Hybrid System for Robust Recognition of Noisy Speech Based on Evolving Fuzzy Neural Networks and Adaptive Filtering

Nikola Kasabov and Georgi Iliev
Department of Information Science, University of Otago, PO Box 56, Dunedin, New Zealand
nkasabov, giliev@infoscience.otago.ac.nz

Abstract - Speech and signal processing technologies need new methods that deal with the problems of noise and adaptation in order for these technologies to become common tools for communication and information processing. This paper is concerned with a method and a system for adaptive speech recognition in a noisy environment (ASN). A system based on the described method can store words and phrases spoken by the user and subsequently recognize them when they are pronounced as connected words in a noisy environment. The method guarantees system robustness in respect to noise, regardless of its origin and level. New words, pronunciations, and languages can be introduced to the system in an incremental, adaptive mode. The method and system are based on novel techniques recently created by the authors, namely: adaptive noise suppression, and evolving connectionist systems. Potential applications are numerous, e.g. voice dialing in a noisy environment, voice command control, improved wireless communications, data entry into databases, helping disabled people, multimedia systems, improved human computer interaction. The method and system are illustrated on the recognition of English and Italian spoken digits in different noisy environments.

Keywords: Speech recognition; Learning; Neural networks; Adaptive systems; Noise cancellation

1 Introduction

Speech recognition is one of the most challenging applications of signal processing [1]. Yet there is no noise-robust, adaptive, speaker-independent speech recognition system capable to maintain a medium, or a large vocabulary, available on the world market. The general problem addressed in this paper is developing adaptive systems working in a noisy environment typical for many applications (e.g. automatic speech recognition (ASR) in offices, vehicles, airplanes etc).

The paper offers solutions to the following specific research problems:

- The noise suppression problem, that is to design novel methods and systems for effective noise suppression in different environments. Examples are vehicle noise (cars, trucks etc.), aircraft noise, industrial noise, office noise.
- The adaptive speech recognition problem, that is the development of methods and systems for speaker-independent recognition with high accuracy, capable to adapt fast to new words, new accents, new speakers for a small-, medium-, to large vocabulary of words, phrases and sentences.
- The problem of integrated speech systems design capable to work reliably in severe noise conditions.

The noise cancellation problem is comparatively well explored [2]. There are methods available that are well situated for different types of noise. The research on the adaptive speech recognition problem is still in its infancy. It has been solved only for speaker dependent systems where the user adjusts the system to their voice.

Here the goal is to develop a method and a system that combine the advantages of several approaches in order to achieve an adaptive system efficient in a range of noisy environments.

The method and system developed and investigated in the paper are based on novel techniques that have been recently created by the authors, namely: adaptive noise suppression [3]; evolving connectionist systems for adaptive learning [4]-[7].

The paper presents first the framework of a noise-robust, adaptive speech recognition system. Then the methods for adaptive noise suppression and adaptive learning and recognition are explained. The paper also gives some experimental results achieved on a case study problem - the recognition of English and Italian spoken digits in different noisy environments. A comparative analysis is also presented when the suggested method and system are compared with other systems on the case study problem.

2 The Framework of the Methodology and the System for Adaptive Speech Recognition in a Noisy Environment

The framework of the proposed ASN methodology and a system is schematically shown in Fig.1. It consists of the following modules and procedures: adaptive noise suppression (ANS), endpoint detection (EPD), acoustic feature extraction (AFE), feature normalization with the use of the Discrete Cosine Transform (DCT) and speech recognition module that uses evolving fuzzy neural networks (EFuNNs). Other modules not shown in the figure are a temporal buffer for storing a sequence of recognized words and a phrase and sentence recognition module that follow the EFuNN module.

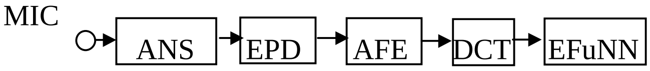


Fig. 1. Block diagram of the framework of the proposed ANS.

The methodology we propose here is based on the following principles:

- Speech recording and noise cancellation with the use of an original method suggested by the authors. The noise is effectively removed from the signal and a 'noise-free' signal is achieved. The method tolerates very low signal-to-noise ratio (close to 0).
- Transforming the 'noise-free' signal with the use of standard transformations tailored for the purpose. That includes: end point detection to determine the boundaries of each word; acoustic feature extraction to extract representative features from the signal; feature normalization to produce a fixed length vector that represents each spoken word and preserves the difference between the words. Here specific to the method, but controllable features and parameters are used.
- The vector that represents the pronounced word or phrase than is fed into a word-storage module created with the use of the general purpose adaptive learning method realized as an EFuNN. The EFuNN allows for adaptive learning. New words and phrases can be added or deleted from the system at any time of its operation, e.g. "go", "one", "connect to the Internet", "start", "end", "find a parking place". New speakers can be introduced to the system, new accents, new languages. In the recognition mode, when speech is entered to the system, the recognized words and phrases at consecutive time moments are stored in the temporal buffer.
- The temporal buffer is fed into an EFuNN-sentence module where multiple word sequences (or a whole sentence) are recognized.
- The recognized word or a sequence of words can be passed to an action module for execution depending on the application of the proposed method and system.

The signal processing part at the beginning is organized as follows: Short-time energy and zero-crossing rate are combined to detect the speech utterance boundaries. Acoustic features of the input speech are extracted over 20 ms frames. Hamming windows having an overlap of 10 ms are used to calculate Mel Frequency Scaled Cepstral Coefficients (MFSCC) and log-energy. A discrete cosine transform (DCT) is applied to the whole segment, retaining as many parameters as it is necessary.

3 Adaptive Filtering

Fig. 2 shows the classical scheme for adaptive noise cancellation using digital filter with finite impulse response (FIR). The primary input consists of speech s(n) and noise $n_2(n)$ while the reference input consists of noise $n_1(n)$ alone. The two noises $n_1(n)$ and $n_2(n)$ are correlated and $h_i(n)$ is the impulse response of the noise path. The system tries to reduce the impact of the noise in the primary input exploring the correlation between the two noise signals. This is equivalent to the minimization of the mean-square error $E[e^2(n)]$ where

$$e(n) = s(n) + n_2(n) – n_3(n). \quad (1)$$

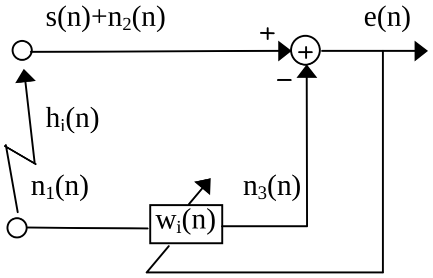


Fig. 2. Adaptive noise cancellation.

Having in mind that by assumption, $s(n)$ is correlated neither with $n_2(n)$ nor with $n_1(n)$ we have

$$E[e^2(n)] = E[s^2(n)] + E[n_2(n) - n_3(n)]^2. \qquad (2)$$

In other words the minimization of $E[e^2(n)]$ is equivalent to the minimization of the difference between $n_2(n)$ and $n_3(n)$. Obviously $E[e^2(n)]$ will be minimal when $n_3(n) \approx n_2(n)$ i.e. when the impulse response of the adaptive filter closely mimics the impulse response of the noise path.

The minimization of $E[e^2(n)]$ can be achieved by updating the filter taps $w_i(n)$. Most often normalized least mean square (NLMS) and recursive least square (RLS) algorithms are used [8].

4 Evolving Connectionist Systems and Evolving Fuzzy Neural Networks for Adaptive Learning and Classification

The proposed methodology and a system for adaptive speech recognition in a noisy environment uses a method for adaptive learning called evolving connectionist systems (ECOS) [4]-[7]. ECOS learn in a continuous, incremental way, adapt to changes in data, are self-organized, are controllable (i.e., through pre-setting appropriate values for certain parameters, certain level of learning and generalization can be achieved).

The ECOS method allows a system to be automatically created. The system consists of nodes (units) that perform pre-defined functions, and connections between them. The system has a minimal initial structure that includes preliminary input and output nodes and few preliminary connections. Data is allowed to flow into the system so that if an input data vector is associated with a desired output vector, the system stores this association into a new node and new connections. Nodes and connections are created automatically to reflect the data distribution. The system's structure is dynamically changing after each data item is introduced. The number of the input and output variables can vary during the learning process thus allowing for more (or less) input and output variables to be introduced at any stage of the learning process. Input and output variables can have 'missing values' at any time of the learning process. If there is no output vector associated with an input vector, the system produces its own output vector (its own solution). If the desired output vector became known afterwards, the system will adjust its structure to produce this output, or one close to it next time the same input vector is presented. The system continuously and adaptively learns from data to associate inputs to outputs and to cluster the data trough allocating nodes to represent exemplars of data.

The learning process in ECOS is achieved through interaction with the environment, which supplies the data flow and reacts to the output produced by the system. In addition, the system provides the knowledge it has learned in the form of IF-THEN rules. The ECOS method implies that a system evolves through its operation in an interactive way. The more data are presented to the system, the more the system evolves. The learning process is on-line, life-long.

Nodes and connections in an ECOS system can be created, modified, merged, and pruned, in a self-organizing manner, similar to how the human brain learns through creating and wiring neuronal structures. The system's structure grows or shrinks depending on the incoming data distribution and preset parameters. Through the process of evolving from data, the system learns the rules of its own behavior. The rules that constitute the system’s knowledge can be reported and/or extracted at any time of the system operation. In this way, an ECOS can be considered as a self-programming environment.

One realization of ECOS is the evolving fuzzy neural network (EFuNN) [4]-[7]. Nodes and connections are created/connected as data examples are presented. An optional short-term memory layer can be used through a feedback connection from the rule (also called, case) node layer (see for example [6]). The layer of feedback connections could be used if temporal relationships of input data are to be memorized structurally. The input layer represents input variables. The second layer of nodes (fuzzy input neurons, or fuzzy inputs) represents fuzzy quantization of each input variable space. For example, two fuzzy input neurons can be used to represent "small" and "large" fuzzy values. Different membership functions (MF) can be attached to these neurons (triangular, Gaussian, etc.). The number and the type of MF can be dynamically modified in an EFuNN. New neurons can evolve in this layer if, for a given input vector, the corresponding variable value does not belong to any of the existing MF to a degree greater than a membership threshold. A new fuzzy input neuron, or an input neuron, can be created during the adaptation phase of an EFuNN. The task of the fuzzy input nodes is to transfer the input values into membership degrees to which they belong to the MFs. The third layer contains rule (case) nodes that evolve through supervised/unsupervised learning. The rule nodes represent prototypes (exemplars, clusters) of input-output data associations, graphically represented as an association of hyper-spheres from the fuzzy input and fuzzy output spaces. Each rule node r is defined by its own values for the system parameters, and by two vectors of connection weights – W1(r) and W2(r), the latter being adjusted through supervised learning based on the output error, and the former being adjusted through unsupervised learning based on similarity measure within a local area of the problem space. The fourth layer of neurons represents fuzzy quantization for the output variables, similar to the input fuzzy neuron representation. The fifth layer represents the real values for the output variables.

An initial version of the EFuNN evolving algorithm is presented in [4]-[7]. The learning process includes also: aggregation of rule nodes (i.e. merging rule nodes); pruning of rule nodes and connections and other operations. The basic EFuNN algorithm, to evolve EFuNNs from incoming examples, is further developed and improved as a method for adaptive learning and self-optimization [9].

The next section describes a case study on the application of the ASN from Fig.1 built to recognize English and Italian spoken words with different sources of noise applied.

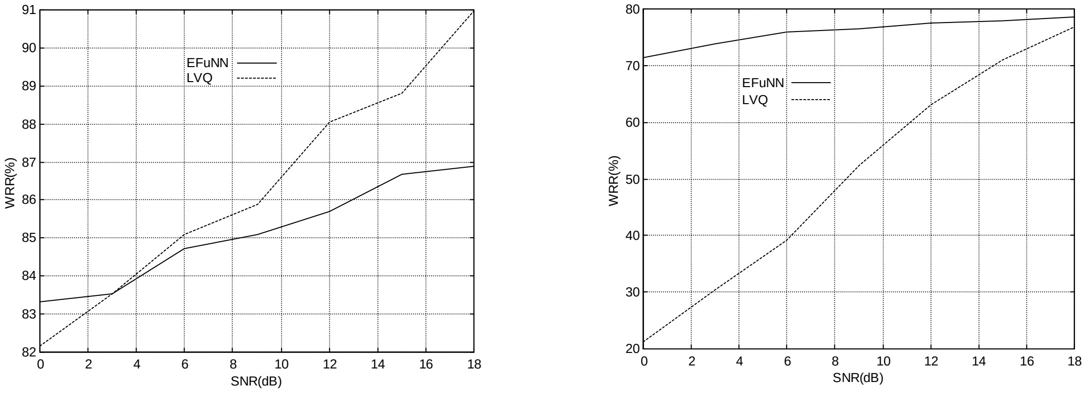
5 Experimental Results

The task is recognition of speaker independent pronunciations of English and Italian digits. English digits are from the Otago Corpus database (http://kel.otago.ac.nz/hyspeech/ corpus/). 17 speakers (12 males and 5 females) are used for training and other 17 speakers (12 males and 5 females) are used for testing. Each speaker utters 30 instances of English digits during recording session in a quiet room (clean data) for a total of 510 training and 510 testing utterances. Italian digits are from the SPK database, collected at ITC-Irst, Trento, Italy. 30 speakers (15 male and 15 female) for training and other 30 (15 male and 15 female) for testing, totally 6000 training and 6000 testing examples. For full details about this database see Ref. [10]. We use 8 Mel Frequency Scaled Cepstral Coefficients (MFSCC) and log-energy as acoustic features.

In order to assess the performance of EFuNN in this application, a comparison with Linear Vector Quantization (LVQ) [11] is accomplish. The clean training speech is used to train both LVQ and EFuNN. Noise is introduced in the clean speech to evaluate behavior of the recognition systems in a noisy environment. Two different experiments are conducted with the use of the standard EFuNN learning method [4]-[7]. In the first instance, car noise is added to the clean speech. In the second instance office noise is introduced over the clean signal. In both cases the SNR ranges from 0 dB to 18 dB.

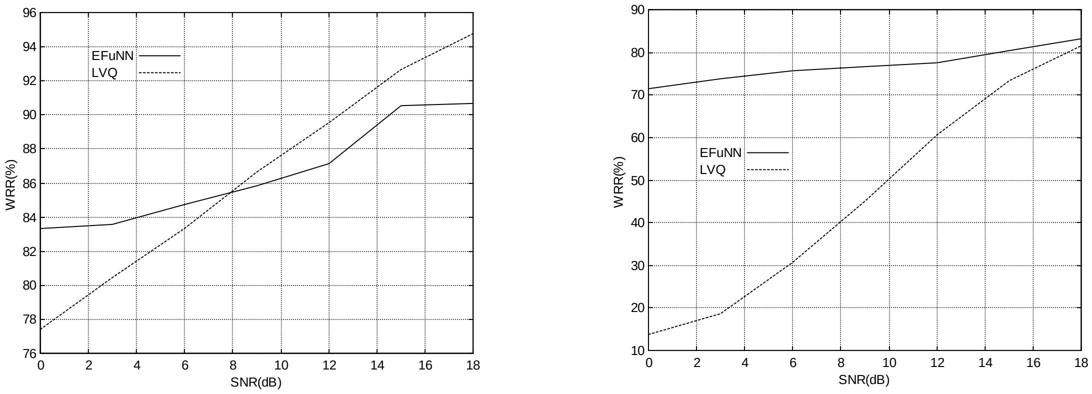
The results for car noise are shown in Fig. 3(a) – English digits and Fig. 4(a) – Italian digits. The word recognition rate (WRR) ranges from 86.87% at 18 dB to 83.33% at 0 dB – English digits and from 90.64% at 18 dB to 83.31% at 0 dB – Italian digits in EFuNN case outperforming LVQ, which achieves WRR=82.16% at 0 dB – English digits and WRR=77.43% at 0 dB – Italian digits.

The results for office noise are presented in Fig. 3(b) – English digits and Fig. 4(b) – Italian digits. The WRR ranges from 78.63% at 18dB to 71.37% at 0 dB – English digits and from 83.22% at 18dB to 71.46% at 0 dB – Italian digits in EFuNN case and is significantly higher than the WRR of LVQ (21.18% at 0 dB – English digits and 13.83% at 0 dB – Italian digits).



(a) (b)

Fig. 3. English digits - word recognition rate (WRR) of two speech recognition systems: LVQ – codebook vectors – 396, training iterations – 15840. EFuNN – 3MF, rule nodes – 157, sthr=0.9, errthr=0.1, lr1=0.01, lr2=0.01, thrw1=0.2, thrw2=0.2, nexa=100, 1 training iteration. (a) Car noise, (b) office noise.



(a) (b)

Fig. 4. Italian digits - word recognition rate (WRR) of two speech recognition systems: LVQ – codebook vectors – 396, training iterations – 15840. EFuNN – 3MF, rule nodes – 139, sthr=0.9, errthr=0.1, lr1=0.01, lr2=0.01, thrw1=0.2, thrw2=0.2, nexa=600, 1 training iteration. (a) Car noise, (b) office noise.

Going further another experiment is conducted with adaptive noise suppression as a pre-processing technique to show the benefit of this arrangement. The results are summarized in Table 1 and they show that our system combining EFuNN and adaptive noise suppression is able to achieve WRR of 91.74% - English digits and WRR of 91.72% - Italian digits in the case of car noise and 86.47% - English digits and 86.56% - Italian digits in the case of office noise at SNR=0 dB. This is well above the results of the LVQ model, which is the recognition system most often used as a benchmark in designing new recognition systems. This experiment illustrates how an ASN system works and what its advantages are. The ASN is based on both the adaptive noise suppression method [3] and on the EFuNN training method. EFuNN outperforms significantly LVQ model both in WRR and especially in the required time for training, which is 3 to 4 orders of magnitude less. It uses one pass of data propagation, while LVQ (as well as all known connectionist methods for learning from data) require many iterations.

Table 1. The performance of an ASN system based on EFuNN and Adaptive Filtering.

Type of noise	SNR (dB)	WRR (%) English digits	WRR (%) Italian digits
car noise	0	91.74	91.72
	9	91.87	92.62
office noise	0	86.47	86.56
	9	86.63	86.68

6 Conclusions

The ASN method and system proposed here are characterized by the following characteristics:

- effective adaptive noise suppression that is regardless of the origin of the noise; the noise suppression module adapts continuously to the noise added to the speech signal;
- a theoretical potential for creating unlimited vocabulary of words in any language (or languages) and accents;
- high recognition and adaptation accuracy;
- different modes of operation depending on the setting of some parameters: (1) speaker-dependent mode, where the system learns to recognize only a single user; (2) multiple speakers mode, where the system is trained on several speakers; (3) speaker independent mode, where the system can potentially recognize any speaker that speaks with a pronunciation and language that is tolerated by the system;
- fast learning and adaptation what concerns adding new words and speakers if necessary.

The ASN method and system are implementation invariant. They can be implemented as software or/and hardware with the use of either conventional or new techniques. The applicability is broad and spans across all application areas of computer and information science where systems that communicate with humans in a spoken language (‘hands-free and eyes-free environment’) are needed.

References

[1] F. Owens, Signal Processing of Speech. London: Macmillan, 1993.

[2] H. Pelton, Noise Control Management. New York: Van Nostrand Reinhold, 1993.

[3] G. Iliev and N. Kasabov, "Adaptive filtering with averaging in noise cancellation for voice and speech recognition," in Proc. ICONIP/ANZIIS/ANNES’99 Workshop, Dunedin, New Zealand, Nov. 22-23,1999, pp. 71-74.

[4] N. Kasabov, "ECOS - A framework for evolving connectionist systems and the 'eco' training method," in Proc. of ICONIP'98, Kitakyushu, Japan, Oct. 21-23,1998, IOS Press, vol.3, pp. 1232-1235.

[5] N. Kasabov, "The ECOS framework and the 'eco' training method for evolving connectionist systems," Journal of Advanced Computational Intelligence, vol.2, No.6, 1998, pp. 195-202.

[6] N. Kasabov, "Evolving connectionist and fuzzy connectionist systems – theory and applications for adaptive, on-line intelligent systems," in Neuro-Fuzzy Techniques for Intelligent Information Systems, N. Kasabov and R.Kozma, eds., Heidelberg: Physica-Verlag, 1999, pp. 111-146.

[7] N. Kasabov and G. Iliev, "A methodology and a system for adaptive recognition in a noisy environment based on adaptive noise cancellation and evolving fuzzy neural networks," Preliminary Patent, University of Otago, 21 December 1999, New Zealand.

[8] B. Widrow and S. Stearns, Adaptive Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[9] N. Kasabov, "Evolving connectionist systems for on-line, knowledge-based learning with self-optimization," IEEE Trans. Systems, Man, and Cybernetics, submitted, 2000.

[10] E. Trentin and M. Matassoni, "Robust segmental-connectionist learning for recognition of noisy speech," in Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, May 25-26, 1999, pp. 159-162.

[11] T. Kohonen, Self-Organizing Maps. Heidelberg: Springer-Verlag, 1995.