Biomorphic Modeling of Phoneme Identification and Classification Based on an Evolving Fuzzy-neural Network

From Hardcomputing to Softcomputing

Mario Malcangi
Università degli Studi di Milano
Department of Computer Science
Milano, Italy
Malcangi@di.unimi.it

Hao Quan, Philip Grew
Università degli Studi di Milano
Department of Computer Science
Milano, Italy
Quan@di.unimi.it, Grew@di.unimi.it

***Abstract*—Speech is dynamic in nature and organized in a complex time-and-frequency structure that makes it very hard to solve the issue of automatic speech recognition (ASR) for diverse speaker conditions. The hardcomputing approach to solving this issue (i.e conventional computing based on precisely-stated, analytical, mathematics-inspired models) pushed processing limits because it is highly computationally time-consuming and intolerant of imprecision, uncertainty or approximation in the data. Softcomputing and its biomorphic implementation is a more natural approach to solving the issue of speaker-independent ASR, given its ability to manage imprecision, uncertainty, and approximation, as well as to reduce system complexity to fit the upcoming requirements of next-generation deeply-embedded systems. This paper reports experiments based on an evolving-fuzzy-neural-network (EFuNN) paradigm trained to process and classify phonemes to drive multimodal (audiovisual) speech-to-text transcription and speaker identification**

Keywords— Phonemes; ASR; hardcomputing; softcomputing; audiovisual speech recognition.

I. INTRODUCTION

Recent advances in microelectronics and neuromorphic engineering [1][12] (e.g. spiking neural networks [2]) have laid the technological and methodological groundwork for a new approach to system development, mainly in the area of human machine interaction (HMI). Biomorphic modeling of Certain challenges such as automatic speech recognition finds in bioinspiration new ideas to solve complex issues [3].

Automatic speech recognition (ASR) and Speaker Identification (SI) are challenging tasks, mainly because of complex the time and frequency structure of speech. Hardcomputing, i.e. conventional computing, approaches this issue with precisely-stated, analytical models that require a lot of computing time and are overly sensitive to imprecision, uncertainty, and approximation. Biological systems, such as human beings, do not apply hardcomputing methods to efficiently process speech signals. Our auditory system implements time-frequency processing of the speech signal with biological organs, such as the cochlea. The speech (air-pressure wave) is first transformed into an analog fluid-pressure wave with the same time structure and features. It is then transformed into a sequence of spatially distributed frequencies. [4]. This biological process is the same as computing a Fourier-transform algorithm, in that it maps the wave's time information onto the corresponding frequency information. To accomplish this, the biological system does not carry out mathematical computations on precise measurements in a hardcomputing fashion. Rather, it executes fuzzy measurements on imprecise information, performing the task of audio-frequency sensing very effectively with minimal energy consumption (its power dissipation is only about 14μW), less complexity (a few cubic millimeters) [5]and fewer errors than does an equivalent hardcomputing system like the audio front end of an automatic speech-recognition system. The cochlea is an example of such optimal biological solution to the problem of feature extraction from time-domain information to the frequency domain that acts as an audio-processing front end for the cerebral cortex's phoneme-recognition function at the stage dedicated to fusing related information such as visemes, gestures, context, etc.

The auditory system is fully softcomputing-based, from the acoustic wave to the symbol (a symbolic description of the phoneme). In earlier work we deployed a multimodal, bio-inspired [6], audiovisual ASR (AVASR) framework intended as a reference design of a complete softcomputing-based AVASR, for which a zero-instruction-set-computing (ZISC) implementation could be feasible. The multistage framework extends from the acoustic wave (the microphone) to symbolic transcription (the grapheme). The first stage, phoneme identification, was fully hardcomputing-based, and did not meet the requirements of a complete ZISC architecture.

The purpose of this investigation is to validate a softcomputing implementation of the first stage the AVASR

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

framework implements as hardcomputing [6], so as to accomplish the goal of full ZISC implementation.

The AVASR softcomputing framework was not fully validated in [6] because the first stage, the audio and visual classifiers, was not implemented through softcomputing but through hardcomputing. The main issue to investigate is how to feed the appropriate softcomputing paradigm with data directly from the sensor (microphone) demanding all needed signal conditioning only of analog circuitry, so that bio-inspired speech-processing modeling could be employed.

The following experiments aim to investigate the best softcomputing paradigm and setup to replace the hardcomputing stage in the AVASR framework [6], and other similar frameworks.

II. EFuNN FOR PHONEME-BASED ISOLATED WORD RECOGNITION

Our chosen softcomputing paradigm of reference is the evolving fuzzy neural network (EFuNN)[7], given its capacity to learn by evolving and adapting through evolution. The first step consisted in checking the EFuNN's ability to learn frequency-domain features directly from time-domain data. Sampled data was extracted from the audio wave of an uttered word and labeled according to its phoneme content. This data was used to train the EFuNN. The trained EFuNN was then tested for its ability to match and recognize the right phoneme sequence in the uttered word.

A. *Evolving Fuzzy Neural Network for Knowledge-based Learning*

This fuzzy neural network (EFuNN) is the implementation of the evolving-connectionist-system (ECOS) [8] paradigm, which enables on-line adaptation and evolution over time. The EFuNN's ability to evolve is incremental, because it adapts to new data, increasing the effectiveness of its learning ability. More important, the EFuNN can learn spatiotemporal data adaptively, in a single pass per learning session [9].

The EFuNN [10] is a connectionist structure, based on fuzzy rules and inference implemented in a five-layer architecture. Connections are created and fixed as labeled input data is presented at the input layer.

The EFuNN architecture has five layers (Fig. 1), the first being the input layer. The second layer implements a fuzzy quantification of the input data, according to fuzzy measurement criteria ("small," "medium" or "large"). The third layer represents the rules through network nodes. Such nodes evolve by supervised or unsupervised learning. The rules are feature models of the input data, so this layer functions as a feature extractor to classify input data into the appropriate domain.

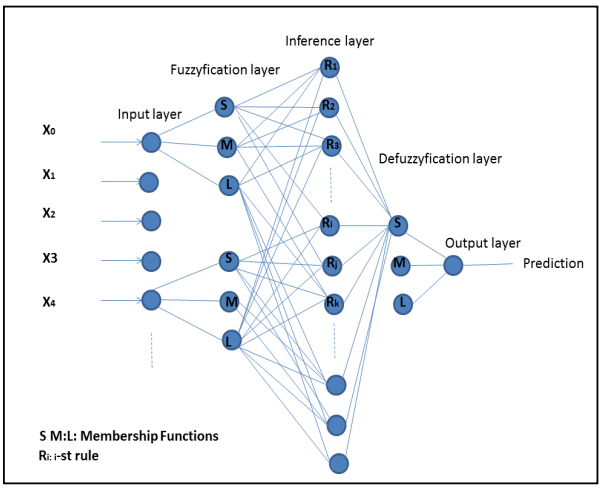

Fig. 1. EFuNN is a softcomputing paradigm that have the ability to evolve incrementally, because it adapts to new data.

The fourth layer takes care of defuzzifying the output data (the inverse operation of the fuzzification at the input layer). The method applied for defuzzification consists of a weighted function and a saturated, linear-activation function that produce the appropriate output (prediction).

B. *Training and testing the data set*

The data set devised to train the EFuNN to learn to recognize single isolated words by matching their phonemes consists of a set of sampled and labeled phonemes derived from the phonetic transcription of the word *hello* (Fig. 3) and the word *fly* (Fig. 4). For training purposes, the phone set is (artificially) machine synthesized (fig. 2)and for test purposes the speech is recorded, naturally produced utterance (Fig. 3, 4).

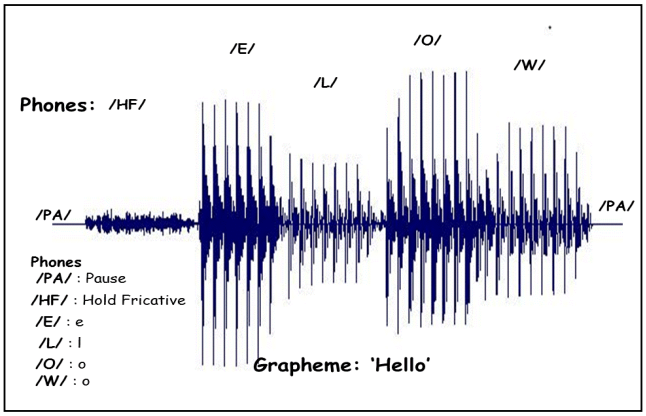

Fig. 2. Word *hello* generated by a formant synthesizer

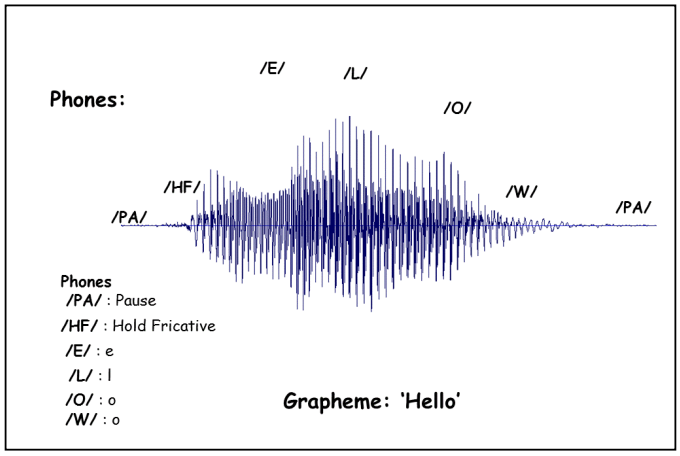

Fig. 3. Word *hello uttered by a haman being.*

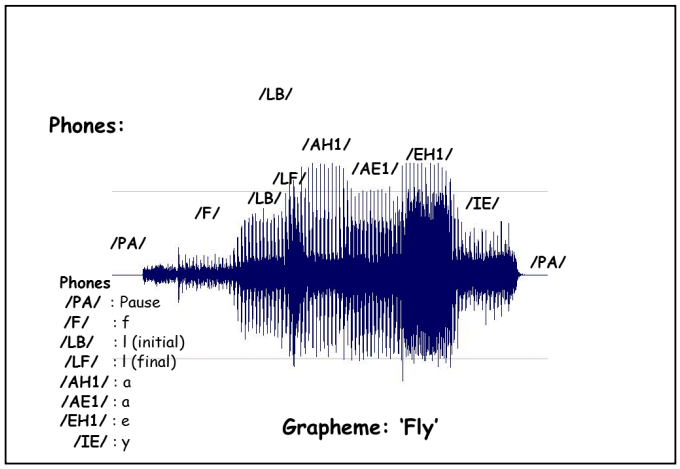

Fig. 4. Word *fly* uttered by a human being..

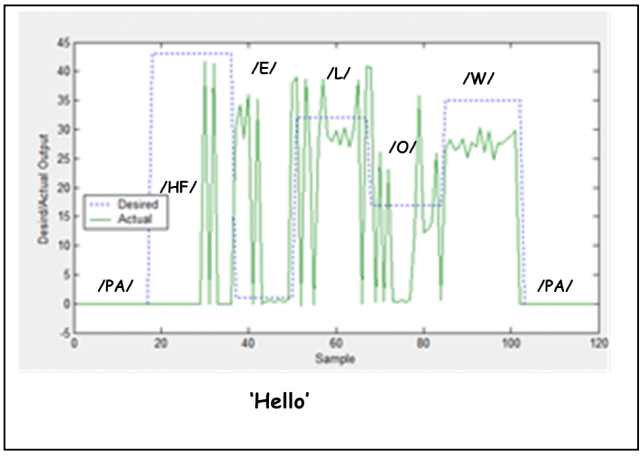

Fig. 5. Recognition of the synthesized word *hello* after training and one evolving step.

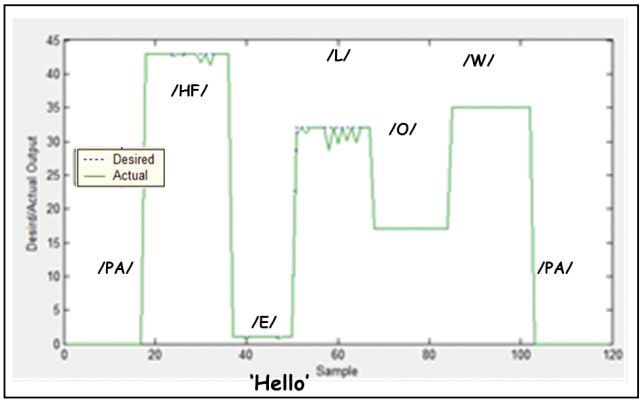

Fig. 6. Recognition of the synthesized word *hello* after training and few evolving steps.

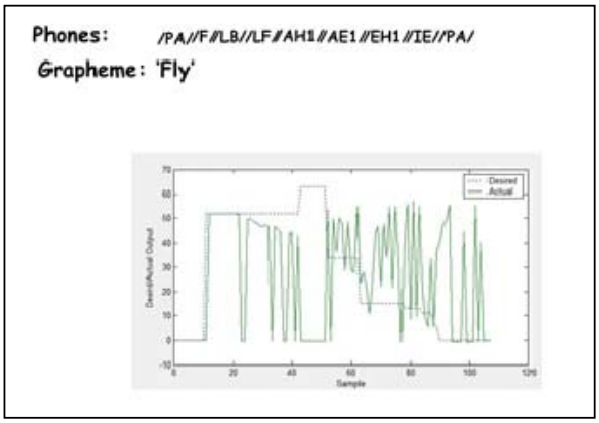

Fig. 7. – Recognition of the word *fly* after the training, without evolving steps

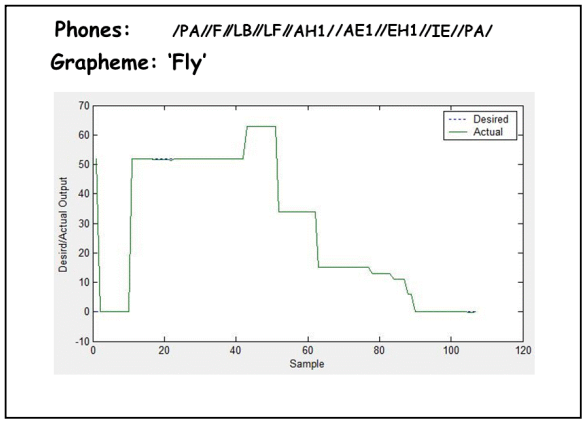

Fig. 9. Recognition of the word *fly* after 3 evolving steps.

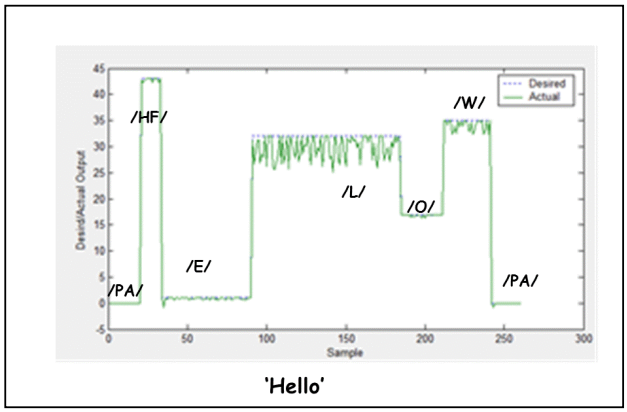

Fig. 8. Recognition of the synthesized word *hello* after several evolving steps.

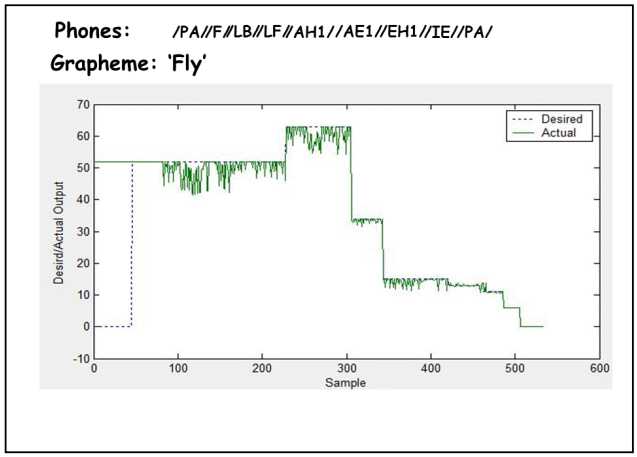

Fig. 10. RecognitIon of the word *fly* after several evolving steps.

The NeuCom [11] environment was used to model and simulate the EFuNN by applying the following setup:

- Sensitivity threshold: 0.95
- Error threshold: 0.05
- Number of membership functions: 5
- Learning rate for W1: 0.1
- Learning rate for W2: 0.1
- Node age: 60.

The phonemes are synthetically generated by a formant speech synthesizer. The word *hello* for testing purposes was uttered by a human being.

After the first training step and one evolving step, the EFuNN was barely able to recognize the correct phoneme sequence (Fig. 5). But, after few more evolving steps , it proved quite able to recognize and classify the phonemes in the word *hello (fig.6)*. The root mean square error dropped to very low values (from 19.89 to 0.52 RMSE). No phoneme mismatch occurs at the recognition stage .(Fig. 6). The same goes for the word *fly* (Fig. 7, 9, 10).

C. Modeling Improvements

After isolated word recognition, the EFuNN was successfully tested for its ability to match any phone. A complete set of English phones, with a few German and French ones as well (64 distinct phones in all), was sampled and labeled to train the EFuNN.

To mimic the cochlea's strategy of dedicated neurodetectors, one neurodetector per frequency (a hair cell), we trained one EFuNN for each phoneme (Fig. 7). This reduced EFuNN complexity (minimizing the number of rules, on the order of tens) and improved test results. This strategy led to a system of 64 parallel EFuNNs, each specialized in recognizing a single phone in a time-domain-sampled sequence (fig.11).

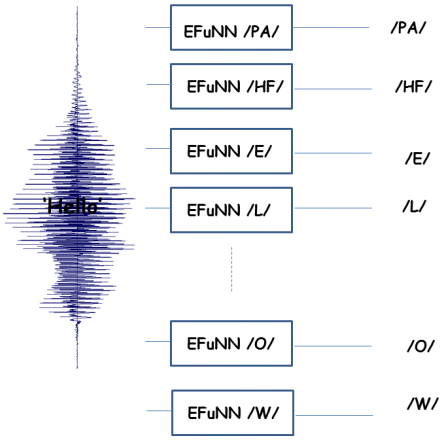

Fig. 11. Architecture of the parallel running dedicated EFuNNs , individually trained to match a single phone of an uttrered word.

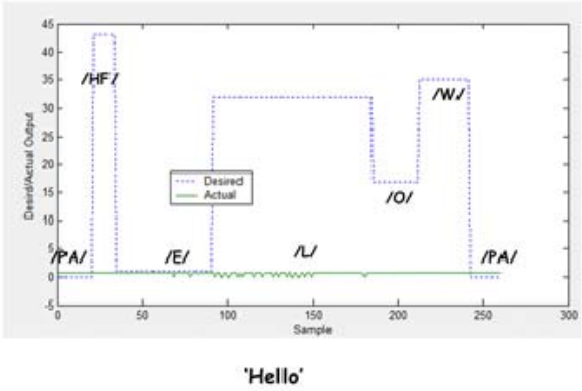

Fig. 11. Matching the phone /E/ and the two pseudoophonemes /PA/ preceeding and following the coda (attack and decay) of the uttered word *hello* with one of the parallel running EFuNNs, the one trained for /E/.

Upon testing, the following performance was achieved (Fig. 8) for the word *hello*: each phone in the uttered word *hello* was successfully predicted by the specific EFuNN trained for that phone. Similar performance was achieved for the word *fly* (fig. 10).

III. Observations on Results of Experiment

The results of experiments confirm that the biomorphic approach to solving the ASR challenge is applicable to a complete softcomputing (ZISC) implementation of an ASR and, even in the above referred multimodal, audiovisual, framework [6]. Careful attention was paid to developing the data set, since the performance of softcomputing paradigms is largely dependent on the learning process. This research demonstrated that learning from time-domain-sampled signals is feasible if a powerful softcomputing paradigm such as EFuNN is applied. This mean that the hardcomputing stage in an AVASR framework can be completely replaced by a softcomputing stage to make its implementation as ZISC feasible.

A. Phonems and phones

For these experiments we applied a set of 64 distinct synthesizable phones that can be combined to reproduce the phonemes needed for the language. Recognition of many phonemes involves multiple items from the phone list. While this obviously applies in the case of diphthongs, which often include phones appearing as glides, it is less obvious for a few articulatory features that make up part of the phone set. These include the /PA/ mentioned above, a time segment with no

vibration, and various forms of closure or aspiration. Closures are distinguished among glottal stops, fricative closure, and other vocal-tract closure. In addition to the /h/ aspiration that is phonemic in the hat-at distinction and allophonic in English syllable-initial stops, distinct nasal and vocal-tract aspirations were identified as specific phones.

REFERENCES

[1] K. Yongtae, Y. Zhang, "A Reconfigurable digital neuromorphic processor with memristive synaptic crossbar for cognitive computing", ACM Journal on Emerging Technologies in Computing Systems, Vol. 11, No. 4, Article 38, April 2015

[2] N. Kasabov, K. Dhoble, N. Nuntalid, G. Indiveri, "Dynamic evolving spiking neural networks for on-linespatio-and spectro-temporal pattern recognition", Neural Networks, n. 41, (2013).

[3] .M.Ferrandez, J.R.A,. Sanchez, F. de la Paz, F.J. Toledo (Eds), "New Challenges on Bioinspired Applications", 4th International Work-Conference on thelinterplay Between Natural and Artificial Computation , IWINAC 2011, May-June 2011, Proceeedings, Part 2, LNCS6687 Springer, 2011.

[4] H. Bourlard, N. Morgan, "Connectionist, speech recognition: a Hybrid approach", Kluwer Academic Publishers, Boston, Mass., 1994.

[5] R. F. Lyon, C. Mead, "An analog Electronic cochlea", IEEE Trans. Acoust., Speech, Signal, Processing.36: 1119-1134, July 1988.

[6] M. Malcangi, H. Quan, "Bio-inspired audio-visual speech recognition – towards the zero instruction set computing", Communications in Computer and Information Science (CCIS), In Engineering Applications of Neural Networks, C. Jayne, L. Iliadis (Eds.), Springer, Switzerland, PP. 326-334, (2016

[7] S. Sutton, B. Barto, "Reinforzed Learning: an Introduction", Adaptive computation and Machine Learning, MIT press, 1998.

[8] Kasabov, N.: Evolving Connectionist Systems: The knoledge engineering approach.Springer, Heidelberg,2007.

[9] N. Kasabov, "EFuNN", IEEE Tr SMC, 2001.

[10] N. Kasabov, "Evolving fuzzy neural networks – algorithms, applications and biological motivation," In: Yamakawa and Matsumoto (eds.), Methodologies for the conception, design and application of the soft computing, World Computing, pp. 271-274, 1998.

[11] http://www.kedri.aut.ac.nz/areas-of-expertise/data-mining-and-decision-support-systems/neuco.

[12] N. Siddique, H. Adeli: Computational intelligence, Synergies of Fuzzy logic, Neural Networks and Evolutionary Computing. Springer, Switzerland, 2016.