc.) need to be combined to generate a complete speech sound (e.g. voiced consonants and unvoiced noisy sounds). Pitch is controlled in terms of fundamental frequency F_0 and variability, according to expression information embedded in the text to be uttered. A set of filters that operate as resonators can\ be programmed to profile the distribution of formants for each sound that needs to be synthesized.

VII. Face synthesis model

Facial synthesis by analysis is a promising approach to the problem of automatically controlling an expressive talking head. A good means of analysis might use markers to read facial movements during utterance [14]. A more advanced method might refer to algorithm-based extraction of semantic description of the face and the components that make up its movement [23] to replay such movement in synthesized form.

We use a two-dimensional face synthesizer to apply speech intensity parameters that control two different components of facial modeling: lip and face modifications during expressive utterance. These are controlled in terms of mouth opening and in terms of the strength of expression-control muscles.

The face-animation system used here has a viseme editor that directly models the set of visemes to be associated with a character. The tongue and lips can be designed by tuning the controls related to mouth position, to lip parameters, and to the strength of mouth-opening muscles. Each viseme is associated with the related phone that will be uttered by the speech synthesizer.

The fuzzy, smoothed control produces variable dynamics during the utterance of stationary speech units such as phonemes and allophones. This dynamic control is used to modulate the amplitude of lip-opening strength, resulting in more natural movement. Expression-control muscles are also dynamically controlled to produce modifications, such as: stretching or relaxing face muscles, frowning with eyebrows, wrinkling the forehead, flaring or contracting the nostrils.

Fuzzy smoothed control of facial expression is used also to blend facial expression with speech expression, above all when an emotion is not compatible with the uttered speech. In such cases the fuzzy control acts as a morpher.

VIII. Conclusion

This research was designed to overcome the problem of synchronizing audio information with video representation of the face of a speaking avatar, avoiding the mismatch between audio stream and face movement. Its was also intended to fully automate the process of controlling face animation during the utterance, starting from pure text and completely independent of any human speaker.

Preliminary results of this research demonstrate that soft computing offers a good solution for the smoothed avatar control during the expressive utterance of text. Using pure text as input information, correct expressive utterance of each word (letter sequence) was achieved. Furthermore, the related expressive avatar face movements were synchronized. The next step will apply a similar approach to automatically extracting high-level expression information related to word sequence.

The importance of high-level expression, as demonstrated in other research [18], [27] into using talking avatars in commercial applications (e-commerce), as well as into general communication processes [21], encourages an extension of the proposed text-to-speech-driven face towards an emotional text-to-speech-driven avatar that would also include body movement, mainly head and harms movement. To this end, more extensive use of softcomputing methods will be made, since most emotional end gesture information is embedded in the text to be uttered by the avatar [19]. A neural network can be trained to analyze the text to be uttered and to classify gestures and emotions embedded in the text.

REFERENCES

[1] Goh O.S., Fung C.C. (2008) The design of interactive conversation agents. WSEAS Transactions on Information Science & Applications, Issue 6, Vol. 5, pp. 901-912

[2] Fujita H., Hakura J., Kurematsu M. (2006) Virtual Cognitive Model for Miyazawa Kenji Based on speech and facial images recognition. WSEAS Transactions on Circuits and Systems, Issue 10, vol. 5, pp. 1536-1543.

[3] Malcangi M., De Tintis R. (2004) Audio based real-time speech animation of embodied conversational agents. Lecture Notes in Artificial Intelligence, LNAI 2915, Springer-Verlag, Berlin, Eidelberg, pp 350-360

[4] Zelezny M., Krnoul Z. (2003) Czech audio-visual speech synthesis with a HMM-trained speech database and enhanced coarticulation. WSEAS Transactions on Computers, Vol. 2, pp. 733-738

[5] Masuko T., Kobayashi T., Tamura M., Masubuchi J., Tokuda K. (1998) Text-to-visual speech synthesis based on parameter generation from HMM. In Proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, May, pp. 3745-3748

[6] Gao W., Xu L.,Yin B., Liu Y., Song Y., Yan J., Zhou J., Chen H. (1997) A text-driven sign language synthesis system. Proceedings of CAD & Graphics'97, December 2-5, Shenzhem, China

[7] Zliekha M. A., Al-Moubayed S., Al-Dakkak O., Ghneim N. (2006) Emotional audio visual arabic text to speech. Proceedings of Eusipco

[8] Beskow J. (1995) Rule-based visual speech synthesis. ESCA, Eurospeech '95, Madrid

[9] Agelfor E., Beskow J., Granstrom B., Lundeberg M., Salvi G., Spens K., Ohman T. (1999) Synthetic visulal speech driven from auditory speech. Proceedings of AVSP 99

[10] Zoric G., Pandzic I. S. (2006) Real-time language independent lip synchronization method using a genetic algorithm. Signal Processing, Vol. 86, Issue 12, December, pp. 3644-3656

[11] Massaro D. W., Beskow J., Cohen M. M., Fry C. L., Rodriguez T (1999) Picture my voice: Audio to visual speech synthesis using artificial neural networks. Proceedings of AVSP'99, Santa Cruz, California

[12] Alotaibi M. B., Rigas D. I. (2008) A usability evaluation of multimodal metaphors for custome knowledge management. International Journal of Computers and Communications, Issue 2, Vol. 2, pp. 59-68

[13] Orhan Z., Gormez Z. (2008) The framework of the Turkish Syllable-based concatenative text-to-speech system with exceptional case handling. WSEAS Transactions on Computers, Issue 10, Vol. 7, pp. 1525-1534

[14] Bottino A., Cumani S. (2008) A fast and robust method for identification of face landmarks in profile images. WSEAS Transactions , Issue 8, Vol. 7, pp. 1250-1259

[15] Cerekovich A., Zoric G., Smid K., Pandzic I. S. (2008) Towards realistic real time speech-driven facial animation. H. Prendingher, J. Lester, M. Ishizuka (Eds.), IVA 2008, LNAI 5208, pp. 476-478, Springer-Verlag Berlin Heidelberg.

[16] Malcangi M., Frontini D. (2009) Language-independent, Neural Network-based, Text-to-phones Conversion", Neurocomputing, Elsevier, Volume 73, Issues 1-3, December, pp. 87-96.

[17] Malcangi M., Grew P. (2009) A Framework for Mixed-language Text-to-speech Synthesis", WSEAS International Conference on Computational Intelligence, Man-machine Systems and Cybernetics (CIMMACS '09), pp. 151-154, Puerto de la Cruz, Tenerife, Spain, December 14-16.

[18] Alotaibi M. B., Rigas D. I. (2009) The role of avatars with facial expressions to communicate customer knowledge. International Journal of Computers, Issue 1, Volume 3, pp. 1-10.

[19] Wong J.-J., Cho S.-Y. (2010) A face emotion tree structure representation with probabilistic recursive neural network modelling. Neural Computing & Applications, Springer-Verlag, 19, pp. 33-54.

[20] Zoric G., Pandzic I. S. (2008) Towards realistic real time speech-based facial animation applications built on HUGE architecture. Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP 2008, Moreton Island, Queensland, Australia September 26-29.

[21] Drahota A., Costall A., Vasudevi R. (2008) The vocal communication of different kind of smile. Speech Communication. Elsevier, issue 50, pp. 278-287.

[22] O'Shaugnessy, D. (1987) Speech Communication – Human and Machine. Addison-Wesley, Reading, MA.

[23] Ravyse I., Sahli H. (2006) Facial analysis and synthesis scheme. In ACIVS 2006, LNCS 4179, J. Blanc-Talon et al. (Edts.), Springer-Verlag Berlin Heidelberg, pp. 810-820.

[24] Ostermann J., Weissenfeld A. (2004) Talking faces – technologies and applications. In 17th International Conference on Pattern Recognition (ICPR'04), Cambridge UK, August 23-26, vol. 3, pp. 826-833

[25] Liu K., Ostermann J. (2009) Optimization of an image-based talking head system. EURASIP Journal on Audio, Speech and Music Processing, Hindawi, vol. 2009.

[26] Mattheyses W., Latacz L., Verhelst W. (2009) On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. In the EURASIP Journal on Audio, Speech and Music Processing, Hindawi, vol. 2009.

[27] Rigas D., Gazepidis J. (2007) A further investigation of facial expressions and body gesture as metaphors in e-commerce. In the seventh WSEAS International Confererence on Applied Informatics and Communications, Athens, Greece.

[28] Albrecht I., Haber J., Seidel H. (2002) Automatic generation of non-verbal facial expressions from speech. In the Procceedings of CGI.

15x Wells J. (1995) Computer-coding the IPA: a proposed extension of SAMPA, http//:www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf.

Prof. Mario Malcangi received his undergraduate and graduate degrees in Electronic Engineering and Computer Science from the Politecnico di Milano in 1981. He is member of the International Neural Network Society and among the founders of the Engineering Applications of Neural Networks Special Interest Group (SIG). His research is in the areas of multimedia communications, digital signal processing, and embedded/real-time systems. His research efforts are mainly targeted at speech- and audio-information processing, with special attention to applying soft-computing methodologies (neural networks and fuzzy logic) to speech synthesis, speech recognition, and speaker identification for implementation on deeply embedded systems. He teaches digital signal processing and digital audio processing at the Università degli Studi di Milano. He has published several papers on topics in digital audio and speech processing.