Bio-inspired Audio-Visual Speech Recognition Towards the Zero Instruction Set Computing

Mario Malcangi(✉) and Hao Quan

Department of Computer Science, Università degli Studi di Milano, Milan, Italy
{malcangi, quan}@di.unimi.it

Abstract. The traditional approach to automatic speech recognition continues to push the limits of its implementation. The multimodal approach to audio-visual speech recognition and its neuromorphic computational modeling is a novel data driven paradigm that will lead towards zero instruction set computing and will enable proactive capabilities in audio-visual recognition systems. An engineering-oriented deployment of the audio-visual processing framework is discussed in this paper, proposing a bimodal speech recognition framework to process speech utterances and lip reading data, applying soft computing paradigms according to a bio-inspired and the holistic modeling of speech.

Keywords: Audio-visual information processing · Automatic speech recognition · Bio-inspired computing · Convolutional neural networks · Evolving fuzzy neural networks

1 Premises

Automatic speech recognition (ASR) will be a key technology for the next generation on information systems, when human-to-machine interaction will be similar to the human-to-human interaction. Experiments to understand speech perception began last century. In 1921, Fletcher and Stainberg had found a functional relations between nonsense's phone sequences (e.g. consonant-vowel-consonant) error-recognition rate and words' recognition rate. This relation demonstrated that the context influences the intelligibility. Allen in his work "How do humans process and recognize speech?" [1] discusses extensively the role of the context in human speech recognition (HSR), citing the famous example of the two questions "How do human recognize speech?" and "How do humans wreck a nice beach?" that can be uttered so that only with appropriate context they can be distinguished. Entropy is higher for simple sounds (phones) and lower for complex words, so two important strategies are in HSR.

2 Introduction

Audio-visual information processing (AVIP) is an interdisciplinary research field that joins computer science and signal processing. It concerns the processing of information that is embedded in physical signals generated by the human beings and by the

C. Jayne and L. Iliadis (Eds.): EANN 2016, CCIS 629, pp. 326–334, 2016.
DOI: 10.1007/978-3-319-44188-7_25

surrounding environment. Most of the research efforts have been targeted audio and visual as individual fields, considering these fields as independent at each other. An emblematic example of this is the ASR problem approached mostly as an audio processing special purpose task. Several investigations [2–9] demonstrated that speech understanding in human beings is a multimodal process where the audio and the visual information concur to successfully complete the correct recognition of communication sounds such as phonemes, phones and words.

The AVIP activity in human beings is not perfect but efficient. This is because it is a biological-based processing model, with evolving inference paradigms performing in adaptive and context aware way. The multimodal nature of both audio production and perception and the relationship between audio and visual information has been investigated and experimental results demonstrates that the bio-inspired approach to the issue of AVIP could be the right way to develop robust and effective AVIP-based applications [10, 11].

There are also several bio-inspired processing processes that are under considered in the development of the ASRs, such as localized time-frequency events, temporal and spatial information (binaural), pitch (for source localization and separation). These processes needs to be considered in order to match the right paradigm to be applied for the ASR development.

Two main bio-inspired soft computing paradigms nicely match audio and visual perception in human beings, the convolutional neural network (CNN) and the evolving connectionst systems (ECOS).

The CNN, a bio inspired variant of the multilayer perceptron (MLP), has been successfully applied to face recognition [12]. CNN embeds the convolution paradigm useful to model spatial and temporal correlations. This apply to speech signal to compensate the translational variance and to capture translational invariance with a reduced set of parameters [13].

CNNs exibit invariance to shifts of speech features along the frequency, dealing with speaker and environment variations. The CNN special network structure (alternation of convolutional and pooling layers) is the main advantage over standard neural network as it demonstrates to be compatible and efficient respect to the way the data can be arranged to be processed efficiently. Considering the voice spectrogram as a 2-D image of features distributed along the frequency and time axes, the same approach of the use of CNN for the image recognition can be extended to speech recognition.

ECOS [14], mainly the evolving fuzzy neural networks (EFuNN), is a bio inspired inference paradigm that meets the capability of the HSR to adapt to noise and the signal filtering by evolving. In a multimodal context such as the audio-visual integration at the higher layers of the HSR hearing model, EFuNN is able to fuse the decision from audio and visual stages [15].

3 Bio Inspired Framework for Audio-Visual Speech Recognition

The bio inspired framework (Fig. 1) for audio-visual speech recognition (AVSR) is a three stage system that apply three bio inspired inferencing (soft computing) paradigms, the convolutional, the evolving and the rule-based. The full framework is completed by two mixed-signal processing (hard computing) frontend and backend stages, to interface the framework by sensing and by actuating towards the physical world.

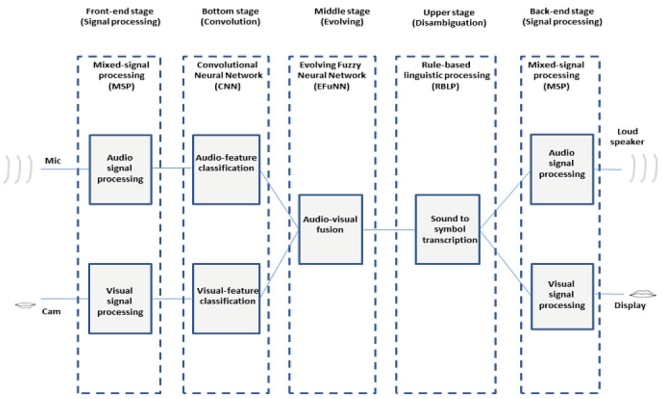

Fig. 1. Framework for the bio-inspired audio-visual speech recognition

3.1 Front End Stage

The front end is a mixed-signal processing (MSP) stage that implements the signal conditioning (linearization, amplification, equalization and filtering) and the extraction of the low level features (time and frequency measurements). It is based on analog and digital signal processing (mixed-signal) models that puts the crisp signal information in a measurement domain suitable to the lower information processing stages. Two distinct MSP front end are available, one for the audio signal, captured by a microphone, and one for the visual signal, captured by a camera.

3.2 Convolutional Stage

The convolutionalstage implements the high-level feature mapping (phoneme and viseme) task by exploiting the temporal and spatial local correlation of the audio and the visual information. Audio and visual low level features from front end stage are inputted to the convolutional layer. This stage consists of two information path, one for audio information and one for visual information. The purpose is to feature the audio and visual information according to the semantic that will be applied at higher stages

(e.g. phonemes and visemes featuring in AVSR systems). Each node of the input layer is connected to the inner layer nodes in a spatially contiguous receptive schema (e.g. each node at layer *n* is connected to only 3 adjacent *n-1* layer nodes) (Fig. 2).

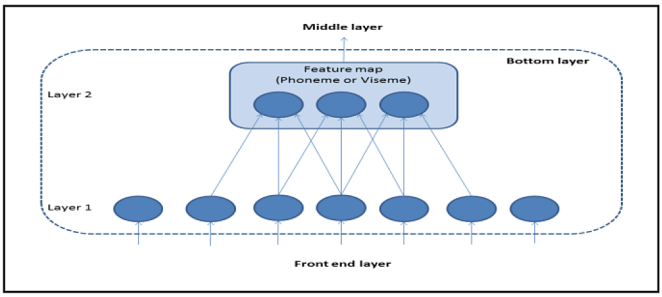

Fig. 2. Feature mapping at convolution stage

3.3 Evolving Stage

The evolving stage implements decision fusion on the audio and visual high-level feature scores by applying the evolving paradigm to enable the adaptation of the AVSR system to the environment variability (e.g. noise) and to the information mismatch (e.g. mismatch of /m/ and /n/ phonemes due to high degree of similarity of time and frequency features).

The evolving stage is implemented by the EFuNN paradigm (Fig. 3).

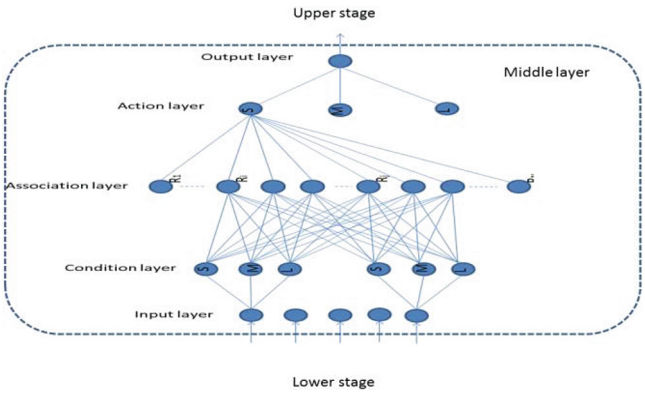

Fig. 3. EFuNN evolving architecture is applied to fuse phoneme-viseme classification and to predict phoneme occurrence

3.4 Linguistic Stage

The linguistic stage implements by rules the process of sound to symbols conversion (e.g. phoneme to grapheme) and of the text disambiguation.

3.5 Back End Stage

The back end is a mixed-signal processing (MSP) stage that implements the signal conditioning (linearization, amplification, equalization and filtering) and physical feature generation to be applied to the physical world (e.g. audio-visual speech synthesis). It is based on digital and analog signal processing (mixed-signal) models that produce the crisp signal information in a measurement domain suitable to be applied to other systems (control, communication, decision, etc.). Two distinct MSP back end are available, one for the audio signal, played by a loudspeaker, one for the visual signal, visualized by a display.

4 Experimental Tests

Some experiments have been executed to test the bio inspired AVSR's ability to adapt to physical context changes (e.g. noise). The test concerned the recovering of the right grapheme from the utterance of a words with two acoustically similar phonemes, /m/ and /n/, and their corresponding visemes.

4.1 Front-End

Two front-end has been programmed to extract the physical features of the captured signal by an audio sensor (microphone) sampled at 16 kHz and 16 bit encoded, and by a visual sensor (camera) 24 fps. A set of digital signal processing (DSP) algorithms has been applied for feature extraction purpose from the audio signal:

- Five frequency bands (tonotopically) ordered onto space (cochlea-like) the feature extracted from the visual signal are:
  - Lips eight (LE)
  - Lips width (LW)

4.2 Convolution Stage

At the convolution stage, the audio the features from the front-end are inputted to the ANNs separately-trained to classify the phonemes. The ANNs scores the phonemes on a frame-by-frame time base, synchronously to the visual framing (2 audio frames (21 ms) for each visual frame (42 ms)). The score is the input for the middle stage (evolving stage). The visual features (visemes) are encoded by LE and LW measurements and passed directly to the middle stage as knowledge related to the current viseme (Fig. 4).

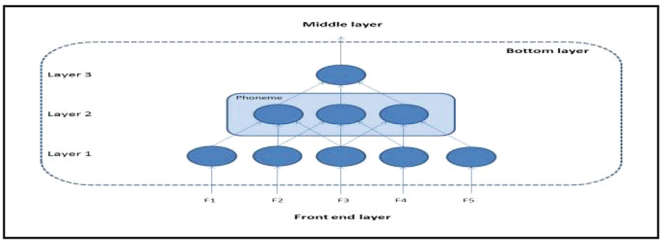

Fig. 4. Phoneme featuring at convolution stage.

4.3 Evolving Stage

At the second stage an evolving fuzzy neural network (EFuNN) has been trainedted to execute the decision fusion of the scoring of the audio and of the visual features. The EFuNN has been trained to predict the phoneme sequence frame-by-frame, streaming the phonetic transcription of the spoken word at output.

The NeuCom [16] 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

4.4 Linguistic Stage

At the third stage the fuzzy logic engine disambiguates the phonetic transcription executing the phoneme-to-grapheme transcription.

4.5 Test Setup

The test has been executed on 100 utterances of phonemes /m/ and /n/under four acoustic changing conditions (increasing additive white noise):

1. 0 dB
2. +6 dB
3. +12 dB
4. +18 dB

The EFuNN was first trained with 80 % of the utterances at 0 dB noisy condition and tested with the remaining 20 %. Then at next test time, the EFuNN evolved using 80 % previous test utterances and fuses decisions using the new noisy classification from audio and visual featured utterances.

4.6 Test Results

Test at 0 dB (Fig. 5) demonstrates good performance in discriminating two similar phonemes /m/ and /n/, not when the audio scoring fails at audio convolution layer (/m/ / n/ mismatch). After evolving, the test (Fig. 6) demonstrates its ability to recover the /m/ /n/ mismatch. The performance fall down when +18 dB additive noise masks the noise-free utterance. After evolving, the test (Fig. 7) demonstrates better performance with noisy utterances.

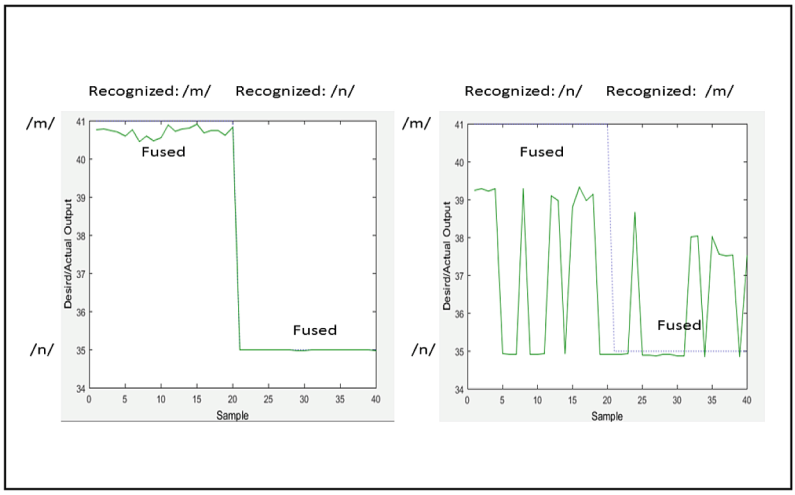

Fig. 5. Decision fusion at evolving stage for /m/ /n/ sequence for correct recognition (left) and for wrong recognition (right).

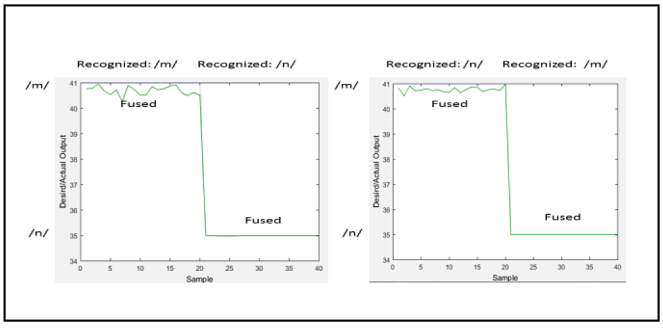

Fig. 6. Decision fusion at evolving stage for /m/ /n/ sequence for correct recognition (left) and for wrong recognition (right) after training on wrong recognition.

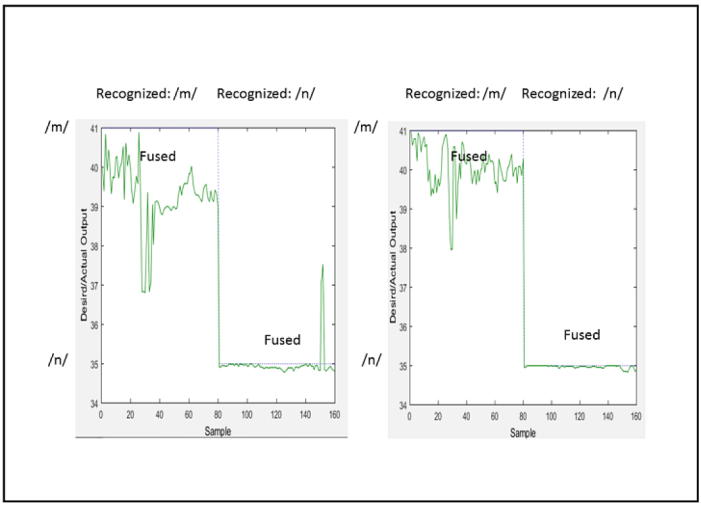

Fig. 7. Decision fusion at evolving stage for /m/ /n/ sequence for noisy input: without adapting (left) and with adapting trought evolving (right).

5 Conclusion and Future Works

The experiment demonstrated that the bio-inspired AVIP framework is effective in harsh conditions keeping low the system complexity. This performance is related the special purpose nature of the subsystems and of their capability to adapt to context changes by data-driven paradigms and intrinsic evolving capabilities.

The purpose of this research is to find which bio-inspired processing and inferencing paradigms could be optimal for the complete computational path from sensing to actuation in audio-visual applications. Bio-inspired signal processing is the next step to extend the paradigm to the front end and the back end of the AVIP framework.

References

1. Allen, J.B.: How do humans process and recognize speech? IEEE Trans. Speech Audio Process. **2**(4), 567–577 (1994)
2. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**, 746–748 (1976)
3. Massaro, D.: Speech perception by ear and eye: A paradigm for psychological enquiry. Erlbaum, London (1987)
4. Norrix, L.W., Green, K.P.: Auditory-visual context effects on the perception of /r/ and /l/ in a stop cluster. J. Acoust. Soc. Am. **99**, 2951 (1996)
5. Bernstain, L.: Visual speech perception. Audio Visual Speech Processing, pp. 21–39 (2012)
6. Cappelletta, L., Harte, N.: Phoneme-to-viseme mapping for visual speech recognition. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods. (IEEE) (2012)

7. Kazemi, A., Boostani, R., Sobhanmanesh, F.: Audio visual speech source separation via improoved context dependent association model. EURASIP J. Adv. Sig. Process., 47 (2014)
8. Vigliocco, G., Perniss, P., Vinson, D.: Language as a multimodal phenomenon: implications for language learning, processing and evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. **369**(1651), 20130292 (2014)
9. Sainath, T.N., Mohamed, A., Kingsbury, B., Ramabhadran, B.; Deep convolutional neural networks for LVCSR. In: Proceedings of ICASSP, May 2013
10. Wysoski, S.G., Benuskova, L., Kasabov, N.: Adaptive spiking neural networks for audiovisual pattern recognition. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 406–415. Springer, Heidelberg (2008)
11. Zouhir, Y., Ouni, K.: A bio-inspired feature extraction for robust speech recognition. SpringerPlus **3**, 651 (2014)
12. Lawrence, S.: Face recognition: a convolutional neural- network approach. IEEE Trans. Neural Netw. **8**(1), 98–113 (1997)
13. Abdel-Hamid, O., Mohamed, A., Jiang, H., Penn, G.: Applying convolutional neural network concepts to Hybrid NN- HMMModel for speech recognition. In: Proceedings of ICASSP (2012)
14. Kasabov, N.: Evolving Connectionist Systems: The knoledge engineering approach. Springer, Heidelberg (2007)
15. Malcangi, M., Grew, P.: Evolving fuzzy-neural method for multimodal speech recognition. In: Iliadis, L., et al. (eds.) EANN 2015. CCIS, vol. 517, pp. 216–227. Springer, Heidelberg (2015). doi:10.1007/978-3-319-23983-5_21
16. http://www.kedri.aut.ac.nz/areas-of-expertise/data-mining-and-decision-support-systems/neuco