ICSV13 - Vienna
The Thirteenth International Congress on Sound and Vibration
Vienna, Austria, July 2-6, 2006

VOICE MESSAGE PRIORITIES USING FUZZY MOOD IDENTIFIER

Onsy Abdel Alim*[1], Hassan Elragal[1], and Heba M. Shehata[1]

[1]Electrical Engineering department, communication and electronics section,
Faculty of Engineering, Alexandria University
Alexandria, Egypt
(E-mail: Onsy20@hotmail.com)

Abstract

Human language carries various kinds of information. In human telephone interaction, the detections of the emotional state of a speaker leaving a voice message on the answer machine as reflected in his or her utterances is crucial for determining the priority of the messages.

This paper based on a neuro-fuzzy network that can identify the mood of the person leaving a message on the answer machine and set a priority of the message. Several basic emotions from human speech including (very serious person (something important is happening like accident ...), regular person, person in a hurry, happy person and sad person). These emotions are classified into three categories according to priority of voice messages into urgent, normal and not urgent.

Classification is carried out using a data corpus obtained from the speech analysis of 20 male and female speakers (with 10 speech samples for each speaker) to obtain the effective parameters which represent the input of the fuzzy inference system.

INTRODUCTION

Speech is the principal mode of communication between humans, both for transfer of information and for social interaction. Learning the mechanisms of speech have been of interest to scientific research, leading to a wealth of knowledge about the production of human speech, and hence to technological system to simulate and to recognize speech electronically.

Variability is the biggest problem encountered in dealing with speech, which has proven to be a major challenge to researchers. With variability, we mean that different speakers say things in different ways at both a verbal and vocal level; there is also considerable variability within the speech of a single speaker.
There are a number of reasons for this variability such as speaking style, emotion and mood, and stress.

By speaking style [1], it means that speakers can alter their way of speaking in response to a number of conditions related to their environment and their status relative to those to whom they are speaking. (A speaker will speak more carefully to a listener with whom they are not familiar) and (a speaker will speak to a child differently from the way they would speak to a grown person).

By emotion and mood [2], it means that different emotional states will affect the speech production mechanism of a speaker in different ways, and lead to acoustical changes in their speech; these changes can be perceived as being due to emotion by the listeners. Generally, emotion refers to short-term states, with mood being longer-term.
Although mood and emotion terms overlap, they are occasionally used synonymously.

By stress [3], it means a number of other factors relating to physiological arousal that contributes to changes in speech. Such factors include fatigue, illness, and the effects workload. Physical stress due to vibration or acceleration may also produce acoustic changes in speech due to direct action on the vocal apparatus itself.

Speaking style, emotion and mood, and stress are essentially independent within the speaker, but all are present to a greater or lesser extent in all speech.
The speech variability produced is generated unconsciously, and even where a speaker adopts a speaking style consciously, the actual vocal changes are made at an unconscious level. It is thus hard to quantify the changes that occur, and produce a robust description of how they are produced.

Most speech scientists have been interested in dealing with "normal speech", that is speech which does not display any of these variability. Successful work on analysis, synthesis and recognition of speech has been achieved under this constraint, but the results break down when natural variability is present (i.e. the performance of speech recognition and verification systems falls dramatically when there is variability in the incoming speech signal). Variability is present in all natural human speech, and thus if we wish to simulate natural speech, we need to incorporate variability in some way.

The motivation of this paper is of how recognition of emotions in speech can be used for sorting voice messages according to the emotions expressed by the caller.

THE PROPOSED NEURO-FUZZY IDENTIFIER SYSTEM

A simple description of the proposed neuro-fuzzy identifier system is shown in Figure 1. The system is composed of two stages; the first stage is the stage where the received

message is recorded on the answering machine and some features is extracted from the recorded message. The second stage is a neuro-fuzzy network of subtractive clustering method of classification that can identify the mood of the recorded message and produce a priority for arranging this message. The used answering machine should have memory and buffers to allow sorting of the messages according to the identified priorities.

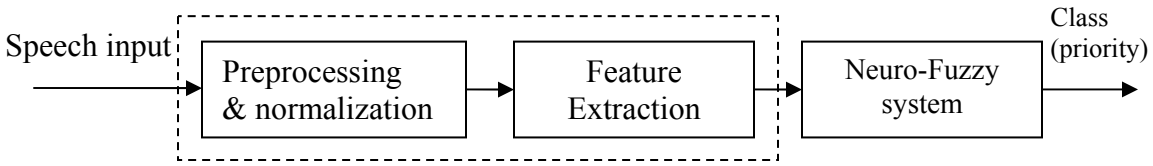

Figure 1 - The proposed neuro- fuzzy identifier system

Speech data and feature extraction

The used speech data are taken in the speech lab. The recording sampling rate was 16 KHz, mono, 16 bit. The recorded sentence was as follows:

1. Khali balak, fih egtma'a bokra. (uttered in urgent and normal moods)
2. Elarabia etelet, elhaqni. (uttered in urgent and normal moods)
3. Kalmni awel ma tewsal. (uttered in urgent and normal moods)
4. Mahadesh shafak men zaman we sama'ana sotak. (uttered in soft and normal moods)
5. Sahbak men ayam el kolya tefteker ana meen? (uttered in soft and normal moods)

These sentences are recorded for 20 different speakers (10 males and 10 females).

Feature extraction

All studies [4-7] in the field of speech analysis point to some important features including:

- The pitch (fundamental frequency) as the main vocal cue for emotion recognition.
- Temporal feature such as speech rate and pausing.
- Statistical parameters such as mean, standard deviation, minimum, maximum, and range.
- Banse and Scherer, 1996 [5] use other acoustic variables contributing to vocal emotion signaling such as vocal energy, frequency spectral features, formants (usually only one or two first formants (F1, F2) are considered).
- Tosa and Nakatsu, 1996 [6] use another approach to feature extraction by considering some derivative features such as LPC (linear predictive coding) parameters of signal.
- Dellaert et al., 1996 [7] use features of the smoothed pitch contour and its derivatives.

In this study, the recorded speech samples were divided into frames. Each frame is 160 samples or 10 milliseconds. The used variables were (energy, pitch, speaking rate

and some of the mel cepstrum coefficients *mcepi* given by relation (1) with X_k is the k[th] log-energy output [8]). Energy and pitch information were extracted form the frames. The speaking rate was calculated as the inverse of the average length of the voiced part of the utterance. Then, for each sentence in the speech data, the following statistics (mean, standard deviation, maximum, minimum and range) were calculated for each variable.

$$mcep_i = \sum_{k=1}^{20} X_k \cos(n(k-\frac{1}{2})\frac{\pi}{20}) \qquad n = 1,2,\ldots\ldots, M \qquad (1)$$

To obtain a speaker independent model a frequency normalization step is done before presenting the parameters to the identifier; to avoid the difference of fundamental frequency between male and female speakers (around 125 Hz for male speakers and 240 Hz for female speakers, the frequency is normalized by 110 Hz for male speakers and 220 Hz for female speakers .Table (1) illustrates an example of the different acoustical features obtained from one sentence of one male speaker.

Table 1 - an example of the obtained acoustical parameters
For one sentence of one male speaker

	Vocal energy	Fundamental frequency (Hz)	Speaking rate ($sec.^{-1}$)	1[st] mel cepstrum coefficient (mcep1)
Urgent mood	0.2605	159.939	1.0715	1.1493
Normal mood	0.1009	155.6758	0.9767	8.6418
Soft mood	0.0827	150.048	0.8747	5.0005

Adaptive neuro-fuzzy network for mood identification

Clustering or classification using neuro-fuzzy system has the advantage of gaining the flexibility of fuzzy inference systems as the fuzzy logic is an effective paradigm to handle uncertainty and the neuro-adaptation in the same classification model. By using given input/output data sets, we can construct a fuzzy inference system (FIS) whose membership function parameters are tuned (adjusted) using either a back propagation algorithm alone, or in combination with a least squares type of method. This allows the fuzzy systems to learn from the data they are modeling and give better results than those obtained from using the FIS alone or any other neural network.

Fuzzy inference systems [9-11] are characterized by fuzzy sets, membership functions and fuzzy if-then rules of the form “IF x is A AND y is B THEN z is C”, where A, B and C are fuzzy sets and x, y and z are members of the sets, respectively. The fuzzy sets A, B and C will each have a membership function associated with it which defines

the distribution of the membership grades for the set. Different shapes of the membership functions are given by Figure (2).

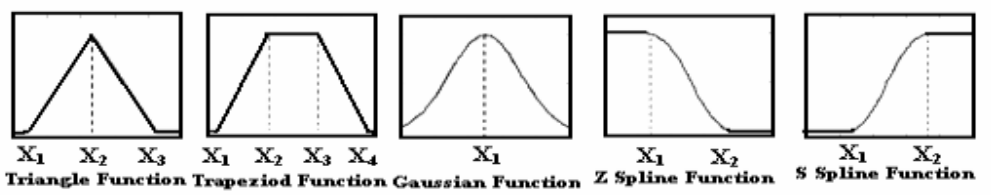

Figure 2 - Five commonly used fuzzy membership functions

Adaptive Neural-Fuzzy Inference System (ANFIS) was proposed by Roger Jang [12-13]. ANFIS is a class of adaptive multi-layer feed-forward networks that is functionally equivalent to a fuzzy inference system. It was proposed in an effort to formalize a systematic approach to generating fuzzy rules from an input-output data set. The architecture of a two-input two-rule ANFIS is shown as Figure (3).

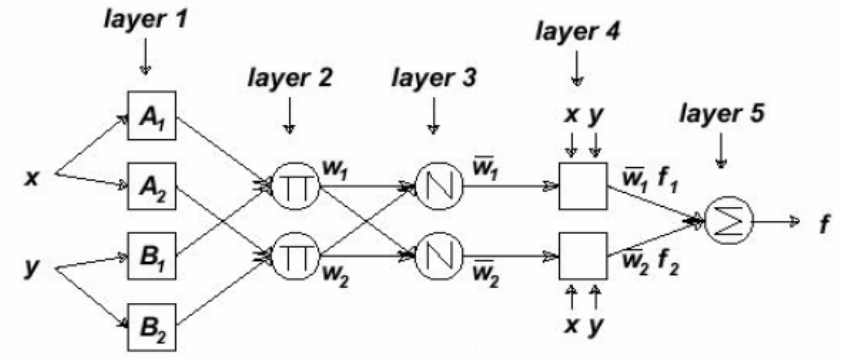

Figure 3 - ANFIS structure

The ANFIS has five layers, in which node functions of the same layer have the same function type. In layer 1; each node in this layer generates a membership grade of a linguistic label. Every node i in this layer is an adaptive node with node function shown by equation (2). Premise parameters $\{a_i,b_i,c_i\}$ updated through Back Propagation. In layer 2; calculates the firing strength of a rule via multiplication according to relation (3)

$$O_{1,i} = \mu_{A_i}(x) = \frac{1}{1+\left(\frac{x-c_i}{a_i}\right)^{2b_i}} \tag{2}$$

$$O_{2,i} = \mu_{A_i}(x) * \mu_{B_i}(y) \quad ,i = 1,2,.....M \tag{3}$$

In layer 3; every node i is a fixed node labeled N. The i^{th} node normalizes the i^{th} rule's firing strength to the sum of all rules' strengths according to relation (4). In Layer 4; node i computes the contribution of i^{th} rule toward the overall output, with node function given by relation (5). Consequent parameters $\{p_i,q_i,r_i\}$ updated through the Recursive Least- Squares Estimation

$$O_{3,i} = \overline{W}_i = \frac{W_i}{W_1 + W_2} \quad ,i = 1,2 \tag{4}$$

$$O_{4,i} = \overline{W}_i \; f_i = \overline{W}_i \; (p_i x + q_i y + r_i) \quad ,i = 1,2 \tag{5}$$

In layer 5: The single node in this layer is a fixed node, which computes the overall output as the summation of contribution from each rule. This is given by relation (6).

$$O_5 = \sum_i \overline{W}_i f_i = \frac{\sum_i W_i f_i}{\sum_i W_i} = \frac{W_1 f_1 + W_2 f_2}{W_1 + W_2} \tag{6}$$

EXPERIMENTAL RESULTS

Applying the previously mentioned extracted features on the subtractive neuro-fuzzy model, the input data are divided into 2 groups the first for training the model and the other for simulation; groupI :150 samples (15 speaker x 10 sentences), and groupII : 200 samples (The same 15 speakers x 10 sentences + 5 new speakers x the same sentences.)

The effect of some chosen acoustical features (vocal energy, fundamental frequency, speaking rate, 1^{st} mel cepstrum coefficient (mcep1) and 2^{nd} mel cepstrum coefficient (mcep2)) on the classification process is summarized in table (2). Table (3) and table (4) show the classification confusion matrix using the FIS with subtractive clustering; the used features are the vocal energy, the fundamental frequency, the speaking rate and the 1^{st} mel cepstrum coefficient.

Table 2 - the percentage of correct classification using Subtractive FIS and different acoustical features

Correct classification / Used features	For training data	For testing data
vocal energy	57 %	56 %
vocal energy and fundamental frequency	60.9 %	58 %
Vocal energy, freq. and speaking rate	92 %	82 %
Energy, freq., speaking rate and mcep1	**100%**	**86.4 %**
Energy, freq., speaking rate, mcep1 and mcep2	94 %	82 %

Table 3 - Confusion Matrix and Percentage of Correct Classification
For the training data

	Urgent	Normal	soft	Correct Classification Rate
Urgent	45	0	0	100 %
Normal	0	75	0	100 %
soft	0	0	30	100 %

Table 4 - Confusion Matrix and Percentage of Correct Classification
For the testing data

	Urgent	Normal	soft	Correct Classification Rate
Urgent	56	4	0	93.3 %
Normal	6	85	9	85 %
soft	2	7	31	85 %

The classification performance of the FIS using subtractive clustering is compared with the performance of the adaptive neuro-fuzzy system using FCM [14] clustering, with grid partitioning and the learning vector quantization (LVQ) neural network [15]. Table (5) shows percentages of the correct classification for the four different networks after 20 training epochs for training and testing data.

Table 5 – comparison between the percentages of correct classification
For the 4 different networks after 20 training epochs

	For training data	For testing data
FIS with Subtractive clustering	100%	86.4%
FIS with FCM clustering	74.33%	68.1%
FIS with grid partitioning	72.3%	66.16%
LVQ neural network	44.59%	30.3333%

CONCLUSION AND FUTURE WORK

The paper proposed a neuro-fuzzy mood identifier that can determine the mood or the emotional state of certain speaker and as a result rearrange the incoming voice massages recorded on the answering machine according to the massage priority.

The paper used the adaptive neuro-fuzzy inference system with subtractive clustering which gave the best results when compared with other systems. This is because of the chosen small value of the cluster radius and the large number of the generated fuzzy rules used in the classification process.

The results of classification using the vocal energy, speaking rate, fundamental frequency and the value of the 1st mel cepstrum coefficient as input parameters showed that using these acoustical features in any ASR system in addition to other known acoustical features will help the robustness of the system and increase its efficiency.

As a recommendation for future work, the used speech samples in the analysis may be obtained from the recordings of different answering machines and different telephone lines to study their effect on the classification process.

REFERENCES

[1] Eskénazi, M. *Trends in speaking styles research*, Proc. Eurospeech '93, Berlin, Germany, pp. 501-509, 1993.

[2] Murray, I.R. and Arnott, J.L. "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", J. Acoust. Soc. of America, 93(2), pp. 1097-1108, 1993

[3] Murray, I.R., Baber, C. and South, A. "Towards a definition and working model of stress and its effects on speech", paper in preparation.

[4] Yindong Yu, Eric Chang, Cong Li "Computer Recognition of Emotion in Speech (2002) http://citeseer.ist.psu.edu/yu02computer.html.

[5] Banse, R. and Scherer, K.R., 1996 *Acoustic profiles in vocal emotion expression*. Journal of Personality and Social Psychology. 70: 614-636.

[6] Tosa, N., Nakatsu, R., 1996 Life-like communication agent - emotion sensing character "MIC" and feeling session character "MUSE". Proc. of IEEE Conference on Multimedia 1996. 12-19.

[7] Frank Dellaert, Thomas Polzin, Alex Waibel "Recognizing Emotion in Speech" (1996) http://citeseer.ist.psu.edu/dellaert96recognizing.html.

[8] Douglas O'Shaughnessy, *Speech Communication Human and Machine*, (second edition, New York), pp. 211-216, 2002

[9] L.A. Zadeh, *Fuzzy sets*, Inf. Contr., Vol. 8, pp.338-353, 1965.

[10] E. H. Mamdani, and S. Assilian, "An Experiment in linguistic synthesis with a fuzzy logic controller," International Journal of Man-Machine Studies, vol.7, no.1, pp.1-12, 1975.

[11] T. Takagi, and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," IEEE Trans. System, Man, and Cybernetics, vol.15, pp.116-132, 1985.

[12] J.S.R. Jang; "ANFIS: Adaptive Network –based Fuzzy Inference System"; IEEE Trans. Syst. Man, Cybern, Vol. 23, No. 3, May/ June 1993.

[13] J.S.R. Jang; "Fuzzy modeling using generalized neural networks and Kalman filter algorithm"; Ninth National Conf. on Artificial Intelligence (AAAI-91), July 1991.

[14] J.C. Bezdek, (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[15] T. Kohonen. Learning vector quantization. In M. Arbib, editor, The Handbook of Brain Theory and Neural Networks, pages 537--540. MIT Press, 1995.