

A Novel Fuzzy Approach to Speech Recognition

Ramin Halavati, Saeed Bagheri Shouraki,
Mahsa Eshraghi, Milad Alemzadeh

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

Abstract. This paper presents a novel approach to speech recognition using fuzzy modeling. The task begins with conversion of speech spectrogram into a linguistic description based on arbitrary colors and lengths. While phonemes are also described using these fuzzy measures, and recognition is done by normal fuzzy reasoning, a genetic algorithm optimizes phoneme definitions so that to classify samples into correct phonemes. The method is tested over a standard speech data base and the results are presented.

Keywords: Fuzzy Modeling, Speech Recognition, Genetic Algorithms

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 phonemes such as Hidden Markov Models (Babaali and Sameti, 2004), Time Delay Neural Networks (Berhold, 1994), Support Vector Classifiers with HMM (Goldwich and Sun, 1998), Independent Component Analysis (Kwina and Lee, 2004), HMM and Neural-Network Hybrid (Schwarz et al, 2004), and more. Although all methods have scored enough to be accepted as a suitable approach, but they have more or less problems such as too much processing or low immunity versus noise and these weaknesses hold the road open for new approaches. Zade proposes computations with words instead of precise numbers as a new paradigm for cognitive problems (Zade, 2002). He insists that more precise computations do not necessarily result in more precise result in cognitive tasks and it may even result in poorer answers.

To reduce the computation effort while keeping the recognition precision, this paper presents a new

approach to phoneme identification using fuzzy computations. To do so, the spectrogram of training speech samples is converted into a fuzzy representation using some arbitrary colors and lengths and all computation take place using these linguistic descriptions. A standard genetic algorithm trains the fuzzy definitions to find the best description for each phoneme and recognition is done by a very fast fuzzy reasoning.

The rest of this paper is organized as follows: In the next section, fuzzy representation of speech data is described. Section three deals with the recognition method, and training approach is described in section four. Section five presents experimental results and the last section makes the conclusions and presents the further works.

2 Linguistic Spectrogram Representation

The first step of both recognition and training processes is conversion of the spectrogram of speech signals¹ into a fuzzy description. The fuzzification approach is based on four major ideas: First, a human recognizer does not read the spectrogram with full precision and pays attention only at local features. Second, a human does not decide based on precise speech amplitudes and only a rough measure is sufficient. Third, we do not count speech frames and use relative lengths such as long or short. Four, we are more sensitive to lower frequencies than higher ones.

Based on these ideas, the frequency axis is separated into 25 ranges according to MEL filter banks. The MEL filter bank frequencies are selected so that the ranges are narrower in lower

¹ Spectrogram is a 2-D Image in which, the vertical axis represents the frequencies and the horizontal axis represents time. The brightness of each point shows the amplitude of a certain frequency at a certain time. A sample is presented in Figure 1.

frequencies and wider in higher ones. Figure 2 shows the spectrogram separated with horizontal lines based on MEL bands and vertical lines based on phoneme positions. Then, to make the local data reduction, each sample column in each MEL band is considered as one block and represented with one

value which is the average of 10% of highest amplitude points in that block. Figure 3 shows the result of this stage. Note that new image has only 25 values in each vertical line (frequency axis) but the time axis is not altered.

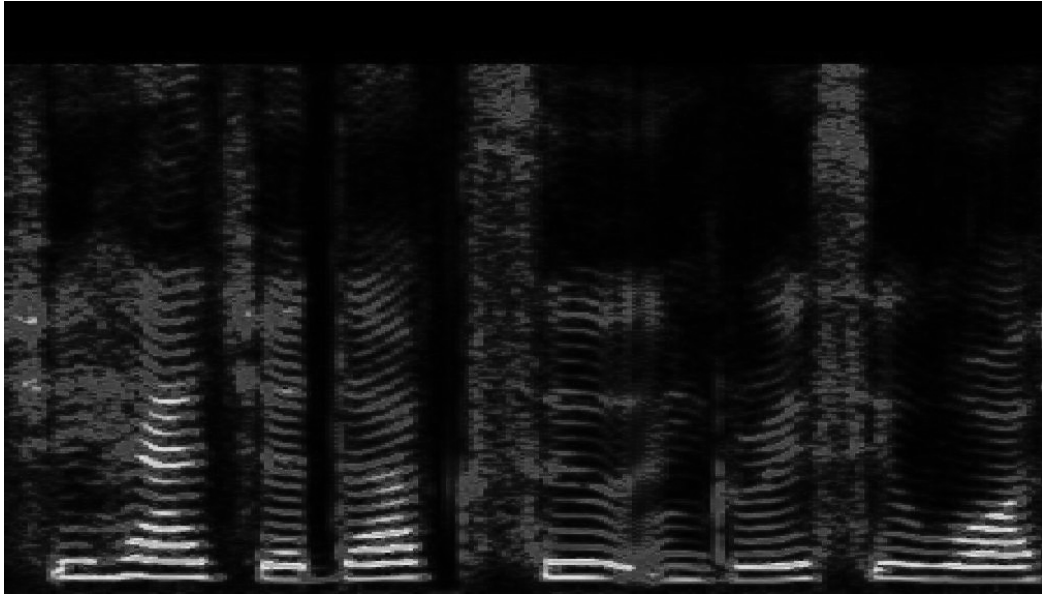


Figure 1 – Spectrogram of voice signals, the vertical axis represents the frequency and the horizontal axis represents time. The color of each point shows the amplitude of a specific frequency at a specific time

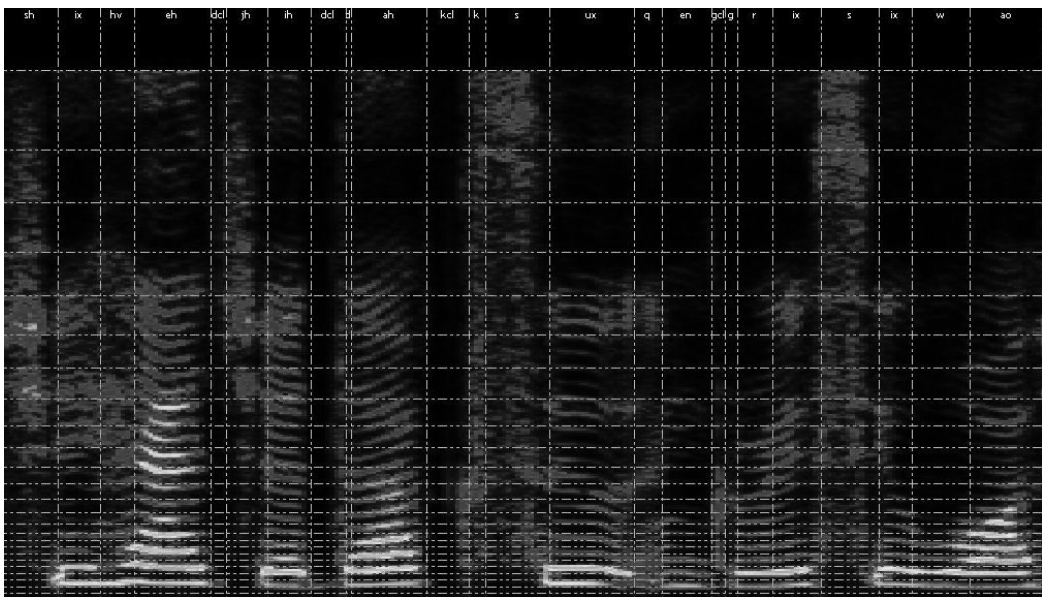


Figure 2, Spectrogram separated with MEL filter bank horizontal lines and phoneme separator vertical lines.

As the next step in fuzzification, a gradient of color is used to represent different ranges of amplitudes which starts from black for lowest amplitude, then blue, magenta, red, yellow, green, cyan and finally fading to white showing the highest value. To describe these eight colors, eight fuzzy sets are used which are each described with a trapezoidal as in Figure 4. Also, as shown in Figures 2 and 3, different phonemes have different lengths. To express this differences, five fuzzy lengths (Very Short, Short, Average, Long and Very Long) are defined. Trapezoidal shapes are used for their definitions, as in colors.

Now, using the above conversions and definitions, each phoneme can be described using an expression stating its length, and its probable colors for each band. For example, one can express a phoneme as stated in Table 1. Note that a disjunction of different colors can be used in expressing each band's color. Based on this type of

phoneme expression, a phoneme recognizer system must have the definitions for fuzzy colors (definition of trapezoids for each of 8 colors), definition of fuzzy lengths, and description of all phonemes using the previously stated colors and lengths which includes a length value for each phoneme and color values for each frequency band of every phoneme.

Table 1, Sample phoneme definition

Phoneme:	/SH/
Range 1:	Black or Blue
Range 2:	Blue
Range 3:	Red or Yellow
...	
Range 25:	Purple or Blue
Length:	Average

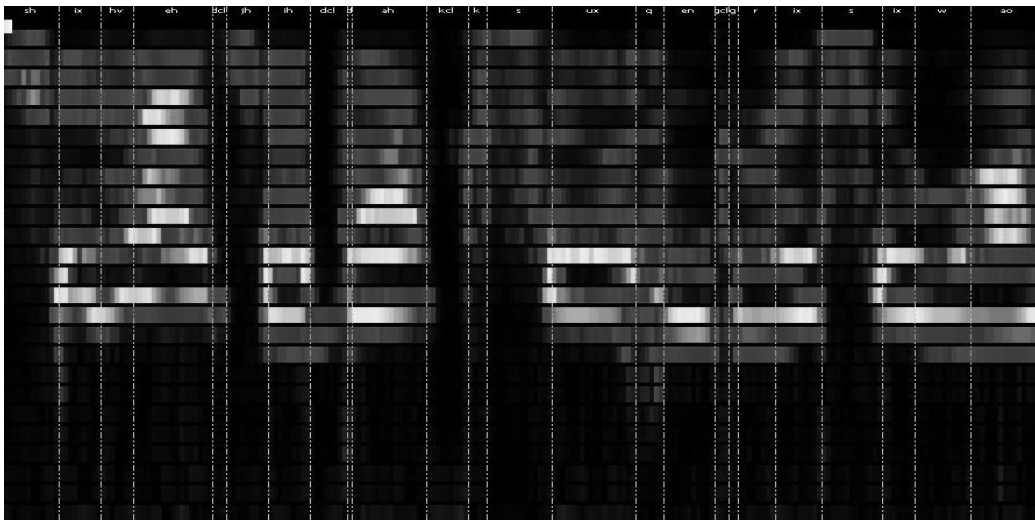


Figure 3 – Fuzzified Spectrogram of Figure 2

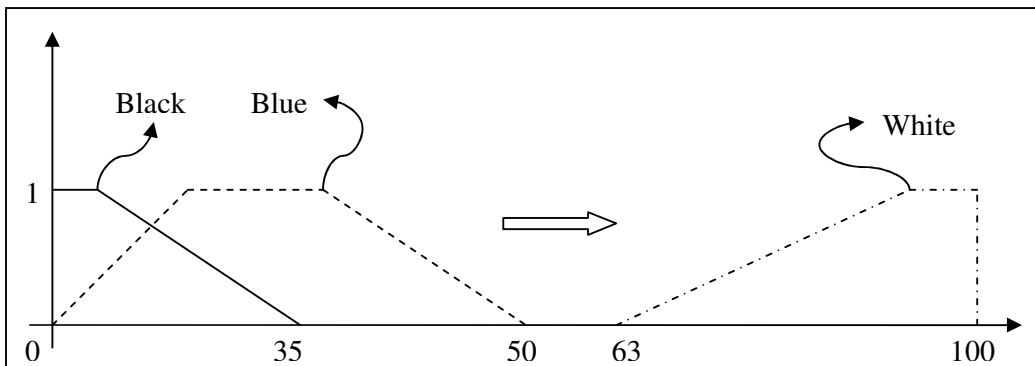


Figure 4 – Color fuzzy sets: The horizontal axis is the amplitudes ranging from 0 to 100, and the vertical axis is the degree of belongingness to each set.

3 The Recognition Process

Assuming that a suitable set of fuzzy colors, fuzzy lengths and phoneme descriptors exist, this part presents the recognition approach which classifies given phonemes. Each input sample may have an arbitrary number of frames, but each frame has exactly 25 values, computed as stated in the previous part.

To recognize, a degree of belongness to each of the phonemes is computed for the given input and the input is classified into the phoneme which has the highest belongness value. This task is done in three steps:

3.1 Step One:

The first step is to compute how much a frame of the given input belongs to the specified pattern of colors. For example, if the input data has 13 frames, the belongness of each frame to current phonemes' expressing colors is computed independent of other frames. To compute this belongness, the value of each frequency band is compared to the described colors of that band and their belongness values are computed. Then, minimum of all these values is considered as the total belongness to that color pattern. Figure 5 shows a sample of this approach.

3.2 Step Two:

After computation of belongness measures for each frame, a small refinement filter is done over the resulting values for neighbor frames. This filter increases low values which are sieged by high values and decreases high values whose neighbors are all low. This task is done to decrease the effect of noise and dissimilarities in identified patterns.

3.3 Step Three:

The input of the last step is the refined belongness measure of each frame to the specified pattern of colors. The longest sequence of frames whose belongnesses are above a certain threshold are found and is called the matching length and their average is called matching value. Then, matching length is compared with the fuzzy set specifying the phoneme's length and its belongness value is computed and multiplied by the so called matching value. This final result specifies the degree of belongness of the given sample to that phoneme pattern. Figure 6 presents a sample for this step.

4 Training Algorithm

The training approach takes place as a normal genetic algorithm (Holland, 1975). A complete recognizer including color definitions, length definitions and all phonemes descriptions is called a genome. The fitness of a genome is defined as how good it can classify phoneme samples. The training process begins with a collection of random genomes. In each iteration, all genomes are sorted based on their fitnesses and 50% of the low ranking ones are thrown away. Then the population is recovered by creating new genomes from the remaining ones using normal genetic algorithm cross-over and mutation operators. This circle is repeated as long as the top scoring genome passes a required limit.

5 Experimental Results

The above algorithm is applied on Timit database with 62 phoneme classes and is tested for single speaker sample sets. In each test, some of one speaker's sentences are used for training and one sentence is used for testing and the results for non-trained or none tested phonemes are removed². The results are presented in Table 2.

Table 2, Experimental Results

Average Training Time:	3-4 hours
1 st correct answers:	85%
3 rd correct answers (out of 62) ³ :	95%
6 th correct answers (out of 62):	98%

² Some sample sets or test sets did not include all phonemes.

³ One of the top three guesses has been correct.

	Input Value	Definition	Max Value
Band 1	34	 Black or Blue	 100 %
Band 2	12	 Blue	 85 %
Band 3	56	 Red or Yellow	 100 %
...	⋮	⋮	⋮
Band 25	89	 Red or White	 90 %
Min Value			Final Value this frame: 85 %

Figure 5. Sample Belongness computation for one frame (Recognition Step 1).

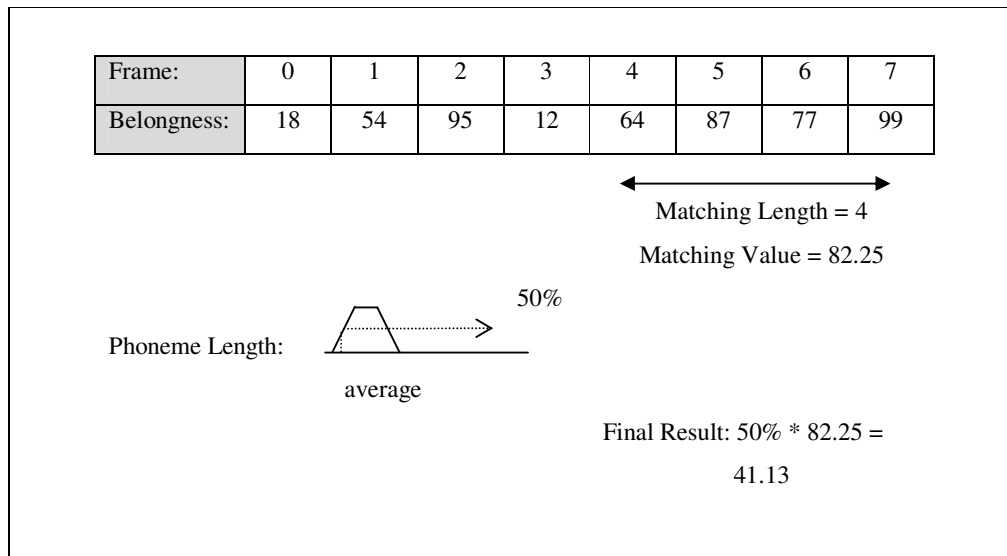


Figure 6. Sample Computation of Step 3 of Recognition

6 Conclusion and Future Work

Despite the existence of several methods for speech recognition, the problem is still open as no algorithm is both fast and accurate enough to be an ultimate answer. Based on fuzzy computations, a novel approach is presented which is believed to be fast and accurate but still needs work to be a complete solution. In this method, the speech signal data are converted into a representation with words and using same representation for phonemes, their correct phoneme classes are identified. The phoneme descriptors are trained using normal genetic algorithms and while having 3 to 4 hours of training time for a set of around 400 phoneme samples, the classification task is very fast due to very little needed computations. The algorithm is tested on Timit database for single-speaker samples and it has resulted in around 85% first correct answers and 95% third correct ones.

As a next step to improve the algorithm, the training algorithm can be altered so that to be faster while having more samples, sequence data can be used so that to consider the relative changes of frames, and word descriptors instead of phoneme descriptors can be designed.

7 References

- 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.
- 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.
- 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.
- 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.
- Schwarz, P., Cernocky, M., and Cernocky, J. (2004), "Phoneme recognition based on TRAPs", *Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, June 2004.
- Zade, 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
- Holland, J. (1975), "Adaptation in Natural and Artificial Systems". Ann Arbor, Michigan: University of Michigan Press