Emotion Recognition Using a Data-Driven Fuzzy Inference System

Chul Min Lee and *Shrikanth Narayanan*

Speech Analysis and Interpretation Laboratory and Integrated Media Systems Center
Department of Electrical Engineering, University of Southern California
{cml,shri}@sipi.usc.edu, http://sail.usc.edu

Abstract

The need and importance of automatically recognizing emotions from human speech has grown with the increasing role of human-computer interaction applications. This paper explores the detection of domain-specific emotions using a fuzzy inference system to detect two emotion categories, *negative* and *non-negative* emotions. The input features are a combination of segmental and suprasegmental acoustic information; feature sets are selected from a 21-dimensional feature set and applied to the fuzzy classifier. Our fuzzy inference system is designed through a data-driven approach. The design of the fuzzy inference system has two phases: one for initialization for which fuzzy c-means method is used, and the other is fine-tuning of parameters of the fuzzy model. For fine-tuning, a well known neuro-fuzzy method are used. Results from on spoken dialog data from a call center application show that the optimized FIS with two rules (FIS-2) improves emotion classification by 63.0% for male data and 73.7% for female over previous results obtained using linear discriminant classifier.

1. Introduction

The need to develop emotion recognition systems from human speech signals has increased as human-computer interaction is increasingly playing a significant role in everyday environment [1][2]. In this regard, prior research on emotion recognition from speech has been done to develop systems using acoustic features such as pitch, energy, and durations of the speech [2][3]. Although such efforts have focused on a few basic emotion classes such as anger, sadness, happiness, and other "universal" emotions, the class boundaries between emotion categories are vague or fuzzy because of the linguistic uncertainties in the definition of the emotional classes in everyday use [4]. In this study, we used two broad emotion categories, *negative* and *non-negative*, and those emotion classes in turn would cover various emotions; e.g., negative emotion class might include "anger", "frustration", and "boredom", and non-negative class would have "neutral", "happiness" and other *positive* emotions. In prior work, we relied on statistical description of class specific data. However, the linguistic vagueness in the definition of emotion categories implies that they can overlap each other in their acoustic information spaces.

To tackle this problem, we applied fuzzy inference (FIS) to emotion recognition. FIS based on fuzzy rules has been applied to numerous engineering applications such as control, signal processing, and pattern classification problems [5][6]. FIS is basically a rule-based system to tackle the language-related uncertainties in the given problems. In pattern classification, fuzzy logic improves classification and decision systems by allowing the use of overlapping class definitions and improves the interpretability of the results by providing more insight into the classifier structure and decision making systems [7]. The Takagi-Sugeno-Kang (TSK) FIS is used in this paper because the TSK model is suitable for generating fuzzy rules from a given input-output data set in a data-driven fashion[8].

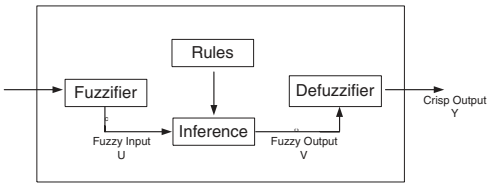


Figure 1: A fuzzy inference system

In modelling FIS, determining the structure, i.e., the number of rules in the system, requires the prior knowledge of the problem. This is usually determined by experts in conventional FIS designs. This work focuses on a data-driven approach since no expert is available to determine the number of rules in the system. For that purpose, unsupervised fuzzy clustering in input space is used to determine the number of rules in the classification problem. Since *Fuzzy C means* (FCM) is a fuzzy version of *K means* algorithm, and a widely adopted method in the literature [9], and significantly reduces the number of rules in a model compared with other partitioning methods [5], it was used to initially identify the structure of FIS in this work. After the initial model is determined, the *antecedent* and consequent parameters in the rules are fine-tuned by a neuro-fuzzy method, *Adaptive-Network-Based Fuzzy Inference System* (ANFIS) [10].

Although a system that can recognize a large variety of emotions is attractive, it may not be necessary or practical in the context of human-machine conversational interfaces. In this regard, we focus on recognizing *negative* and *non-negative* emotions from speech signals, which are defined in our previous work [2], using data collected from a commercially deployed automatic call center dialog system. This study also focuses on acoustic information in speech, especially, prosody-based information.

The rest of paper is organized as follows: Section 2 describes the basic theory of fuzzy inference system and explains methods in the design of FIS. Section 3 reports on the experimental results of classification of two emotion classes by FIS, and comparison between the results from FIS and other classification methods such as *linear discriminant classifier* (LDC) and *nearest neighborhood classifier* (NN). Finally, Section 4 concludes the paper.

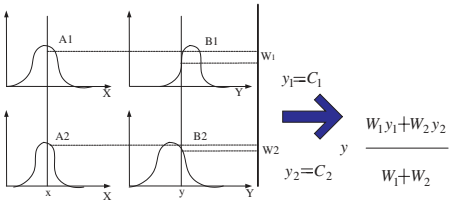


Figure 2: A two input zeroth-order TSK fuzzy model with two rules.

2. Fuzzy Inference System

Fuzzy inference systems, which are also called fuzzy rule-based systems, or fuzzy models, are composed of 4 blocks (see Figure 1).

- *Fuzzifier*: Transforms the crisp inputs into fuzzy inputs by membership functions (MFs) that represent fuzzy sets of input vectors. In this work, singleton fuzzifier is assumed; i.e., $\mu_A(x) = 1$ for $x = x'$ and $\mu_A(x) = 0$ for all $x \in U$ with $x \neq x'$;
- *Rules*: Consists of fuzzy IF-THEN rules;
- *Inference*: Inference engine for fuzzy rules;
- *Defuzzifier*: Transforms the fuzzy output into crisp output. Defuzzification process requires the most computational complexity in FIS, and center-of-gravity or height defuzzification method is common. In TSK FIS, the final output is a weighted average of each rule output; therefore, it does not require defuzzification process.

The major component in an FIS is "Rules", and Rules are expressed in the form of IF-THEN statements. Let U and V be universe of discourses for *antecedent* and *consequent* of the rules, then the rule of *if x is A, then y is B*, where $x \in U$, and $y \in V$, represents a relation between A and B, and extension to multiple rules and multiple antecedents can be easily done by specifying both *composition* and *inference* methods [11]. Throughout the paper, we adopt *product* composition for "and" operation, and *min* inference, which is most commonly used composition and inference methods in engineering applications [11].

Denote $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ as an input feature vector in the classification problem, then typical TSK FIS consists of IF-THEN rules where consequent parts are constant (zeroth-order) or linear function (first-order) of inputs, and has the form of:

$$R^l : \text{If } x_1 \text{ is } F_{1l} \text{ and } \ldots, x_p \text{ is } F_{pl}, \text{ then } y_l = b_{i1}x_1 + \ldots + b_{ip}x_p + b_{i(p+1)} \quad l = 1, \ldots, M \quad (1)$$

where M is the number of rules, y_l is the output of the lth rule, and $F_{1l}, \ldots, F_{pl}$ are the antecedent fuzzy sets. The overall output of the model is computed by

$$y = \frac{\sum_{l=1}^{M} w_l y_l}{\sum_{l=1}^{M} w_l} = \frac{\sum_{l=1}^{M} w_l C_l}{\sum_{l=1}^{M} w_l} \quad (2)$$

where w_l is the degree of activation of the antecedent in the lth rule

$$w_l = \prod_{i=1}^{p} F_{il}(x_i), \;\; l = 1, 2, \ldots, M \quad (3)$$

An example of a zeroth-order TSK model is shown in Figure 2 with two input features and two rules.

Number of Clusters	Partition Coefficient
2	**0.5446**
3	0.3639
4	0.2723
5	0.2499

Table 1: An example of Partition coefficients (PC) for given number of clusters with a 21-dim base feature set (male data). The whole dataset is used to compute the PC. Both PC and the cluster number where PC is maximized are highlighted.

2.1. Initial Rule Base Generation

Fuzzy clustering is used to generate the antecedent of the initial fuzzy rule base, and *fuzzy c-means* (FCM) algorithm is used for fuzzy clustering [12]. Suppose that input data, $X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\} \subset R^p$ consists of n vectors in p dimensional space, then FCM defines a soft clustering into $c < n$ clusters, minimizing the following objective function:

$$J = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \quad (4)$$

$$d_{ij}^2 = \|\mathbf{x}_j - \mathbf{c}_i\|^2 \quad (5)$$

$$(6)$$

where m is weighting exponent, and set to 2 in this work. By minimizing J, FCM is characterized by cluster centers:

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} u_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}^m}, \;\; i = 1, \ldots, p \quad (7)$$

and a MF denoting degree of *belongingness* of the data $\mathbf{x}_j$ to the ith cluster u_{ij}:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{kj}}{d_{ij}})^{2/(m-1)}} \quad (8)$$

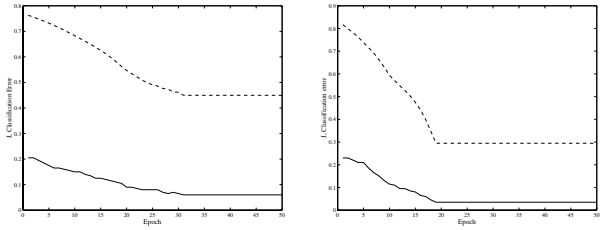
Two difficulties arise in the application of FCM; one is how to set the initial cluster center, and the other is the optimal number of clusters. For the former, we randomly select the data points and set them as initial centers. It is known that the objective function J increases as c increases. To resolve the latter, a validity measure of the number of clusters is used [9][12]. Usually, it can be useful to make the partition as crisp as possible, and Bezdek suggested *partition coefficient* (PC) as a suitable measure in this case [12]:

$$PC(u) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^2(\mathbf{x}_j)}{n} \quad (9)$$

The number of clusters is determined by the point corresponding to the maximum PC. A example of PC results for each number of cluster in FCM is shown in Table 1.

The consequent parts of the rules in TSK FIS are determined by least square estimation method after antecedent MFs are obtained from FCM. Let X be an input matrix with rows $[x_k \; 1]$ and let W_l denote a diagonal matrix of lth rule with the degree of activation $w_l(x_k) \in R^{n \times n}$ along the diagonal. Then the consequent parameter of lth rule b_l is the least square solution of $y_l = Xb_l + \mathbf{e}$ where $\mathbf{e}$ represent error vector:

$$b_l = [X^T W_l X]^{-1} X^T W y_l \quad (10)$$



(a) b)

Figure 3: The convergence of FIS in learning (shown for male data) (a) with four rules and (b) 2 rules. Both classification error (solid line) and the ANFIS objective function (dashed line) are given.

In this work, constants are used as the consequents of the rules (zeroth-order TSK), which could represent class labels. The consequents are interpolation of the rules; therefore they are not the same as the class labels. However, if the resulting consequent values are close to given class labels, we can interpret each rule as the representation of a class [5].

2.2. Fine Tuning of the Rules

After the initial rule base is set, fine tuning of the parameters is needed. The initial model is not optimal because the MFs of the antecedents in the rules are established from the partition of input data only; accordingly, the model cannot appropriately represent the input-output relationship. Adaptive-Network-Based Fuzzy Inference System (ANFIS) method is adopted in the fine tuning of the parameters [10]. ANFIS is a neuro-fuzzy approach; i.e., the combination of fuzzy system with neural networks. `MATLAB` by MathWorks has ANFIS in its standard learning method of FIS parameters, and is used in this work. Like the initialization of FIS, hybrid learning method is used. First, the parameters of MFs in the antecedent are optimized by gradient descent method (backpropagation algorithm). Once the antecedent parameters are fixed, the consequent parameters are estimated by least square method as in the case of the initialization. This alternate step is repeated every epoch; one epoch represents a run of the whole training data set in learning the parameters. Figure 3 shows the convergence curve of classification error and objective function. The objective function is mean-squared error between the desired output and the system output.

3. Experimental Results

3.1. Data Corpus

The speech data used in the experiments were obtained from real users engaged in a spoken dialog with a machine agent over the telephone for a call center application deployed by SpeechWorks [13]. The first step was to mine this data using objective measures such as ASR accuracy, total number of dialog turns, and rejection rate to narrow down the inventory to potentially useful dialogs for our experiments. This was followed by subjective tagging of the data into one of two possible emotion categories - negative, and non-negative - by four different human listeners. One reason for considering only two classes was class-specific data sparsity. In our study, *negative* emotions represent anger and frustration in human speech, whereas *non-negative* emotions are the complement of that, i.e., they represent neutral or positive emotions such as happiness or delight. The order of utterances was randomly chosen in order for listeners not to be influenced in guessing the emotions by the situation in the dialogs (minimizing thus the effect of discourse context). Details including inter-labeller agreement are given in [14]. After the human listening tests, it turned out that most *non-negative* emotion utterances were neutral in nature, i.e., they had no apparent display of emotions. After the database preparation, we obtained 776 utterances for female speakers with 575 non-negative and 201 negative utterances and 591 for male (452 non-negative and 139 negative emotion-tagged utterances). Based on previous investigation [2], the male and female in the data corpus are considered separately.

3.2. Input Features

In our work, we computed 21 acoustic correlates including those providing prosodic information of the speech signal. The chosen features comprised utterance-level statistics corresponding to pitch (F0), energy, duration, and the first and second formant frequencies. These features are a superset of features used in published literature [1].

1. **Pitch (F0)**: mean, median, standard deviation, maximum, minimum, range (max - min), and linear regression coefficient.
2. **Energy**: mean, median, standard deviation, maximum, minimum, range, and linear regression coefficient.
3. **Duration**: speech-rate, ratio of duration of voiced and unvoiced region, and duration of the longest voiced speech
4. **Formant**: first and second formant frequencies (F1, F2), and their bandwidths (BW1, BW2).

In this study, we experimented with two sets of rank-ordered selected features; the first one had 10 best features (F10), and the other, 15 best features (F15). The best 15 chosen features, for the male and female data, were

Male Ratio of duration of voiced and unvoiced region, F0 median, energy median, energy min, F0 mean, F0 max, F0 range, energy mean, F0 STD, F0 regression coefficient, energy regression coefficient, energy STD, speech-rate, duration of the longest voiced speech, F1.

Female Ratio of duration of voiced and unvoiced region, F0 median, F0 mean, energy min, energy median, speech-rate, F1, energy regression coefficient, F0 regression coefficient, energy STD, energy max, energy mean, energy range, F0 STD, BW1, duration of the longest voiced speech, F1.

And F10 includes the first 10 best features from F15.

3.3. Experiments

In the experiments, we first downsampled the data set to equalize the number of negative and non-negative emotion classes. This downsampling procedure is rationalized by the fact that in most applications, we do not have prior knowledge of distribution such that equal prior probability is commonly assumed. 240 data set is selected 20 times in random manner, from which 200 data points are used as training and the other 40 are used as test set. The classification errors are estimated from the test data set, and the final errors are obtained by averaging over 20 independent test errors.

	Full	F15	F10
LDC	30.87 ± 17.36	13.25 ± 5.91	14.87 ± 5.29
NN	31.75 ± 6.08	27.0 ± 8.65	27.38 ± 7.41
FIS-4	21.88 ± 8.62	9.62 ± 4.75	5.38 ± 5.69
FIS-2	14.62 ± 7.88	5.63 ± 4.99	3.13 ± 3.13

(a)

	Full	F15	F10
LDC	30.87 ± 17.36	20.12 ± 10.59	12.75 ± 6.38
NN	30.5 ± 5.99	31.38 ± 6.26	23.62 ± 6.04
FIS-4	14.75 ± 8.73	6.12 ± 3.58	5.25 ± 4.13
FIS-2	10.62 ± 8.11	5.38 ± 3.74	2.25 ± 2.68

(b)

Table 2: Comparison of performance of classifiers with respect to classification errors and their standard deviations(%) in (a) male data and (b) female data, 4 classifiers are compared, i.e., linear discriminant classifier (LDC), nearest-neighborhood classifier (NN), and fuzzy inference classifier with 2 rules and 4 rules (FIS-2 and FIS-4). The results are for 3 feature sets; Full (21 full feature set), F15 (15 best feature set), and F10 (10 best feature set)

Fuzzy c-means clustering methods are applied to initialize TSK fuzzy inference system with constant consequents. The desired classification outputs are denoted 1 for negative emotion class, and -1 for non-negative emotion class. The decision c for the output of TSK fuzzy model, is classified by the following classification rule:

$$c = \begin{cases} 1, & y \geq 0 \\ -1, & y < 0 \end{cases}$$

The number of rules in the initial models are determined by maximizing PC, and each independent run generates 2 rules in its initial model. Each rule represents an emotion class, either negative or non-negative emotion. To investigate how an FIS can perform with different number of rules, FIS with four rules (FIS-4) are also generated. This initial model is optimized using hybrid method of ANFIS explained in Section 2.2.

In Table 2, we compare the results obtained from various classifiers. The results are calculated by averaging over 20 independent classification errors obtained from the test data.

4. Conclusions

Emotion recognition using a fuzzy inference system is explored in the context of automated call center applications in this paper. FIS provides improved performance compared with other classifiers such as linear discriminant and nearest-neighborhood classifiers. Results show that the optimized FIS with two rules (FIS-2) improves emotion classification by 63.0% for male data and 73.7% for female over the results from LDC. Also, as the number of input features decreases, the performance with respect to classification error has reduced. This explains that there is clearly need to apply signal processing techniques to examine the optimal input features in given applications. Interestingly, the consequent constants for rules are $y_1 = 1.25$ and $y_2 = -0.96$ from one of the optimized FIS-2, where subscripts denote the rule number; in other words, they are close to class labels for each emotion. This fact shows that each rule closely represents one of the emotion classes, which enhances the interpretation of the rules in FIS.

Since FIS is a promising tool for emotion recognition in speech, this research direction should be further explored. In the regard of future work, other sources of information in the data should be investigated and seamlessly combined with acoustic information in the context of fuzzy model. Those other kinds of information that show promise include language information, and discourse information, which were explored in a previous work under a statistical framework [15].

5. Acknowledgement

This work was supported in part by the National Science Foundation under IIS-0238514, Cooperative Agreement No. EEC-9529152 and by the Department of the Army under contract DAAD 19-99-D-0046.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Sig. Proc. Mag.*, vol. 18(1), pp. 32–80, Jan 2001.

[2] C.M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. Automatic Speech Recognition and Understanding*, Dec 2001.

[3] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Artificial Neu. Net. In Engr.(ANNIE '99)*, 1999.

[4] R. Plutchik, *The Psychology and Biology of Emotion*, HarperCollins College, New York, NY, 1994.

[5] M. Setnes and H. Roubos, "Ga-fuzzy modeling and classification: Complexity and performance," *IEEE Trans. on Fuzzy Systems*, vol. 8(5), pp. 509–522, Oct 2000.

[6] H. Wu and J.M. Mendel, "Binary classification of ground vehicles based on the acoustic data using fuzzy logic rule-based classifiers," Tech. Rep. 356, USC-SIPI, Oct. 2002.

[7] J.A. Roubos, M. Setnes, and J. Abonyi, "Learning fuzzy classification rules from data," in *RASC*.

[8] T. Takagi and M. Sugeno, "Fuzzy idenfication of systems and its application to modeling and control," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 15, pp. 116–132, 1985.

[9] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, Wiley, New York, NY, 1999.

[10] J.R. Jang, "Anfis: Adaptive-network-based fuzzy inference system," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 23(3), pp. 665–685, 1993.

[11] J.M. Mendel, "Fuzzy logic sytems for engineering: A tutorial," *Proceeding of the IEEE*, vol. 83(3), pp. 345–377, 1995.

[12] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Functions*, Plenum, New York, NY, 1981.

[13] SpeechWorks, "http://www.speechworks.com/index flash.cfm," .

[14] C.M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," Submitted to IEEE Trans. on Speech and Audio Processing.

[15] C.M. Lee, S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. ICSLP 2002*, Denver, Co, Sep 2002.