A Novel Noise Immune Fuzzy Approach to Speaker Independent, Isolated Word Speech Recognition

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

Computer Engineering Department, Sharif University of Technology,
Tehran, Iran

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

Abstract. This paper presents a novel approach to speaker independent isolated word speech recognition using fuzzy modeling which is specifically designed to ignore noise. The task is based on conversion of speech spectrogram into a linguistic fuzzy description and comparison of this representation with similar linguistic descriptions of words. The method is tested and compared with a widely used speech recognition approach and has shown a significant higher robustness versus noise.

1 Introduction

Recognition of human speech is a problem with many solutions, but still open because none of the current methods are fast and precise enough to be comparable with a human recognizer. Several methods exist for recognition of human words 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], and more. But they have more or less problems such as too much required processing time or low immunity versus noise and these weaknesses hold the road open for new approaches.

On the other hand, Zadeh proposes computations with words instead of precise numbers as a new paradigm for cognitive problems [6]. He insists that more precise computations do not necessarily result in more correct answer in cognitive tasks and it may even result in poorer answers. Also, we know that humans do not try to find and remove noise in their cognitive skills [7] and usually just try to ignore it. Based on these ideas, several speech recognizer systems use fuzzy computation to achieve a higher noise immunity level. For example, [8] and [9] have used a hybrid Neuro-Fuzzy approach to recognize speech commands or [10] and [11] have exploited a fuzzy Hidden Markov Model to recognize speech tokens. But in all such cases, fuzzy is used as a high level decision making approach and runs over the previously computed precise data.

In [12], another fuzzy method for isolated phoneme recognition is proposed in which the input signal is represented by a fuzzy description language and compared with similar descriptions for phonemes. A major difference between this approach and the other cited approaches is the fact that this model doesn't use conventional acoustic parameters and speech features and instead, uses fuzzy features which are specifically designed to overcome noise and it is shown that it has much better resistance versus some other widely used approaches.

To extend this approach to word level, one must confront the following problems:

- First, there is no accurate algorithm for separation of words into phonemes and therefore it is not possible to simply define words as sequences of phonemes.
- Second, the exact definition of phoneme is '*The smallest phonetic unit in a language that is capable of conveying a distinction in meaning such as m in mat and b in bat for English.*'[13] But this distinction is not sufficient for speech recognition as many phonemes have different acoustic features based on their preceding and succeeding letters.
- Third, to have a large-vocabulary word-recognizer system, the training and recognition time are two important concerns while this is not a very important fact for phoneme recognizer systems as the set of phonemes are quite limited and few (around 70 in English).

To overcome the above problems and to have a completely-fuzzy isolated word speech recognition system, this paper presents a word recognition system which is based of fuzzy-gesture recognition and fuzzy rule base. The major rationales behind this approach are:

- When we aim to recognize gestures (the visual representations) instead of phonemes (linguistic definitions), we eliminate the problem of phonemes with several different acoustic features.
- Using gestures with equal lengths removes the requirement of segmenting words into phonemes.
- By recognizing gestures to classify words, we have a wider range of sub-words in compare with word recognition using phonemes, resulting in more possible discrimination.

The rest of this paper is organized as follows: Sections 2 and 3 respectively represent input conversion and rule base generation approaches, Section 4 presents the experimental results and at last comes the conclusions and future works.

2 Fuzzy Representation of Input Signal and Words

The first step of the process is the conversion of speech spectrogram into a fuzzy description. The fuzzification approach is based on three major ideas:

- A human recognizer does not read the spectrogram with full precision and pays attention only at local features.

- Human do not decide based on precise speech amplitudes and only a rough measure is sufficient.
- We are more sensitive to lower frequencies than higher ones.

Based on these ideas, the frequency axis of speech spectrogram is segmented according to MEL filter banks [14]. The MEL filter bank frequencies are designed so that the ranges are narrower in lower frequencies where human ears are more sensitive and wider in higher frequencies where we are less sensitive. And the time axis is segmented into 23 equal ranges. This is number is chosen heuristically and based on some trial and errors on recognition performance. Figure 1-left shows the spectrogram and its segmentations.

Then, to get the general view of each block, we need to represent all amplitudes in one block with one value which gives the general feeling of that block. This can be done by sorted all values of each block in descending order and taking the average of top 10% as the representator of all values. The rationale behind this approach is the fact that we seek to find the almost highest amplitude in each block, and if we just choose the highest point, we become too much sensitive to noise. But averaging the top 10% gives a higher level of resistance versus noise. Figure 1-right presents a sample of the simplified image.

Once the figure is simplified, it can be represented by a set of linguistic terms using 5 fuzzy sets representing VERY LOW, LOW, AVERAGE, HIGH, and VERY HIGH amplitudes. Therefore, a word can be represented by a matrix of 25×23 fuzzy sets whose T-Norm (we used product) gives the degree of similarity between the input and the word. Each fuzzy set is assumed as a triangle that is represented with 3 points, as shown in Figure 2.

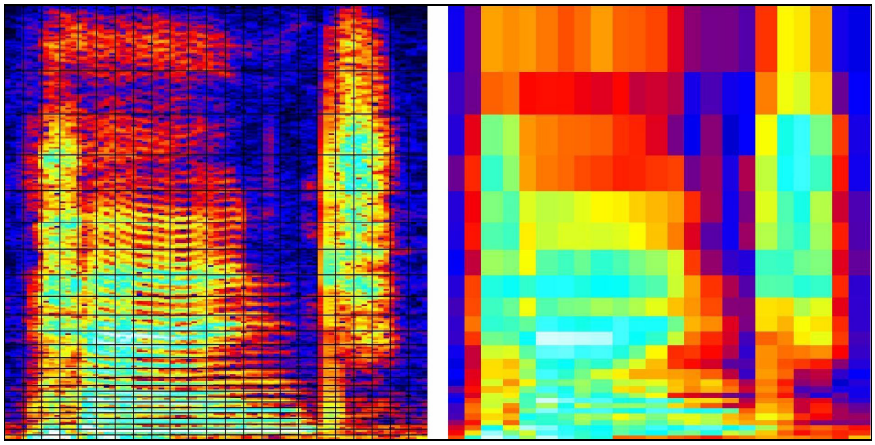

Fig. 1. Left: A sample of speech spectrogram and its horizontal and vertical segmentations. The vertical axis represents frequencies and the horizontal axis represents time. The lighter values present higher amplitude. **Right:** The simplified form of the left image

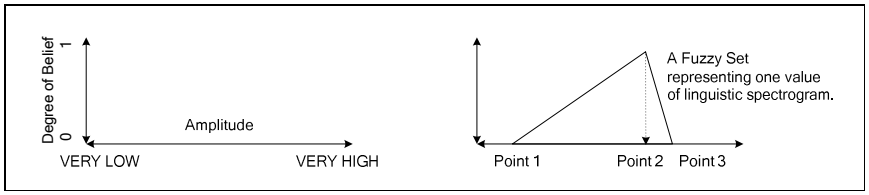

Fig. 2. A sample fuzzy set, representing one value of the linguistic spectrogram.

3 Generation of Fuzzy Rule Base

To generate the fuzzy rule base to recognize words, we assume a rule base with one rule for each word and each rule consisting of 23×25 fuzzy sets, called a word description. The algorithm initializes by creating random descriptions. In each iteration of the training period, one sample word is chosen randomly and compared with all word descriptions. If the description which shows the maximum degree of similarity is not the correct one, the correct description is altered so that it gains a little more similarity to the given sample.

These modifications can be done by moving each fuzzy set of the description so that it would match more with the input to be learned. This is done by moving one or more points from the set definition points (Figure 2) and in cases where more than one choice is available, one is chosen randomly.

When all samples are compared with descriptions and the mistaken ones are updated, if the total error rate is above a certain threshold, the entire process is repeated and otherwise the training process is assumed finished. Figure 3 presents a diagram of the training sequence.

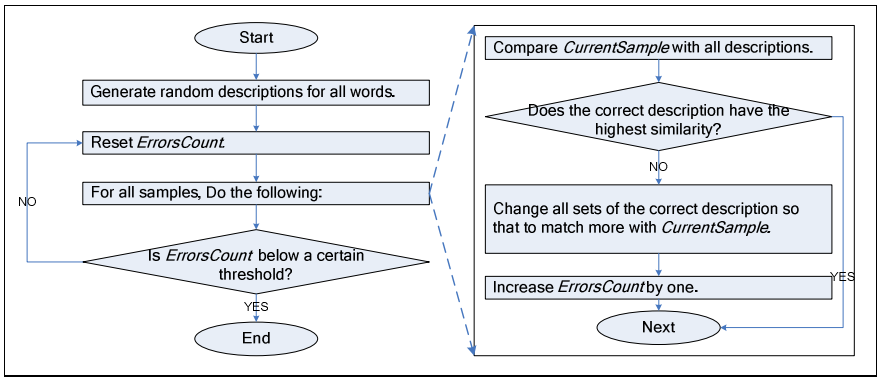

Fig. 3. Diagram of the training algorithm

4 Experimental Results

To compare the above system with conventional speech recognition approaches, we used a Hidden Markov Model system with MFCC acoustic features as the benchmark. Both systems were tested with two sets of speech samples: the first set composed of 100 words, each recorded 10 times by one speaker and the second set composed of 250 words, each recorded 10 times with 10 different speakers. The tests where run with clean samples and 20, 10 and 0 db white and babble noise.

As depicted in Figure 3, the proposed algorithm has a recognition rate of 5 to 10 percent less than MFCC-HMM benchmark, but shows better results when the noise level exceeds 20db and reaches 60 to 70 percent better results in 0 db noise. This much higher results stays similar both for white noise where the noise is distributed uniformly among all frequency bands and babble noise which has a more concentration in lower frequencies.

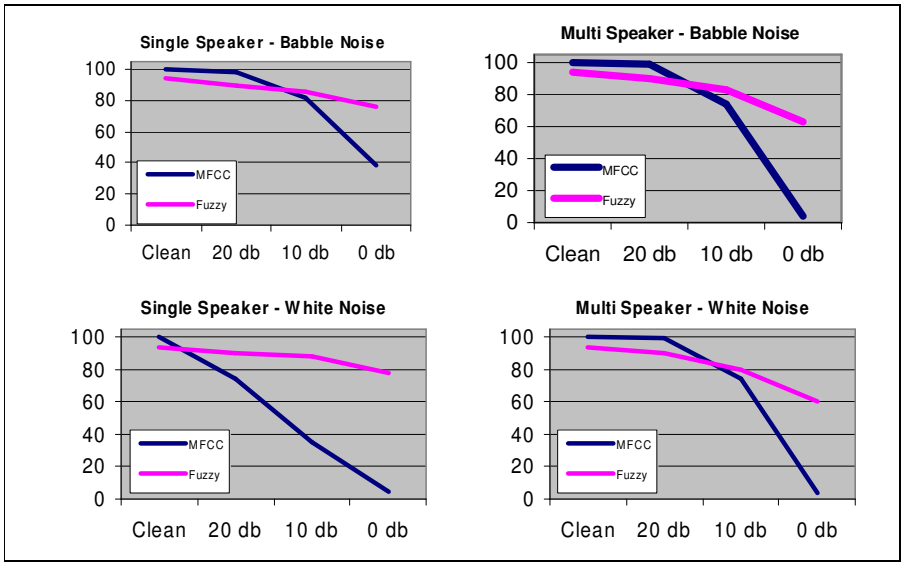

Fig. 4. Comparison results of this approach and MFCC+HMM benchmark system

5 Conclusions and Future Works.

One of the major problems with most of the current speech recognition systems is noise sensibility which drastically drops the recognition rates in presence of high amounts of noise. To overcome this problem, this paper presents a novel modeling of speech spectrogram using fuzzy descriptive terms, based on some insights from human auditory system to recognize isolated words.

To do so, the speech spectrogram is first simplified and fuzzified to be represented with a set of linguistic terms and a fuzzy rule base is generated to classify these representations into word classes.

The final recognition system is compared with a HMM system which uses MFCC features and is tested in multi speaker and single speaker tasks with different levels of additive babble or white noise. As presented in previous section, the results presents a slight (at most 10%) lower result in clean test cases and much better results (up to 75% better) in test cases with high amounts of noise. Thus it can be concluded that the system has been quite successful in dealing with noise and ignoring it.

As a next step to this contribution, we are working on an appropriate elastic fuzzy rule base to deal with different stretching of words and also on an adaptation algorithm which will improve the results when a single speaker is focused.

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. 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.
7. Dreyfus, H.L., "What Computers Still Can't Do, A Critique of Artificial Reason", MIT Press, 1972, page 121.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. The American Heritage® Dictionary of the English Language, Fourth Edition. Copyright © 2000 by Houghton Mifflin Company. Published by Houghton Mifflin Company. All rights reserved.
14. Stevens, S.S., and Volksman, J., "The relation of pitch to frequency", American Journal of Psychology, vol. 53, p.329, 1940.