Evolving Systems (2017) 8:85–94
DOI 10.1007/s12530-016-9156-6

ORIGINAL PAPER

Evolving connectionist method for adaptive audiovisual speech recognition

Mario Malcangi[1] · **Philip Grew**[1]

Received: 15 January 2016 / Accepted: 18 June 2016 / Published online: 7 July 2016

Abstract Reliability is the primary requirement in noisy conditions and for highly variable utterances. Integrating the recognition of visual signals with the recognition of audio signals is indispensable for many applications that require automatic speech recognition (ASR) in harsh conditions. Several important experiments have shown that integrating and adapting to multiple behavioral end-context information during the speech-recognition task significantly improves its success rate. By integrating audio and visual data from speech information, we can improve the performance of an ASR system by differentiating between the most critical cases of phonetic-unit mismatch that occur when processing audio or visual input alone. The evolving fuzzy neural-network (EFuNN) inference method is applied at the decision layer to accomplish this task. This is done through a paradigm that adapts to the environment by changing structure. The EFuNN's capacity to learn quickly from incoming data and to adapt while on line lowers the ASR system's complexity and enhances its performance in harsh conditions. Two independent feature extractors were developed, one for speech phonetics (listening to the speech) and the other for speech visemics (lip-reading the spoken input). The EFuNN network was trained to fuse decisions made disjointly by the audio unit and the visual unit. Our experiments have confirmed that the proposed method is reliable for developing a robust, automatic, speech-recognition system.

✉ Mario Malcangi
malcangi@di.unimi.it

Philip Grew
grew@di.unimi.it

[1] Department of Computer Science, Università degli Studi di Milano, Milan, Italy

Keywords Audiovisual speech recognition (AVSR) · Evolving fuzzy neural network (EFuNN) · Speech-to-text (STT) · Decision fusion · Multimodal speech recognition

1 Introduction

HAL, the computer in the movie "2001: A Space Odyssey," tells a crew member: "Dave, although you took very thorough precautions in the pod against my hearing you, **I could see your lips move."** This depicts how speech is understood multimodally because each perceptual function, by itself, distinguishes the utterance. When the auditory and the visual perceptual functions are active jointly, they fuse separate recognition results at a higher level to yield optimal speech understanding. Emotion also contributes to this fusion process, because muscles move during utterance, depending on mood and prosodic features.

We know that speech is more intelligible when the listener can see the speaker's face. This is due primarily to lip movement. Some phonemes may be confused in the audio domain (e.g.,*/m/*and*/n/*) but not in the visual domain (Fig. 1), where they correspond to distinct visemes. Phoneme-to-viseme mapping highlights this peculiarity in the human understanding of speech. Because there are fewer visemes than phonemes, the mapping between phoneme subsets and a given viseme is many-to-one. As a result, the visual interpretation of speech, in order to be effective, requires understanding speech acoustically. However, mere audio interpretation also proves inadequate for distinguishing similar sounds whose meanings differ, even when those sounds have different visemes.

There are various application contexts for automatic speech recognition (ASR): speech-to-text transcription in harsh, noisy conditions, natural interfaces in automobiles,

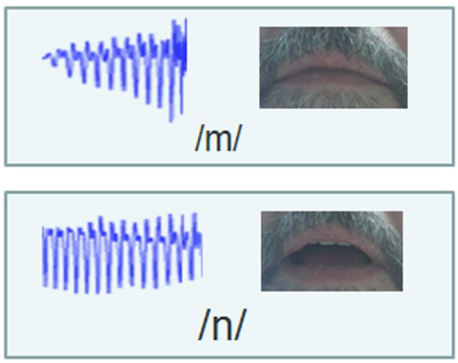

Fig. 1 Phonemes that may be confused in the audio domain (e.g.,*/m/* and*/n/*) might correspond to distinct visemes in the visual domain

access to systems for the visually or hearing impaired, automatic captioning of audiovisual streams, audiovisual data mining, natural user interfaces in handheld and wearable devices, etc. Such applications demand a multimodal approach to ASR but need implementations whose computational paradigms offer flexibility and adaptability.

Speech is the most effective way for human beings to communicate. It is also ideal between a human being and a machine. However, the performance gap between natural, human speech recognition and artificial, machine speech recognition is still huge, especially in noisy environments. Personal communication and information systems are rapidly evolving toward wearable systems. Consequently, the human–machine interface and the interaction paradigms need to feel natural and reliable, given the harsh operating conditions of these new, deeply embedded computing systems. Multimodal, speech-recognition systems will be one of the most important enabling technologies for such soon-to-come, deeply embedded systems (Kölsch et al. 2006; Marshall and Tennent 2013).

2 Related work for AVSR

The AVSR approach to ASR sytems from Petajan's work (Petajan 1984) is still an active research field. Its hot topics concern lip localization, feature extraction, and methods for fusing audio and visual information. Hidden Markov models (HMMs) were proposed first, followed by soft computing paradigms, mainly artificial neural networks (ANNs). A combination of HMMs and ANNs were applied by Noda (Noda et al. 2015), who experimented with a connectionist-hidden Markov model system for noise-robust AVSR. Dupont (Dupont and Luettin 2000) developed a sensor-fusion module responsible for joint modeling through time of acoustic and visual feature streams that uses multistream hidden Markov models (MHMMs). Kasabov (Kasabov 1998; Watts 2009) demonstrated that the evolving connectionist paradigm is well suited to the challenge of fusing auditory and visual information at the decision stage (Kasabov et al. 2000).

Recognizing speech presents challenges because speech signals vary greatly due to many factors related to how the utterance is produced (different speakers, different ways of speaking, different acoustics, different emotional states, different physical states, different ages, physiology changed by aging, etc). Speech is more than mere acoustics; it involves collateral visual communication (moving lips, facial expressions, and some body language), as well as explicit gesture (motion and movement of hands and arms). Two main challenges need to be addressed: adapting to variability and fusing multimodal data.

Multimodal speech recognition is founded on human beings' natural ability to communicate by integrating various sensory signals, while using context of situation to make decisions about what utterances they are hearing. For example, the McGurk and MacDonald experiment (McGurk and MacDonald 1976) showed that human beings combine auditory and visual information during interpretation (Wright and Wareham 2005). As a result, decisions about the meaning of a speech sound may differ according to situational audiovisual context. This experiment shows that audiovisual speech processing at recognition time enables two embedded capabilities, one for merger and one for combining. Merging and combining at the phoneme-recognition stage are powerful abilities that enable the AVSR to fuzzily remedy ambiguities in deciding on the most probable phoneme-to-grapheme transcription of the uttered speech. Implementing such fusion through soft computing proved effective, especially when the AVSR system uses two stages, a lower stage to extract features and an upper stage to fuse decisions (Noda et al. 2015; Malcangi et al. 2013; Patel et al. 2005; Stork et al. 1992).

Audiovisual speech recognition (AVSR) (Basu et al. 1999; Massaro 1996; Benoît et al. 1996; Salama et al. 2014; Bernstein and Auer 1996) enhances speech recognition by combining it with image recognition so that, e.g., heard utterances are supplemented by lip reading, which is especially helpful in harsh audio environments. The common approach to implementing AVSR systems is to merge audio features and video features into a single pattern-matching framework. This approach leads to highly complex systems that are hard to tune when they are actually up and running (Kaucic et al. 1996; Yang and Waibel 1996; Steifelhagen et al. 1997; Malcangi and de Tintis 2004). An alternative is to run the utterance-recognition system and the lip-reading system independently at the feature-matching stage and then fuse the decision at a later stage. Using such two-stage AVSR frameworks offers several advantages. First, it makes for more flexible and reliable AVSR. Second, it enables us to follow soft computing paradigms.

This not only lowers complexity but also enhances performance under harsh conditions, since it relies on fuzzy logic and neural networks.

The fuzzy-logic approach to AVSR is proposed by (Badura et al. 2014), who find it effective. But fuzzy logic has some drawbacks in terms of how it is to be modeled. For example, we need to establish how knowledge is to be developed (the rule set and membership functions). We need to choose a method for rule explosion. Various approaches have attempted to address these issues (Joo 2003), and knowledge development and rule explosions have both been efficiently optimized under the evolving paradigm (Kasabov 1998).

Methods from computational intelligence and techniques for adaptive machine learning have been successfully applied to AVSR, but certain problems with the evolving nature of the speech-recognition process remain open. One concerns the best choice of architecture to guarantee lifelong learning. Excessive training time is another important issue, given the real-time nature of the AVSR task.

Approaches that rely on an evolving connectionist system (ECoS) show promise for developing AVSR suited to highly variable phenomena. The simple evolving connectionist system (SECoS), a minimal implementation of ECoS, did a reasonably good job of recognizing isolated phonemes (Watts and Kasabov 2000). Its ability to learn and make generalizations was tested on the Otago Speech Corpus (Sinclair and Watson 1995), a body of segmented words representing 45 phonemes. The SECoS model's performance was evaluated compared to the more traditional connectionist structure, the multilayer perceptron (MLP), a model widely adopted in deploying ASRs and AVSRs, using the same datasets. SECoS outperformed MLP, showing good data recall and good adaptability to new data. The cost of this performance is seen in the large number of nodes in its hidden layer.

Because fuzzy neural networks are an optimal connectionist paradigm for modeling linguistic rules through the behavior of a process, we applied the evolving fuzzy neural-network (EFuNN) paradigm (Kasabov 2001) to implement the decision layer for a previously developed fuzzy-based AVSR (Malcangi et al. 2013). The purpose of that research was to develop an intermediate stage between the stage that matches phonemes to visemes and the stage that transcribes speech to text. This enables merger and combination to be completed before matching errors caused by noise are propagated to the stage that transcribes phonemes to graphemes.

The remainder of the paper is organized as follows. Section 2 describes related work on AVSRs. Section 3 presents the framework, the proposed evolving adaptive AVSR system architecture and the feature-extraction units. Section 4 discusses the fusion method, i.e., applying the EFuNN evolving architecture to fuse phoneme-viseme classification and predict phoneme occurrence. Section 5 describes experimental simulations and performance evaluation. Finally, Sect. 6 gives conclusions and future development.

3 Framework

Differently from the existing works we propose a new framework for adaptive speech recognition was defined and set up, according to the model for evolving intelligent systems (EIS). Among ECoS paradigms, we opted for EFuNN because of its ability to generate evolving rules that can be deployed in a fuzzy logic engine. Adaptation is driven by an analysis module that acts on the feature-decision layer, evaluating output from the decision-fusion layer and evolving in response to changes in surrounding context.

3.1 The proposed adaptive AVSR system architecture

The architecture for the proposed AVSR system consists of three feed-forward stages with feedback that enables its evolving functionality (Fig. 2). The first stage has three parallel operating units devoted to extracting and classifying features. This hard computing stage is based on (audio and video) digital signal-processing algorithms (DSP). The second stage, a soft computing fusion and combination unit, is based on a fuzzy logic engine (FLE). The third stage is a speech-to-text-transcription unit based on artificial intelligence (AI). The feedback goes through a transversal layer that exploits the EFuNN's capacity for quickly generating a set of evolved rules, which are applied to the fuzzy engine at runtime.

System input consists of audio and video streams. Audio information is captured by an array microphone (STMicroelectronics MEMS microphones). It is then conditioned and digitalized at 16 kHz/16 bit. Video information is captured by a camera recording at 24 fps. Application-specific software was developed to jointly capture audio and video on a frame-by-frame basis. The hardware and software setup for the audio-visual front-end is shown in Fig. 3. A MATLAB-based graphical user interface was developed to support system development and testing.

3.2 The feature-extraction unit

Phonemes and visemes (Table 1) are matched, scored, and encoded at the lower stage. Matching is based on signal-processing methods so as to classify, score, and store the utterance and its related visual sequence, frame by frame.

The feature-extraction unit consists of three distinct subsystems, one that processes the utterance (Fig. 4a) to

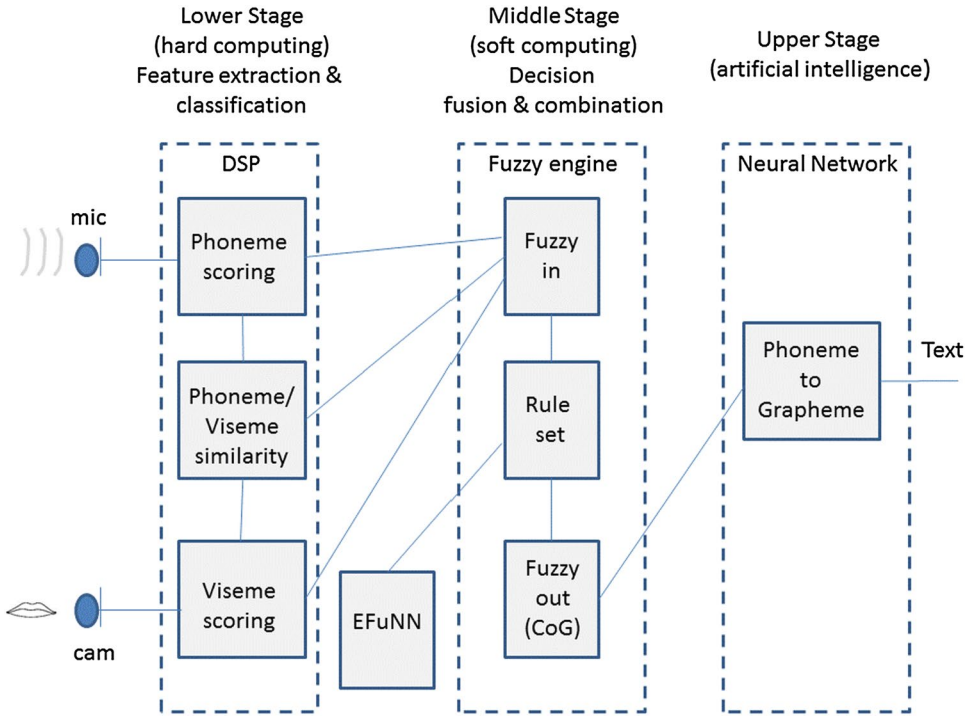

Fig. 2 Audiovisual speech-recognition system with decision layer based on fuzzy logic engine fed with rules tuned using EFuNN paradigm

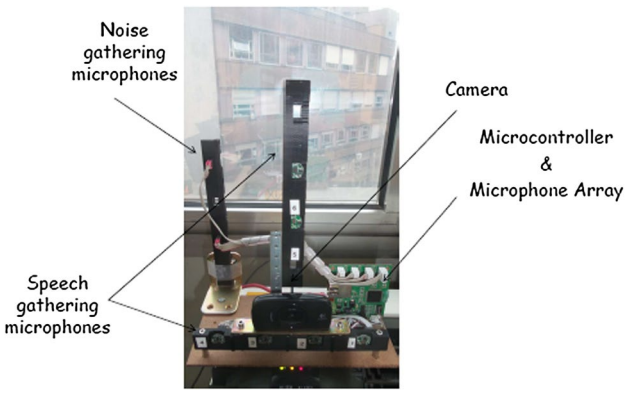

Fig. 3 Audio-visual front-end hardware setup

match phonemes, a second that processes video frames (Fig. 4b) to match visemes, and a third that measures the similarity of the matched phonemes and visemes. Phoneme and viseme matching units are independent systems. The similarity-scoring subsystem depends on the matching and scoring subsystems for phonemes and visemes.

The phoneme-extraction unit segments the audio stream into short intervals (20.85 ms), measures the features (pitch, formants, and intensity), and executes its classification (phoneme: score, duration).

The following features were used:

Root mean square (RMS):

$$RMS_j = \sqrt{\frac{1}{N}\sum_{m=0}^{N-1} s_j^2(m)} \tag{1}$$

m: sample number
N: total samples in a frame
s: frame
j: frame index
Zero-crossing rate (ZCR):

$$ZCR_j = \sum_{m=0}^{N-1} 0.5\left|sign(s_j(m)) - sign(s_j(m-1))\right| \tag{2}$$

Auto correlation (AC):

$$AC_j = \sum_{i=1}^{N}\sum_{j=1}^{N+1-i} s_j(j)s_j(i+j-1) \tag{3}$$

Cepstral linear prediction coefficients (CLPC):

$$CLPC_j = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \tag{4}$$

Frame-by-frame uttered speech (j) is encoded in feature vectors that are matched against a set of phoneme templates to classify and score the j[th] frame. Euclidean distance

Table 1 Viseme-class coding and the corresponding phoneme symbols

Viseme-classcode	Phoneme(s) symbol
0	Pause
1	ae, ax, ah
2	aa
3	ao
4	ey, eh, uh
5	er
6	y, iy, ih, ix
7	w, uw
8	ow
9	aw
10	oy
11	ay
12	h
13	r
14	1
15	s, z
16	sh, ch, jh, zh
17	th, dh
I8	f, v
19	d, t, n
20	K g, ng
21	p, b, m

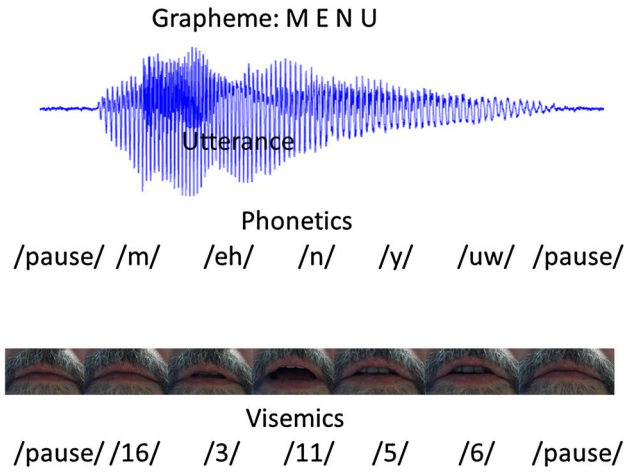

Fig. 4 Utterance of the word *menu*, its phonemic transcription, and corresponding visemes

metrics are applied to match each frame. Phoneme duration is measured as the number of contiguous audio frames (windows) that the current phoneme matches.

The viseme-extraction unit measures lip features (height, width, and duration) on each video frame (1/24 s) and yields a classification (viseme: score, duration). Visemes

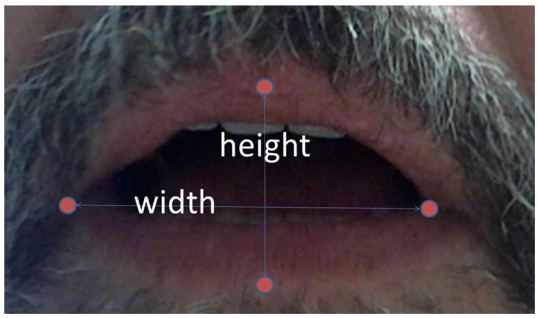

Fig. 5 After four key lip points were tagged, height and width are measured to index from the ratio of height to width

are identified by the height-to-width ratio of lip contour. To measure lip features, mouth contour is located after the face has been detected, whereupon lip position is determined. Four key lip points (Fig. 5) are pinpointed, two for vertical and two for horizontal, delimiting effective lip contour. Height and width are measured to create a relative index from the ratio of height to width. The viseme is then identified and scored, employing a matching method based on a set of templates and its Euclidean distance metrics. Viseme duration is then measured as the number of contiguous visual frames that the viseme matches.

The phoneme-viseme-comparison unit is a lookup table that scores how much the current phoneme is phonetically similar to the current viseme (Cappelletta and Harte 2012). The merge-and-combine unit predicts the phoneme on a window-by-window basis, running a fuzzy logic engine tuned according to the EFuNN paradigm. The output of this unit is a stream of phonemes (one phoneme per window) ready for phoneme-to-grapheme transcription. The phoneme-to-text-transcription unit applies a ruleset to transform each phoneme into the corresponding alphabetic representation, yielding the final text transcription of the uttered word (or a homophone). EFuNN-based feedback updates the fuzzy engine's rule set when changes in context occur (e.g., noise increases, new speaker, etc.) or errors are found at the higher transcription layer.

4 Evolving Fuzzy modeling of AVSR

The fuzzy logic-based inference paradigm was applied to draw inferences about phonemes from a set of audio and visual features, handling uncertainty due to noise and great variation in both audio and visual information. The main issue in designing the fuzzy logic engine was setting the rules. This was accomplished by applying the evolving neuro-fuzzy EFuNN paradigm, a neuro-fuzzy structure that evolves by creating and modifying its nodes and connections after learning from input.

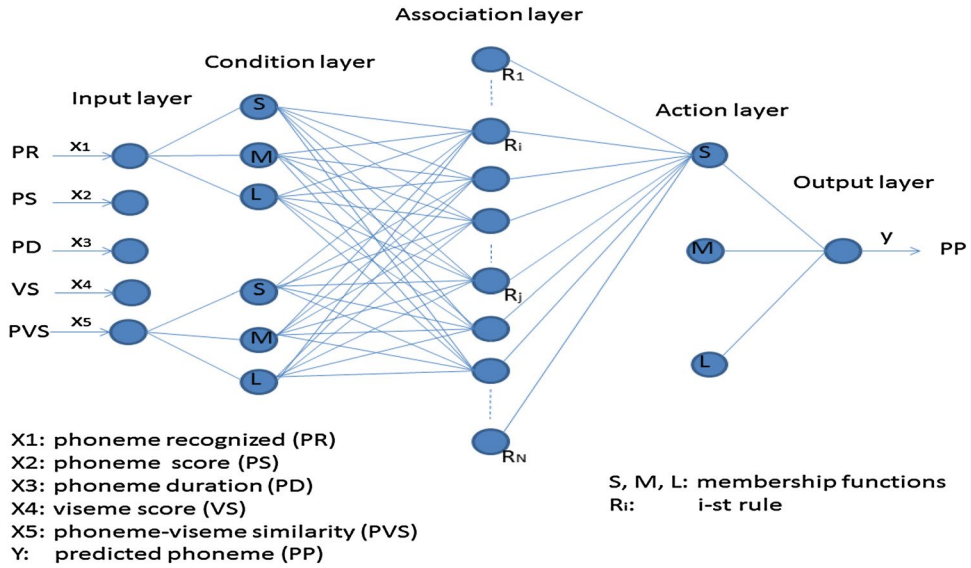

Fig. 6 EFuNN evolving architecture applied to fuse phoneme-viseme classification and to predict phoneme occurrence

The EFuNN paradigm connects using a feed-forward architecture of five layers of neurons and can be trained with neural-network methods (Kasabov 1998, 2001). By evolving it obviates the need to adapt to an a priori architecture, because it starts with a minimal set of initial nodes and then grows or shrinks at training and learning time, depending on its data input. This strategy avoids the problem of catastrophic forgetting and enables the network to be further trained with new data, retaining the effects of previous learning because new nodes are created without removing the old ones, thus preserving previous knowledge. Pruning and aggregation at training time avoid overtraining during learning by removing weak connections and their nodes.

The EFuNN is a five-layered, feed-forward, artificial neural network, in which each layer performs one specialized function in the fuzzy logic engine: input, condition, rule, action, and output (Fig. 6). The input layer (layer 1) represents (crisp) input variables that are presented to the nodes on the condition layer. The nodes on the condition layer (layer 2) are fuzzy membership functions that perform fuzzification on crisp input. The rule layer is the evolving layer (layer 3), which can create and aggregate the nodes, adapting them to changes in fuzzified input data. The nodes in this layer shape the rules that embed map the correspondence of input to output. The action layer (layer 4) consists of fixed-shape, fuzzy membership functions that fuzzily quantify output values. This layer computes the degree to which an output vector belongs to an output membership function (MF). The output layer (layer 5) defuzzifies the action output.

The layers perform their functions as follows:

- Layer 1: input (crisp values).
- Layer 2: condition (input membership functions).
- Layer 3: association (rules).
- Layer 4: action (output membership functions).
- Layer 5: output (crisp values).

The learning algorithm consists mainly of certain key actions, such as updating connections, aggregating nodes, pruning nodes, and extracting rules. At layer 3, the rule nodes cluster input–output data associations. Two connection weights, W1 and W2, are adjusted so that W1 is related to the fuzzified input vector and W2 is related to the corresponding output vector. To adjust W1, supervised learning based on output error is applied. To adjust W2, similarity is applied, using the cluster method.

To train and test the fuzzy engine, a sequence of pattern data was recorded from the output of the phoneme-extractor and viseme-extractor units. Data vectors *x(t)* of the input with the corresponding output were assembled to train and test the EFuNN. The data vector consists of five input measurements and one output. The five input measurements are: phoneme recognized (PR), phoneme score (PS), phoneme duration (PD), viseme recognized (VR), viseme score (VS), and phoneme-viseme similarity (PVS). The output item is the predicted phoneme (PP). Thus:

$$x(n) = [PR, PS, PD, VR, VS, PVS, PP] \tag{5}$$

$$n = t/Tw$$

The vector, indexed by time-window number n, is compiled throughout the utterance. The size of Tw is compatible with the quasi-stationary characterization of speech (from 20 to 40 ms). However, it must also be compatible with the duration of a visual frame (1/24 s, i.e., 41.7 ms). Hence, the time window was set to 20.85 ms, half the duration of a visual frame. The dataset was generated from basic uttered phoneme sequences (syllables), with and without added

background noise. It consists of 520 patterns x(n), one per frame, each fully describing the association between input and output. The dataset was split randomly to yield two data subsets, one with 80 % of the vectors for training, the other with 20 % for testing. The NeuCom (2016) environment was used to model and simulate the EFuNN by applying the following setup:

- Sensitivity threshold: 0.99
- Error threshold: 0.01
- Number of membership functions: 3
- Learning rate for W1: 0.1
- Learning rate for W2: 0.1
- Node age: 60

The sensitivity and error thresholds affect the generation of new rule nodes. If the sensitivity among inputs increases, then the network is more likely to create new rule nodes. If the threshold for error between actual output and desired output decreases, then the network is more likely to increase rule nodes. As the threshold increases, the network tends to retain its learning over a longer time. If pruning is on, i.e., the network's ability to remove connections between the layers while maintaining its original training performance, the network is less likely to reproduce a rule node that was pruned previously. The learning rate influences the training process. As the learning rate increases, the node will saturate faster, reducing its capacity to generalize. As the age threshold increases, the network's ability to retain what it has learned over the time increases. If aggregation is on, the network tries to aggregate the rules to form global behavior descriptions, thus avoiding increases in size that would make it unwieldy. The number and shape of membership functions depends directly on the dynamics of the input and output data and on how the functions are measuring data in the crisp domain. The more functions there are, the more the interconnections at the input and output layers.

5 Performance evaluation

The adaptability of the AVSR experimental setup was evaluated on the basis of the evolving functionality yielded by the EFuNN. Two sets of tests were conducted, the first to check the AVSR's ability to fuse the AV decision and the second to check adaptation through the evolving method. To run the first test, the word *menu*, with the right phoneme sequence, was first entered. Then, the same word with the phonemes/m/and/n/swapped was entered. To run the second test, environmental conditions for the utterance of the word *menu* were altered by adding audio noise. The EFuNN's performance was tested by checking its ability to recover from having confused similar phonemes in the two conditions, noise-free and noisy. The word was uttered and put into the AVSR forty times. The audio and the visual scores were collected to train the EFuNN by tuning the rules to be loaded into the fuzzy engine, then new input of forty utterances of word *menu* was been sent to the AVSR system. The fused decisions were presented graphically, grouping the forty utterances of the word *menu* by phoneme class.

5.1 Decision-fusion test

One hundred utterances of the word *menu* were uttered in noise-free conditions. The EFuNN was trained from scratch with the output from the audio-visual scoring layer, then tested. The results (Fig. 7) showed that the EFuNN's self-teaching ability is adequate to learn from data, recovering from confusion over similar phonemes (e.g.,/m/and/n/in the uttered word *menu*). The fuzzy engine was trained once (Fig. 7a) and twice (Fig. 7b) without any knowledge of the/m/-and-/n/phoneme mismatch. Then (Fig. 7c), the/m/-and-/n/phoneme mismatch was tested. The fuzzy engine learned how to fuse and combine phonemes and visemes recognized by independent audio and visual units.

The previously trained AVSR evolved by acquiring more knowledge about fusing audio and visual data. Its ability to recognize right the/m/-/n/phoneme sequence (Fig. 8a) and to recover from the mismatched one (Fig. 8b) improved.

5.2 Adaptation test

Additive noise (24 dB) was mixed, in linear fashion, into the utterance of the word *menu* and fed as input to the AVSR. No special noise-recovery strategy was applied at the (hard computing) lower stage. The EFuNN was allowed to evolve incrementally with the new set of noisy decision. The test sequence was then executed. The evolved ruleset for decision fusion was tested. Before evolving, the AVSR mismatched the/m/-/n/phoneme sequence (Fig. 9a, b). After evolving with new knowledge about the noisy conditions (24 dB noise), its recognition rate was quite good. It never confused the two similar phonemes/m/and/n/at 0 dB noise level (Fig. 10a, b). It performed only slightly less well at 24 dB noise (Fig. 10c, d), thus demonstrating its ability to evolve without forgetting.

6 Conclusion and future development

We proposed a new framework for adaptive speech recognition based on the model for evolving intelligent systems (EIS). We opted for EFuNN because of its ability to generate evolving rules that can be deployed in a fuzzy logic

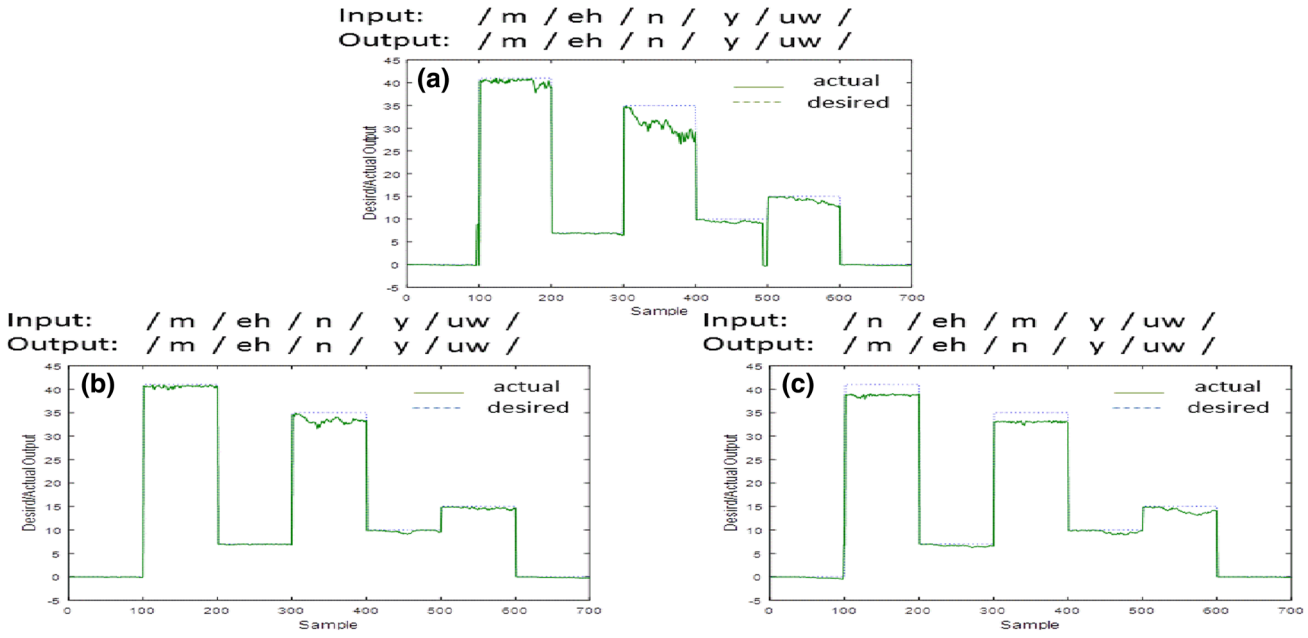

Fig. 7 Desired and actual output values of the trained by EFuNN for noise-free utterance of the word *menu*: the fuzzy engine was trained (**a**) once and (**b**) twice without knowledge of any/m/-/n/mismatch; then (**c**) the/m/-/n/phoneme mismatch was tested

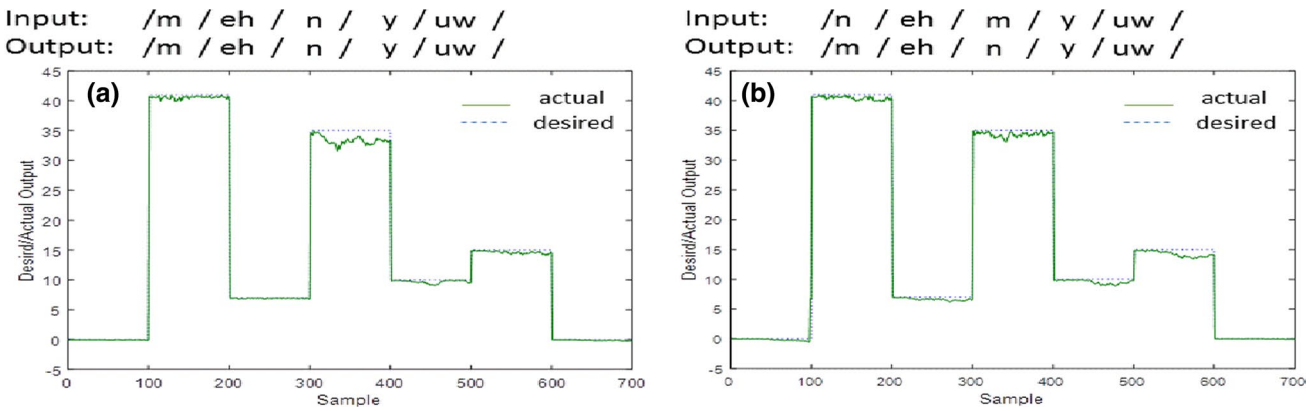

Fig. 8 The evolved AVSR system performed better at recognizing the right/m/-/n/phoneme sequence (**a**) and at recovering from the mismatched one (**b**)

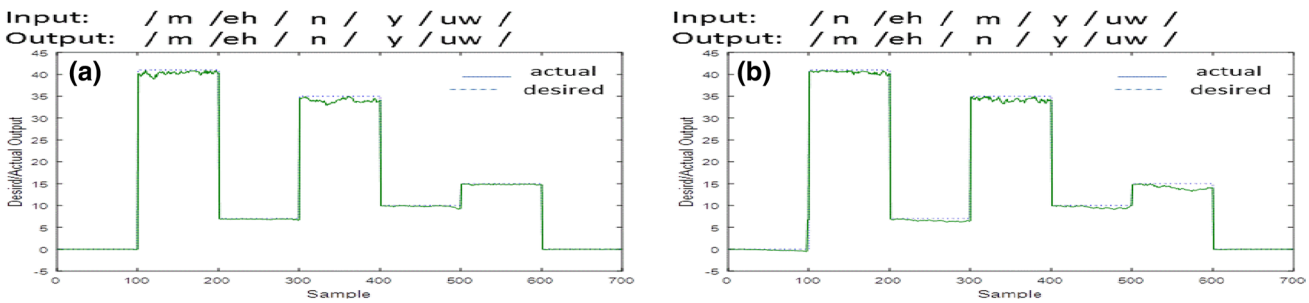

Fig. 9 Under noisy conditions, the AVSR performs badly at the feature and scoring layer

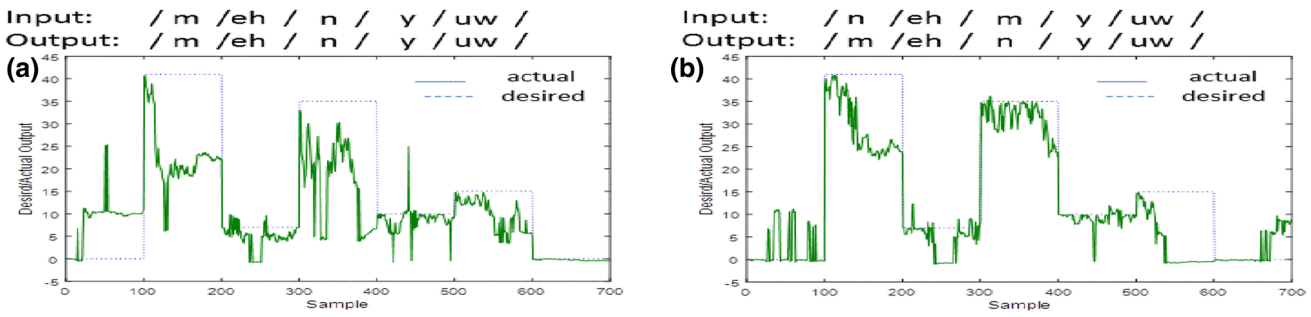

Fig. 10 After a first evolutionary step based only on the right/m/-/n/phoneme sequence under noisy conditions (24 dB noise), the AVSR systems performs very well **a**, **b** with 0 dB noise and well **c**, **d** with 24 dB noise

engine. Adaptation is driven by evaluating output from the decision-fusion layer and evolving in response to changes in surrounding context.

The fuzzy logic-based inference paradigm was applied to draw inferences about phonemes from a set of audio and visual features, handling uncertainty due to noise and great variation in both audio and visual information. The evolving neuro-fuzzy EFuNN paradigm, a neuro-fuzzy structure that evolves by creating and modifying its nodes and connections after learning from input, has been applied.

These results show that the evolving fuzzy neural-network (EFuNN) paradigm can be successfully applied to develop a fuzzy logic-based inference engine for merging and combining phonemes and visemes at the intermediate stages of a layered AVSR system. Several advantages were found, mostly in performance. These included an increase in reliability because system complexity was reduced.

Future development will focus on extending the evolving and adapting capabilities of the ECoS paradigm to the upper and lower stages of the AVSR system. One remaining issue is how to integrate the dynamic, evolving fuzzy neural-network paradigm into the AVSR in a proactive fashion. This would allow evolving capabilities to be embedded in the system. Another issue is how to apply the EFuNN paradigm to the system's lower layer, scaling it according to that layer's pattern-matching nature, and husbanding hard computing power for the important task of conditioning and feature extraction. Disambiguation is also a key issue that will affect the AVSR's upper layer at the phoneme-to-grapheme transcription stage and the syntactic transcription of the utterance.

Acknowledgments The audio and video hardware and software data acquisition setup was provided by STMicroelectronics. Special acknowledgement is due Dr. Claudio Marchisio for expertise on the visual components of the system and Dr. Roberto Sannino for expertise on the audio components.

References

Badura S, Frátrik M, Škvarek O, Klimo M (2014) Bimodal vowel recognition using fuzzy logic networks - naive approach. ELEKTRO, 2014, pp. 22–25, IEEE

Basu S, Neti C, Senior A, Rajput N, Subramaniam A, Verma A (1999) Audio-visual large vocabulary continuous speech recognition in the broadcast domain. In: IEEE Workshop on Multimedia Signal Processing. pp. 475–481

Benoît C, Guiard-Marigny T, Le Goff B, Adjoudani A (1996) Which components of the face do humans and machines best speechread? In: Stork DG, Hennecke ME (eds) speechreading by humans and machines: models, systems, and applications. Springer-Verlag, New York, pp 315–328

Bernstein LE, Auer ET Jr (1996) Word Recognition in Speechreading. In: Stork DG, Hennecke ME (eds) In speechreading by humans and machines: models, systems, and applications. Springer-Verlag, New York, pp 17–26

Cappelletta L, Harte N (2012) Phoneme-to viseme mapping for visual speech recognition. Proceeding of the 2012 International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012). (2012)

Dupont S, Luettin J (2000) Audio-visual speech modeling for continuous speech recognition. IEEE Trans Multimedia 2(3):141–151

http://www.kedri.aut.ac.nz. Accessed 6 July 2016

Joo MG (2003) A method of converting conventional fuzzy logic system to 2 layered hierarchical fuzzy system. In: Proceedings of the IEEE International Conference on Fuzzy Systems, pp. 1357–1362

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

Kasabov N (2001) Evolving fuzzy neural networks for online, adaptive, knowledge-based learning. IEEE Trans Syst Man Cybern B 31(6):902–918

Kasabov N, Postma K, Van den Herik EJ (2000) AVIS: a connectionist-based framework for integrated auditory and visual information processing. Info Sci J 123(1–2):127–148

Kaucic R, Dalton R, Blake A (1996) Real-time lip tracking for audio-visual speech recognition applications. Proc Eur Conf Comput Vision II:376–387

Kölsch M, Bane R, Höllerer T, Turk M (2006) Touching the visualized invisible: wearable AR with a multimodal interface. IEEE Computer Graphics and Applications, May/June 2006

Malcangi M, de Tintis R (2004) Audio based real-time speech animation of embodied conversational agents. In: Lecture Notes in

Computer Science, Vol. 2915, Gesture-based communication in human-computer interaction, pp. 429–430

Malcangi M, Ouazzane K, Patel P (2013) Audio-visual fuzzy fusion for robust speech recognition. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 582–589

Marshall J, Tennent P (2013) "Mobile interaction does not exist" in chi '13 extended abstracts on human factors in computing systems (CHI EA '13). ACM, New York, pp 2069–2078

Massaro DW (1996) Bimodal Speech Perception: A Progress Report. In: Stork DG, Hennecke ME (eds) In speechreading by humans and machines: models, systems, and applications. Springer-Verlag, New York, pp 79–101

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

Noda K, Yamaguchi Y, Hiroshi K, Okuno G, Ogata T (2015a) Audio-visual speech recognition using deep learning. Appl Intell Springer 42(4):722–737

Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015b) An asynchronous DBN for audio-visual speech recognition. Applied Intelligence 42(4):722–737

Patel P, Ouazzane K, Whitrow R (2005) Automated visual feature extraction for bimodal speech recognition. In: Proceedings of IADAT-micv2005. pp. 118–122

Petajan ED (1984) Automatic lipreading to enhance speech recognition. In: IEEE Communication Society Global Telecommunications Conference

Salama ES, El-khoribi RA, Shoman ME (2014) Audio-visual speech recognition for people with speech disorders. Int J Comput Appl 96(2):51–56

Sinclair S, Watson C (1995) The development of the Otago speech database. In: Kasabov K, Coghill G (eds). Proceedings of ANNES'95, IEEE Computer Society Press

Steifelhagen R, Yang J, Meier U (1997) Real time lip tracking for lipreading. In: Proceedings of Eurospeech

Stork DG, Wolff GJ, Levine EP (1992) Neural network lipreading system for improved speech recognition. In Proceedings International Joint Conf. on Neural Networks, vol. 2, pp. 289–295

Watts MJ (2009) A decade of Kasabov's Evolving Connectionist Systems: a Review. IEEE Transactions on Systems, Man and Cybernetics Part C. Applications and Reviews (2009) 39(3):253–269

Watts M, Kasabov N (2000) Simple evolving connectionist systems and experiments on isolated phoneme recognition. In: Proceedings of the first IEEE conference on evolutionary computation and neural networks, San Antonio, pp. 232–239, IEEE Press

Wright D, Wareham G (2005) Mixing sound and vision: the interaction of auditory and visual information for earwitnesses of a crime scene. Legal Criminol Psychol 10(1):103–108

Yang J, Waibel A (1996) A real-time face tracker. In: Proc WACV. pp. 142–147