GFIS: Genetic Fuzzy Inference System for Speech Recognition

Washington Luis Santos Silva and Ginalber Luiz de Oliveira Serra
Federal Institute of Education, Science and Technology
Department of Electroelectronics, Laboratory of Computational Intelligence Applied to Technology
Av. Getulio Vargas, 04, Monte Castelo, CEP: 65030-005, São Luis, Maranhão, Brazil

Keywords: Recognition Speech, Fuzzy Systems, Optimization, Genetic Algorithm, Discrete Cosine Transform.

Abstract: The concept of fuzzy sets and fuzzy logic is widely used to propose of several methods applied to systems modeling, classification and pattern recognition problem. This paper proposes a genetic-fuzzy recognition system for speech recognition. In addition to pre-processing, with mel-cepstral coefficients, the Discrete Cosine Transform (DCT) is used to generate a two-dimensional time matrix for each pattern to be recognized. A genetic algorithms is used to optimize a Mamdani fuzzy inference system in order to obtain the best model for final recognition. The speech recognition system used in this paper was named Genetic Fuzzy Inference System for Speech Recognition **(GFIS)**. Experimental results for speech recognition applied to brazilian language show the efficiency of the proposed methodology compared to methodologies widely used and cited in the literature.

1 INTRODUCTION

Parameterization of an analog speech signal is the first step in speech recognition process. Several popular signal analysis techniques have emerged as standards in the literature. These algorithms are intended to produce a perceptually meaningful parametric representation of the speech signal, parameters that can emulate some behavior observed in human auditory and perceptual systems. Actually, these algorithms are also designed to maximize recognition performance (Picone, 1991),(Rabiner and Hwang, 1993). The problem of pattern recognition might be formulated as follows: Let S_k classes, where k = 1,2,3...K, and $S_k \subset \Re^n$. If any pattern space is take with dimension $\Re^x$, where $x \leq n$, it should transform this space into a new pattern space with dimension $\Re^a$, where $a < x \leq n$. Then assuming a statistical measure or second order model for each S_k, through a covariance function represented by $\left[\Phi_x^{(k)}\right]$, the covariance matrix of the general pattern recognition problem becomes:

$$[\Phi_x] = \sum_{k=1}^{K} P(S_k) \left[\Phi_x^{(k)}\right] \quad (1)$$

where $P(S_k)$ is a distribution function of the class S_k, *a priori*, with $0 \leq P(S_k) \leq 1$. A linear transformation operator through the matrix $\mathbf{A}$ maps the pattern space in a transformed space where the columns are orthogonal basis vectors of this matrix $\mathbf{A}$. The patterns of the new space are linear combinations of the original axes as structure of the matrix $\mathbf{A}$. The statistics of second order in the transformed space are given by:

$$\Phi_{\mathbf{A}} = \mathbf{A}^T [\Phi_x] \mathbf{A} \quad (2)$$

where $\Phi_{\mathbf{A}}$ is the covariance matrix which corresponds to the space generated by the matrix $\mathbf{A}$ and the operator $[\cdot]^T$ corresponds to the transpose of a matrix. Thus, it can extract features that provide greater discriminatory power for classification from the dimension of the space generated (Andrews, 1971).

One of the most widespread techniques for pattern speech recognition is the "Hidden Markov Model" (HMM) (Shenouda et al., 2006). A well known deficiency of the classical HMMs is the poor modeling of the acoustic events related to each state. Since the probability of recursion to the same state is constant, the probability of the acoustic event related to the state is exponentially decreasing. A second weakness of the HMMs is that the observation vectors within each state are assumed uncorrelated, and these vectors are correlated (Wachter et al., 2007). To overcome these drawbacks, robust recognizer has been proposed, since it has been experimentally shown that spectral variations are discriminant features for similar sounds (Fissore et al., 1997).

Luis Santos Silva W. and Serra G..
GFIS: Genetic Fuzzy Inference System for Speech Recognition.
DOI: 10.5220/0004045705360541
In *Proceedings of the 9th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2012), pages 536-541
ISBN: 978-989-8565-21-1

1.1 Proposed Methodology

In this proposal, a speech signal is encoded and parameterized in a two-dimensional time matrix with four parameters of the speech signal. After coding, the mean and variance of each pattern are used to generate the rule base of Mamdani fuzzy inference system. The mean and variance are optimized using genetic algorithm in order to have the best performance of the recognition system. This paper consider as patterns the brazilian locutions (digits): **'0 ','1 ','2 ','3 ','4 ','5 ','6 ','7 ','8 ','9 '** . The Discrete Cosine Transform (DCT) (Ahmed et al., 1974),(Zhou and Chen, 2009) is used to encoding the speech patterns. The use of DCT in data compression and pattern classification has been increase in recent years, mainly due to the fact its performance is much closer to the results obtained by the Karhunen-Loève transform which is considered optimal for a variety of criteria such as mean square error of truncation and entropy (Fu, 1968). This paper demonstrates the potential of DCT and fuzzy inference system in speech recognition (Milner et al., 1994),(Silva and Serra, 2011).

2 SPEECH RECOGNITION SYSTEM

The proposed recognition system GFIS block diagram is depicted in Fig.1.

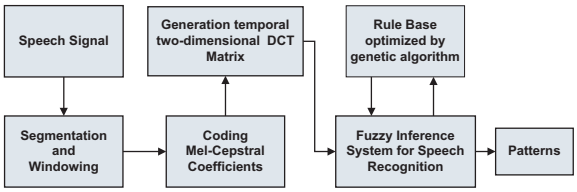

Figure 1: Block diagram of the proposed recongnition system GFIS.

2.1 Two-Dimensional Time Matrix DCT Coding

Initially, the speech signal is digitizing, so it is divided in frames which are windowed and encoded in a set of parameters defined by the order of mel-cepstral coefficients (MFCC). The DCT coefficients are computed and the two-dimensional time DCT matrix is generated, based on each speech signal to be recognized (Ariki et al., 1989),(Milner et al., 1994). Let $mfcc$ the mel-cepstral coefficients. The two-dimensional time matrix is the result of DCT in a sequence of T mel-cepstral coefficients observation vectors on the time axis, given by:

$$C_k(n,T) = \frac{1}{N}\sum_{t=1}^{T} mfcc_k(t)cos\frac{(2t-1)n\pi}{2T} \quad (3)$$

where $mfcc_k$ are the mel-cepstral coefficients, and $k, 1 \leq k \leq K$, is the k-th (line) component of t-th frame of the matrix and $n, 1 \leq n \leq N$ (column) is the order of DCT. Thus, the two-dimensional time matrix (Azar and Razzazi, 2008), where the interesting low-order coefficients k and n that encode the long-term variations of the spectral envelope of the speech signal is obtained (Fissore et al., 1997). Thus, there is a two-dimensional time matrix $C_k(n,T)$ for each input speech signal. The elements of the matrix are obtained as follows:

1. For a given spoken word ***P***(digit), ten examples of utterances of ***P***are gotten. Each examples is properly encoded in T frames distributed along the time axis;
2. Each frame of a given example of the word ***P*** generates a total of K mel-cepstral coefficients and the significant features are taken for each frame along time. The *N-th* order DCT is computed for each mel-cepstral coefficient of same order within the frames distributed along the time axis, i.e., c_1 of the frame t_1, c_1 of the frame $t_2,...,$ c_1 of the frame t_T, c_2 of the frame t_1, c_2 of the frame $t_2,...,$ c_2 of the frame t_T, and so on, generating elements $\{c_{11},c_{12},c_{13},...,c_{1N}\}$, $\{c_{21},c_{22},c_{23},...,c_{2N}\}$, $\{c_{K1},c_{K2},c_{K3},...,c_{KN}\}$, and the matrix given in equation (3). Thus, a two-dimensional time matrix DCT is generated for each example of the word ***P***, represented by: C_{kn}, where $k = 1,2,...K$ and $n = 1,2,...,N$. In this paper, a two-dimensional time matrix is generated by C^j_{kn} for each spoken word, where $j = 0,1,2...9$ and $K = N = 2$.
3. Finally, a matrix of mean and variances, for all matrices C^j_{kn} from the ten examples of spoken words used as patterns, is generated in order to produce gaussians to be used as fundamental information for implementation of the fuzzy recognition system. The means and variances to be optimized by genetic algorithm maximize the total of hits from the fuzzy recognition system.

2.2 Fuzzy Inference System for Speech Recognition Decision

Given the fuzzy set A input, the fuzzy set B output, should be obtained by the relational max-t composition(Monserrat et al., 2007). This relationship is given

by.

$$B = A \circ Ru \tag{4}$$

where Ru is a fuzzy relational rules base.

The fuzzy rule base of practical systems usually consists of more than one rule. In this paper the compositional inference is used (Wang, 1994),(Gang, 2010).

$$Ru^l{:}IF\ \ x_1\ \ is\ \ A_1^l\ \ and...and\ \ x_n\ \ is\ \ A_n^l\ \ THEN\ \ y\ \ is\ \ B^l \tag{5}$$

where A_i^l and B^l are fuzzy set in $U_i \subset \Re$ and $V \subset \Re$, and $x = (x_1, x_2, ..., x_n)^T \in U$ and $y \in V$ are input and output variables of fuzzy system, respectively. Let M be the number of rules in the fuzzy rule base; that is, $l = 1, 2, ...M$.

From the coefficients of the matrices $\mathbf{C}_{kn}^j$ with $j = 0, 1, 2, ..., 9$, $k = 1, 2$ and $n = 1, 2$ generated during the training process, representing the mean and variance of each pattern j a rule base with $M = 40$ individual rules is obtained and given by:

$$Ru^j : IF\ \ C_{kn}^j\ \ THEN\ \ y^j \tag{6}$$

In this paper, the training process is based on the fuzzy relation Ru^j using the Mamdani implication. The rule base Ru^j should be considered a relation $R(X \times Y) \rightarrow [0, 1]$, computed by:

$$\mu_{Ru}(x, y) = I(\mu_A(x), \mu_B(y)) \tag{7}$$

where the operator I should be any t-norm (Babuska, 1998), (Seki et al., 2010), (Gosztolya et al., 2009). Given the fuzzy set A' input, the fuzzy set B' output might be obtained by ***max-min*** composition, (Wang, 1994). For a minimum t-norm and max-min composition it yields:

$$\mu_{(B')} = max_x min_{x,y}(\mu_{A'}(x), \mu_{(Ru)}(x, y)) \tag{8}$$

The elements of the matrix C_{kn}^j were used to generate gaussians membership functions in the process of fuzzification. For each trained model j the gaussians memberships functions $\mu_{c_{kn}^j}$ are generated, corresponding to the elements c_{kn}^j of the two-dimensional time matrix $\mathbf{C}_{kn}^j$ with $j = 0.1, 2, 3, 4, 5, 6, 7, 8, 9$, where j is the model used in training. The training system for generation of fuzzy patterns is based on the encoding of the speech signal ***s(t)***, generating the parameters of the matrix C_{kn}^j. Then, these parameters are fuzzified, and they are related to properly fuzzified output y^j by the relational implications, generating a relational surface $\mu_{(Ru)}$, given by:

$$\mu_{Ru} = \mu_{c_{kn}^j} \circ \mu_{yj} \tag{9}$$

This relational surface is the fuzzy system rule base for recognition optimized by genetic algorithm to maximize the speech recognition. The training system is shown in Fig.2.

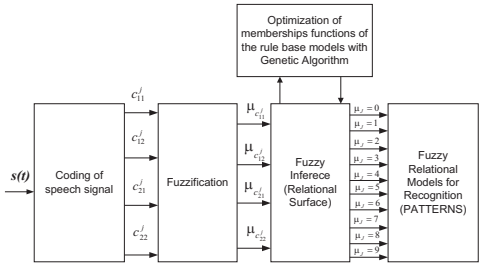

Figure 2: Generation Systems Fuzzifieds Models.

The decision phase is performed by a fuzzy inference system based on the set of rules obtained from the mean and variance matrices of two dimensions time of each spoken digit. In this paper, a matrix with minimum number of parameters (2×2) in order to allow a satisfactory performance compared to pattern recognizers available in the literature. The elements of the matrices $\mathbf{C}_{kn}^j$ are used by the fuzzy inference system to generate four gaussian membership functions corresponding to each element $c_{kn}^j \Big|^{k=1,2;n=1,2}$ of the matrix. The set of rules of the fuzzy relation is given by:

Rule Bases

$$\textbf{\textit{IF}}\ \ c_{kn}^j \Big|^{k=1,2;n=1,2}\ \ \textbf{\textit{THEN}}\ \ y^j \tag{10}$$

Modus Ponens

$$\textbf{\textit{IF}}\ \ c_{kn}^{'j} \Big|^{k=1,2;n=1,2}\ \ \textbf{\textit{THEN}}\ \ y^{'j} \tag{11}$$

From the set of rules of the fuzzy relation between antecedent and consequent, a data matrix for the given implication is obtained. After the training process, the relational surfaces is generated based on the rule base and implication method. The speech signal is encoded to be recognized and their parameters are evaluated in relation to the functions of each patterns on the relational surfaces and the degree of membership is obtained. The final decision for the pattern is taken according to the $max - min$ composition between the input parameters and the data contained in the relational surfaces.

$$\mu_{y'j} = \mu_{c_{kn}^{'j}} \circ \mu_{(Ru)} \tag{12}$$

2.3 Optimization of Relational Surface with Genetic Algorithm

The continuous genetic algorithm (Haupt and Haupt, 2004),(Tang et al., 1997) is configured with a population size of 100, generations of 300, with mutations

probability of 15% and two chromosomes, with 40 genes each, to optimize a cost function with 80 variables, which are the means and variances of the patterns to be recognized by the proposed fuzzy recognition system. The genetic algorithm was used to optimize the variations of mean and variances of each pattern in order to maximize the successful recognition process.

3 EXPERIMENTAL RESULTS

3.1 System Training

The patterns to be used in the recognition process were obtained from ten speakers who are speaking the digits 0 until 9. After pre-processing of the speech signal and fuzzification of the matrix C_{kn}^{j}, its fuzzifieds components $\mu_{c_{kn}^{j}}$ had been optimized by the GA that maximize the total of successful recognition. The optimization process was performed with 16 realizations of the genetic algorithm. The best result of the recognition processing by GFIS is shown in Fig.3. The total number of hits using GA was 92 digits correctly identified in the training process. The relational surface generated for this result was used for validation process. The best individual in the first generation is shown in Fig.4. In this case the total number of correct answers was 46 digits. The ralational surface of the best individual in the first generation is shows in Fig.5.

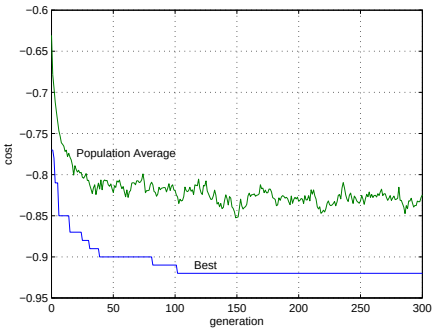

Figure 3: Plot of the best results obtained in the training process.

The optimum individual, GFIS, presents the features in Fig.6 and Fig.7.

3.2 System Test - Validation

In this step, 100 locutions uttered in a room with controlled noise level and 500 locutions uttered in an environment without any kind of noise control were used. For every ten examples of each spoken digit,

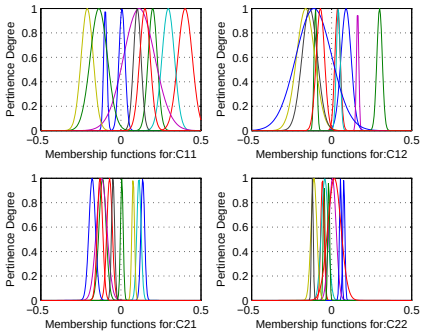

Figure 4: Membership functions for c_{kn}^{j} in the 1st generation.

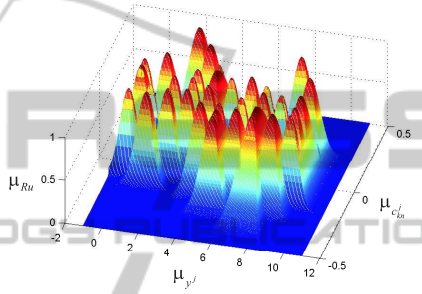

Figure 5: Relational surface (μ_{Ru}) in the 1st generation.

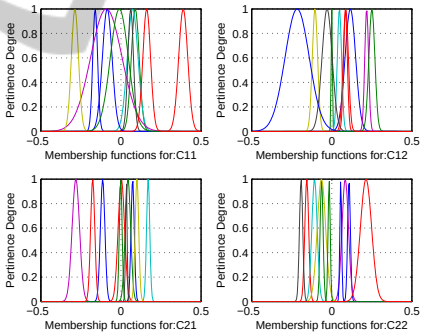

Figure 6: Membership functions for c_{kn}^{j} optmized by GA.

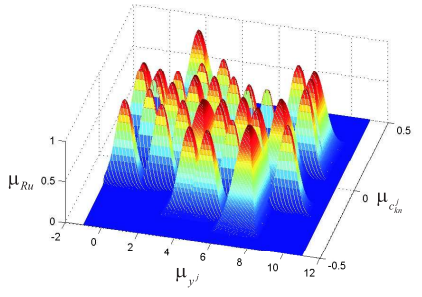

Figure 7: Relational surface (μ_{Ru}) optmized by GA.

was generated two-dimensional time matrix cepstral coefficients $\mathbf{C}_{kn}^{j}$ and they were used in the test procedure. Where performed six types of tests:

Training: Recognition Optimized by GFIS (5 Female and 5 Male Speakers)

TEST 1: Validation - Strictly speaker dependent recognition, where the words used for training and testing were spoken by a same group of 10 speakers(5 Female and 5 Male Speakers).

TEST 2: Validation test- Recognition based on the partial dependence of the speaker with two examples for each ten examples of each digit(Female Speaker).

TEST 3: Validation test- Recognition based on the partial dependence of the speaker with two examples for each ten examples of each digit(Male Speaker).

TEST 4: Validation test- Recognition independent of the Speaker, where the speaker does not have influence in the training process(Female Speaker).

TEST 5: Validation test- Recognition independent of the Speaker, where the speaker does not have influence in the training process(Male Speaker).

The tables from 1 to 6 presents the comparative analysis of the HMM with three state, three gaussian mixture by state and order analysis, i.e., the number of mel-cepstral parameter equal 8 and 12. The number of hits, using the HMM and GFIS for speech recognition. In the table for HMM, *sn = state number*, and *pn = parameters number(Mel-cepstrais coefficients)*.

Table 1: Results for the digits used in the training.

Brazilian Digits	English Digits	HMM		HMM		GFIS
		sn=3	*pn=8*	*sn=3*	*pn=12*	*pn=4*
ZERO	(zero)	9		10		10
UM	(one)	9		10		8
DOIS	(two)	7		8		10
TRES	(three)	8		9		9
QUATRO	(four)	7		8		8
CINCO	(five)	10		8		10
SEIS	(six)	7		10		10
SETE	(seven)	9		10		9
OITO	(eight)	10		10		10
NOVE	(nine)	9		8		8
Total(%)		85%		91%		92%

Table 2: Validation Test 1.

Brazilian Digits	English Digits	HMM		HMM		GFIS
		sn=3	*pn=8*	*sn=3*	*pn=12*	*pn=4*
ZERO	(zero)	9		10		9
UM	(one)	9		10		8
DOIS	(two)	7		7		9
TRES	(three)	8		8		8
QUATRO	(four)	7		8		9
CINCO	(five)	10		10		10
SEIS	(six)	7		8		9
SETE	(seven)	9		9		9
OITO	(eight)	9		9		10
NOVE	(nine)	9		9		9
Total(%)		84%		88%		90%

Table 3: Validation Test 2.

Brazilian Digits	English Digits	HMM		HMM		GFIS
		sn=3	*pn=8*	*sn=3*	*pn=12*	*pn=4*
ZERO	(zero)	9		9		10
UM	(one)	9		9		7
DOIS	(two)	6		6		7
TRES	(three)	10		9		8
QUATRO	(four)	9		9		8
CINCO	(five)	6		7		10
SEIS	(six)	6		7		6
SETE	(seven)	6		7		9
OITO	(eight)	7		8		7
NOVE	(nine)	9		9		9
Total(%)		77%		80%		81%

Table 4: Validation Test 3.

Brazilian Digits	English Digits	HMM		HMM		GFIS
		sn=3	*pn=8*	*sn=3*	*pn=12*	*pn=4*
ZERO	(zero)	7		9		8
UM	(one)	8		9		8
DOIS	(two)	7		8		10
TRES	(three)	6		8		7
QUATRO	(four)	7		8		9
CINCO	(five)	8		8		8
SEIS	(six)	8		7		9
SETE	(seven)	7		8		9
OITO	(eight)	9		9		8
NOVE	(nine)	8		9		8
Total(%)		75%		83%		84%

Table 5: Validation Test 4.

Brazilian Digits	English Digits	HMM		HMM		GFIS
		sn=3	*pn=8*	*sn=3*	*pn=12*	*pn=4*
ZERO	(zero)	6		6		10
UM	(one)	2		3		2
DOIS	(two)	4		5		5
TRES	(three)	5		5		8
QUATRO	(four)	5		7		4
CINCO	(five)	7		8		10
SEIS	(six)	4		8		5
SETE	(seven)	5		6		9
OITO	(eight)	4		6		4
NOVE	(nine)	5		5		9
Total(%)		49%		59%		66%

4 CONCLUSIONS

Evaluating the results, it is observed that the proposed speech recognizer GFIS, even with a minimal parameters number in the generated patterns was able to extract more reliably the temporal characteristics of the speech signal and produce good recognition results compared with the traditional HMM. To obtain equivalent results with HMM is necessary to increase the state number and/or mixture number. An increase in the order of the analysis above 12 does not improve significantly the performance of HMM. Any particu-

Table 6: Validation Test 5.

Brazilian Digits	English Digits	HMM		HMM		GFIS
		sn=3	*pn=8*	*sn=3*	*pn=12*	*pn=4*
ZERO	(zero)	4		5		8
UM	(one)	5		9		4
DOIS	(two)	9		9		4
TRES	(three)	3		4		3
QUATRO	(four)	4		5		5
CINCO	(five)	9		7		10
SEIS	(six)	5		6		5
SETE	(seven)	8		6		8
OITO	(eight)	9		8		10
NOVE	(nine)	6		6		10
Total(%)		62%		65%		67%

lar technique of noise reduction, such as those commonly used in HMM-based recognizers, was not used during the development of this paper. It is believed that with proper treatment of the signal to noise ratio in the process of training and testing, the GFIS Recognizer may improve its performance:

1. Increase the speech bank with different accents;
2. Improve the performance of genetic algorithm to 100% recognition in the training process;
3. Use Nonlinear Predicitve Coding for feature extraction in speech recognition;
4. Use Digital Filter in the speech signal to be recognized.
5. Increase the parameters number used.

ACKNOWLEDGEMENTS

The authors would like to thank FAPEMA for financial support, research group of computational intelligence applied to technology at the IFMA by its infrastructure for this research and experimental results, and the Master and PhD program in Eletrical Engineering at the Federal University of Maranhão (UFMA).

REFERENCES

Ahmed, N., Natajaran, T., and Rao, K. R. (1974). Discrete cosine transform. *IEEE Transaction on Computers*, C-23:90–93.

Andrews, H. C. (1971). Multidimensional rotations in feature selection. *IEEE Transaction on Computers*, C-20:1045–1051.

Ariki, Y., Mizuta, S., Nagata, M., and Sakai, T. (1989). Spoken- word recognition using dynamic features analysed by two-dimensional cepstrum. *IEEE Proceedings*, 136(v.2):133–140.

Azar, M. Y. and Razzazi, F. (2008). A dct based nonlinear predictive coding for feature extraction in speech recognition systems. *IEE International Conference on Computational Intelligente for Measurement Systems and Applications*, pages 19 – 22.

Babuska, R. (1998). *Fuzzy Modeling for Control*. Kluwer Academic Publishers.

Fissore, L., Laface, P., and Rivera, E. (1997). Using word temporal structure in hmm speech recongnition. *ICASSP 97*, v.2:975–978.

Fu, K. S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning*. Acadmic Press, New York.

Gang, C. (2010). Discussion of approximation properties of minimum inference fuzzy system. *Proceedings of the 29th Chinese Control Conference*, pages 2540–2546.

Gosztolya, G., Dombi, J., and Kocsor, A. (2009). Applying the generalized dombi operator family to the speech recognition task. *Journal of Computing and Information Technology*, pages 285–293.

Haupt, R. L. and Haupt, S. E. (2004). *Pratical Genetic Algorithms*. John Wiley and Sons, New York.

Milner, B. P., Conner, P. N., and Vaseghi, S. V. (1994). Speech modeling using cepstral-time feature and hidden markov models. *Communications, Speech and Vision, IEE Proceedings I*, v.140(5):601–604.

Monserrat, M., Torrens, J., and Trillas, E. (2007). A survey on fuzzy implication functions. *IEEE Transactions on Fuzzy Systems*, v.15(6):1107–1121.

Picone, J. W. (1991). Signal modeling techiniques in speech recognition. *IEEE Transactions*, v.79:1214–1247.

Rabiner, L. and Hwang, J. B. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey.

Seki, H., Ishii, H., and Mizumoto, M. (2010). On the monotonicity of fuzzy inference methods related to ts inference method. *IEEE Transactions on Fuzzy Systems*, v.18(3):629–634.

Shenouda, S. D., Zaki, F. W., and Goneid, A. M. R. (2006). Hybrid fuzzy hmm system for arabic connectionist speech recognition. *The 23rd National U.Jio Science Conference (NRSC 2006)*, v.0:1–8.

Silva, W. L. S. and Serra, G. L. O. (2011). Proposta de metodologia tcd-fuzzy para reconhecimento de voz. *X SBAI:Simposio Brasileiro de Automacao Inteligente*, pages 1054–1059.

Tang, C., Lai, E., and Wang, Y. C. (1997). Distributed fuzzy rules for preprocessing of speech segmentation with genetic algorithm. *Fuzzy Systems, Proceedings of the Sixth IEEE International Conference on*, v.1:427–431.

Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., and Compernolle, D. V. (2007). Template-based continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, v.15:1377–1390.

Wang, L. X. (1994). *A course in Fuzzy Systems and Control*. Prentice Hall.

Zhou, J. and Chen, P. (2009). Generalized discrete cosine transform. *Circuits, Communications and Systems, PACCS 2009, Pacific Asia Conference on*, pages 449–452.