

# A Novel Approach to Very Fast and Noise Robust, Isolated Word Speech Recognition

Ramin Halavati, Saeed Bagheri Shouraki,  
Hossein Tajik, Arpinch Cholakian, Mina Razaghpour

Computer Engineering Department  
Sharif University of Technology  
Tehran, Iran

{halavati, sbagheri, tajik, cholakian, razaghpour} @ce.sharif.edu

## Abstract

*A novel very light weight approach to isolated word speech recognition is introduced. The approach uses a new simplistic feature set and a neural network recognition system. The algorithm's main processing requirements are FFT computation and a simple neural network comparison, making the method a suitable solution for low price embedded devices.*

*The proposed method is tested on single speaker and multiple speaker test sets and the results are compared with a widely used speech recognition approach, presenting very fast recognition and quite good recognition rate.*

## 1. Introduction

Recognition of human speech is a research field with more than 30 years of history and several different solutions are presented to cope with its complexities and difficulties, such as Hidden Markov Models [1], Time Delay Neural Networks [2], Support Vector Classifiers with HMM [3], Independent Component Analysis [4], HMM and Neural-Network Hybrid [5], Hybrid Neuro Fuzzy approaches like [6] and [7], and Fuzzy Hidden Markov Models as in [8] and [9]. But despite the existence of so many approaches, they all have more or less problems such as too much required processing time or low immunity versus noise and these problems have not been quite solved with the increasing of computational resources.

On the other hand, as Zadeh insists in [10], more precise computations do not necessarily result in more correct answer in cognitive tasks and it may even result in poorer answers. Based on this idea and Zadeh's new paradigm for cognitive tasks (i.e. linguistic processing) a novel feature set for speaker dependent phoneme

recognition was introduced in [11] to achieve a high level robustness versus noise. In this approach, the speech spectrogram was converted into a linguistic representation and this representation was recognized using a conventional fuzzy rule base, which was trained by genetic algorithms. The major advantage of the specified approach versus the previous contributions was in its simplicity and its noise robustness for single speaker test sets but on the other hand, it had the disadvantage of very low training speed and inability to learn speaker independent cases.

To develop an isolated word speech recognition system for embedded devices, two main requirements are low computation and high resistance versus noise. To achieve these goals, this paper presents a simplistic feature extraction approach similar to that of [11], with the relative advantages that it does not require fuzzification, it can be used with a simple single layer neural network, and it can be trained and recognized very fast.

In the rest of this paper, section 2 and 3 present the feature extraction and training algorithm. Then the experimental results and comparisons are represented, and at last come the conclusions and future works.

## 2. Feature Extraction

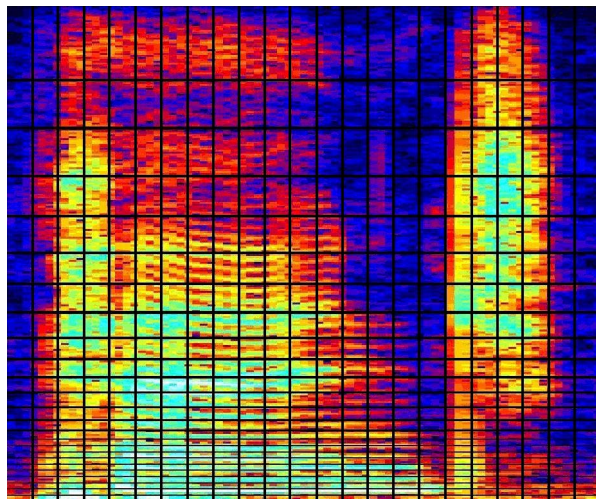
To extract the features, we used three major ideas from human auditory system:

1. We are more sensitive to lower frequencies than higher ones.
2. Our sensitivity changes due to recent inputs.
3. A human recognizer does not read the spectrogram with full precision and only pays attention to local rough features.

Based on these ideas, we first try to have a rough sensitivity adaptation by applying a horizontal mean

filter to the spectrogram. To do so, the average of each frequency band is computed and subtracted from all values of that band. The values that drop below zero are changed to zero and the remaining values are normalized to 0 to 1 range.

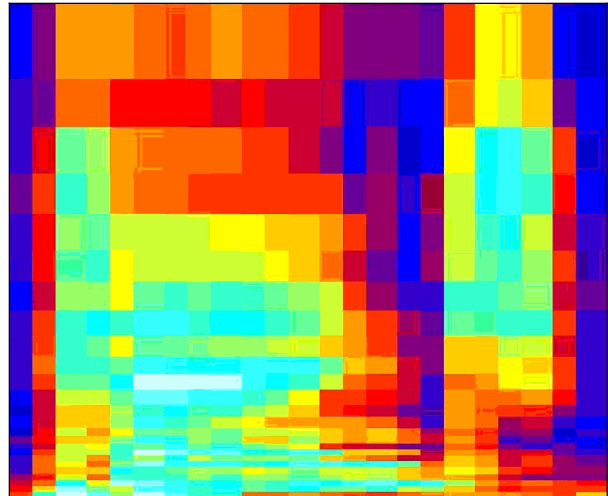
At the next step and to extract local rough measures from the filtered spectrogram, the frequency axis of the spectrogram is segmented into 25 ranges and the time axis is segmented into 23 ranges. To have more sensitivity to lower frequencies, the frequency ranges are selected based on MEL filter banks [12] which are narrower for lower frequencies and wider for higher ones. But the time ranges have equal widths ( $1/23^{\text{rd}}$  of the sample length which is selected based on experiments). Figure 1 presents a sample spectrogram and its segments.



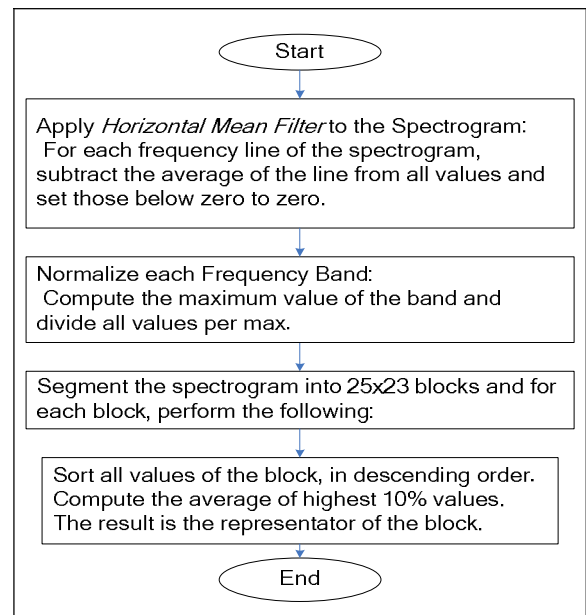
**Figure 1.** A sample of speech spectrogram with the specified segmentations; the vertical axis represents frequency and the horizontal axis represents time.

Once the spectrogram is segmented, we wish to extract one feature from each of the  $25 \times 23$  blocks, so that it represents a rough measure of the highest value of that block. If we just choose the maximum, the value will be very much sensitive to noise. To avoid this sensitivity, the *Average Top 10% (AT10)* operator is used which sorts the values inside one block and chooses the average of the highest 10% values as the representing feature of that block. Figure 2 depicts the spectrogram after application of *AT10* operator.

Once the spectrogram is simplified, each cell represents just one value and we have a  $25 \times 23$  Matrix of values as the feature set. Figure 3 presents the entire feature extraction process.



**Figure 2.** The spectrogram after simplification



**Figure 3.** Diagram of the feature extraction algorithm

### 3. The Training Algorithm

To train the recognition system using *AT10* features, we assume each set of  $23 \times 25$  features as a vector and used one Linear Vector Quantizer neural network for each class. Each network has  $23 \times 25$  weights and the similarity between the input vector and the network weights represents the similarity of the input to that class.

In the training phase,  $N$  random networks (based on the number of classes) are created, and for a fixed number of iterations, the following steps are performed: A random input vector is selected and com-

pared with all networks. If the most similar network is the correct class, no training is required and we move to the next random training sample. But if the correct class gains lower similarity than a wrong class, both networks are adjusted so that the wrong one becomes less similar to the training vector and the correct one gets more similar. The entire training sequence is depicted in Figure 4.

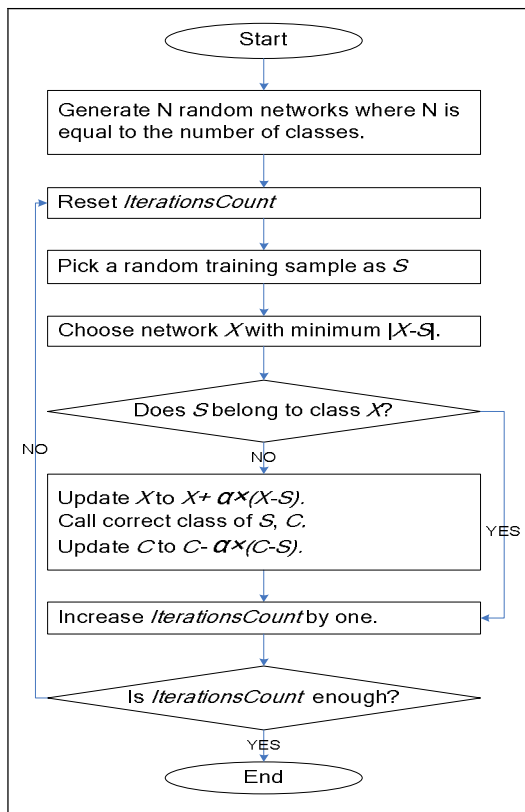


Figure 4. Diagram of the training algorithm

## 4. Experimental Results

To test the proposed algorithm, we used two sets of single speaker and multiple speaker test cases. The single speaker set included 100 words, each recorded 10 times by one female speaker, and the multiple speaker set included 250 words, each pronounced by 10 speakers, 5 male and 5 female, extracted from Fars.Dat speech database.<sup>1</sup>

<sup>1</sup> FARSDAT includes a variety of Farsi speech data uttered by 304 native speakers who differ from each other with regards to age, gender, dialect, and educational level. Each speaker uttered twenty sentences in two sessions. The speech was collected in acoustic booth of the Linguistics Laboratory of the University of Tehran.

In all tests, training is performed by clean data, and the tests are done in existence of 0 to 20 db of additive White or Babble noise. The benchmark system is an HMM recognizer with MFCC feature set.

As depicted in Figures 5 and 6, in all cases, the new approach has shown almost similar results in compare with MFCC+HMM benchmark in clean or low noise environments and much better results in cases where the noise level, whether white or babble, increases.

Also, Figure 7 presents the comparison of required processing time for this approach versus the benchmark system. The tests have been done on 10 minutes of recorded speech, segmented into 1000 separate sample files. As it is seen, the AT10 approach requires half the processing time the MFCC features require and the used Neural Networks consume about 240 times less processing time than HMM model.

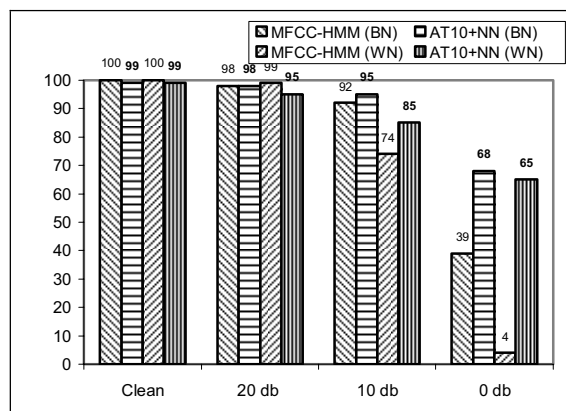


Figure 5. Single speaker recognition results

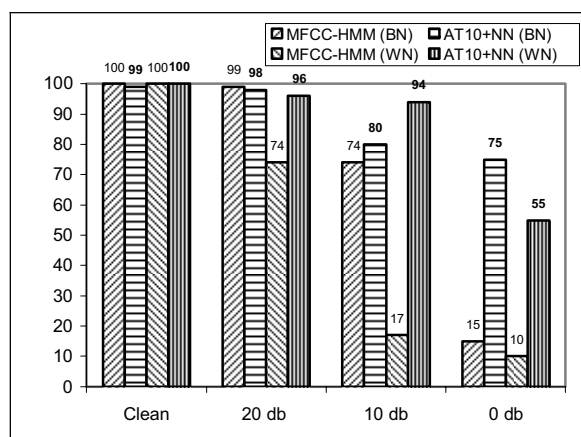
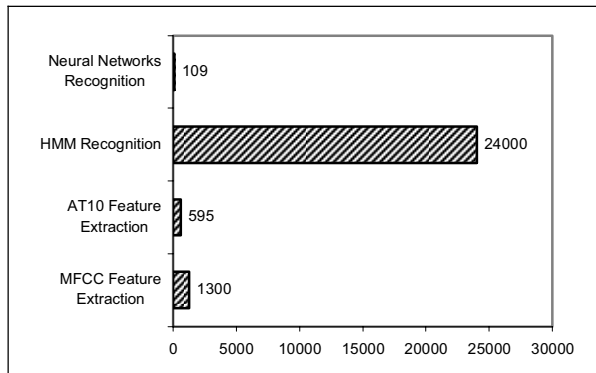


Figure 6. Multiple speaker recognition results



**Figure 7.** Processing time for classifying 10 minutes of recorded speech in milliseconds

## 5. Conclusions and Future Works

A major problem of most ASR systems is noise sensitivity and high computational requirements. To develop a fast, yet noise robust isolated word speech recognition system, this paper presented a novel feature set from speech spectrogram, which is based on a simple operator, called *AT10*.

The result of speech recognition using this feature set and 1 layer neural networks, show that in compare with HMM-MFCC benchmark, this method is more than 200 times faster. Also, it can resist 0 db additive white or babble noise and have 55% to 75% recognition rate in such environment.

Noting that most of the required processing of the proposed feature extraction algorithm is in computation of FFT which has low price, optimized hardware chips and the recognition is done with simple neural networks, which also has hardware implementations, the good recognition results and noise robustness makes this approach a very desirable solution for low price light weight embedded ASRs.

As a next step to this contribution, we are working on algorithms for fast adaptation with speakers, dealing with more elasticity in words lengths, and simultaneous voice activity detection and recognition.

## 6. References

- [1] Babaali, B., and Sameti, H. (2004), "The Sharif Speaker-Independent Large Vocabulary Speech Recognition System", The 2nd Workshop on Information Technology & Its Disciplines (WITID 2004), Feb. 24-26, 2004, Kish Island, Iran.
- [2] Berthold, M.R. (1994), "A Time Delay Radial Basis Function Network for Phoneme Recognition", Proceedings of the IEEE International Conference on Neural Networks, vol. 7, pp.4470-4473, Orlando, 1994.
- [3] Golowich, S.E., and Sun, D.X. (1998), "A Support Vector/Hidden Markov Model Approach to Phoneme Recognition", ASA Proceedings of the Statistical Computing Section, pp. 125-130.
- [4] Kwona, O.W., and Lee, T.W. (2004), "Phoneme recognition using ICA-based feature extraction and transformation", Signal Processing, Vol. 84, No. 6, pp. 1005-1019, June 2004.
- [5] Schwarz, P., Cernocky, M., and Cernocky, J. (2004), "Phoneme recognition based on TRAPs", Workshop on Multimodal Interaction and Related Machine Learning Algorithms, June 2004.
- [6] Leung, K.F., Leung, F.H.F., Lam, H.K., Tam, P.K.S. (2003), "Recognition of speech commands using a modified neural fuzzy network and an improved GA", IEEE International Conference on Fuzzy Systems, v 1, 2003, p 190-195.
- [7] Doye, D.D., Kulkarni, U.V., Sontakke, T.R. (2002), "Speech recognition using modified fuzzy hypersphere neural network", Proceedings of the International Joint Conference on Neural Networks, v 1, 2002, p 65-68.
- [8] Cheok, A.D., Chevalier, S., Kaynak, M., Sengupta, K., Chung, K. (2002), "Use of a novel generalized fuzzy hidden Markov model for speech recognition", IEEE International Conference on Fuzzy Systems, v 3, 2002, p 1207-1210.
- [9] Tran, D., Wagner, M. (1999), "Fuzzy hidden Markov models for speech and speaker recognition", Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS, 1999, p 426-430.
- [10] Zadeh, L.A. (2002), "From Computing with Numbers, to Computing With Words, A New Paradigm", International Journal on Applied Mathematics, 2002, Vol.12, No.3, 307-324.
- [11] Halavati, R., Shouraki, S.B., Sameti, H., Zadeh, S.H., Babali, B. (2005), "A Novel Noise Immune Approach to Speech Recognition", Proceedings of IFSA World Congress 2005, p 972-976.
- [12] Stevens, S.S., and Volksman, J., "The relation of pitch to frequency", American Journal of Psychology, vol. 53, p.329, 1940.