Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx'05), Madrid, Spain, September 20-22, 2005

ADAPTIVE NETWORK-BASED FUZZY INFERENCE SYSTEM FOR AUTOMATIC SPEECH/MUSIC DISCRIMINATION

J.E. Muñoz-Expósito, S. Garcia-Galán, N. Ruiz-Reyes, P. Vera-Candeas, F. Rivas-Peña

Electronics and Telecommunication Engineering Department. University of Jaén
Polytechnic School, Linares, Jaén, SPAIN
{jemunoz,sgalan,nicolas,pvera,rivas}@ujaen.es

ABSTRACT

Automatic discrimination of speech and music is an important tool in many multimedia applications. The paper presents an effective approach based on an Adaptive Network-Based Fuzzy Inference System (ANFIS) for the classification stage required in a speech/music discrimination system. A new simple feature, called Warped LPC-based Spectral Centroid (WLPC-SC), is also proposed. Comparison between WLPC-SC and some of the classical features proposed in [11] is performed, aiming to assess the good discriminatory power of the proposed feature. The length of the vector for describing the proposed psychoacoustic-based feature is reduced to a few statistical values (mean, variance and skewness). To evaluate the performance of the ANFIS system for speech/music discrimination, comparison to other commonly used classifiers is reported. The classification results for different types of music and speech show the good discriminating power of the proposed approach.

1. INTRODUCTION

Automatic discrimination between speech and music has become a research topic of interest in the last few years. Several approaches have been described in the recent literature for different applications [1][2] [3][4][5]. Each of these uses different features and pattern classification techniques and describes results on different material.

Saunders [1] proposed a real-time speech/music discriminator, which was used to automatically monitor the audio content of FM audio channels. Four statistical features on the zero-crossing rate and one energy-related feature were extracted, a multivariate-Gaussian classifier was applied, which resulted in an accuracy of 98%.

In automatic speech recognition (ASR) of broadcast news, it's desirable to disable the input to the speech recognizer during the non-speech portion of the audio stream. Scheirer and Slaney [2] developed a speech/music discrimination system for ASR of audio sound tracks. Thirteen features to characterize distinct properties of speech and music, and three classification schemes (MAP Gaussian, GMM and k-NN classifiers) were exploited, resulting in an accuracy of over 90%.

Another application that can benefit from distinguishing speech from music is low bit-rate audio coding. Designing an universal coder to reproduce well both speech and music is the best approach. However, it is not a trivial problem. An alternative approach is to design a multi-mode coder that can accommodate different signals. The appropriate module is selected using the output of a speech-music classifier [6] [7].

Automatic discrimination of speech and music is an important tool in many multimedia applications. Khaled El-Maleh et al. [3] combined the line spectral frequencies and zero-crossings-based features for frame-level narrowband speech/music discrimination. The classification system operates using only a frame delay of 20 ms, making it suitable for real-time multimedia applications. An emerging multimedia application is content-based indexing and retrieval of audiovisual data. Audio content analysis is an important task for such application [8]. Minami et al. [9] proposed an audio-based approach to video indexing, where a speech/music detector is used to help users to browse a video database.

Comparative view of the value of different types of features in speech music discrimination is provided in [10], where four types of features (amplitudes, cepstra, pitch and zero-crossings) are compared for discriminating speech and music signals. Experimental results showed cepstra and delta cepstra bring the best performance. Mel Frequencies Spectral or Cepstral Coefficients (MFSC or MFCC) are very often used features for audio classification tasks, providing quite good results. In [4], MFSC's first order statistics are combined with neural networks to form a speech music classifier that is able to generalize from a little amount of learning data. MFCC are a compact representation of the spectrum of an audio signal taking into account the nonlinear human perception of pitch, as described by the mel scale. They are one of the most used features in speech recognition and have recently proposed in musical genre classification of audio signals [11][12].

Unlike the previous works, speech/music discrimination approaches based on only one type of features are presented in [13] and [5], which result in fast and robust classification systems. The approach in [13] takes psychoacoustic knowledge into account in that it uses the low frequency modulation amplitudes over 20 critical bands to form a good discriminator for the task, while the approach in [5] exploits a new energy-related feature, called modified low energy ratio, that improves the results obtained with the classical low energy ratio.

We present here our contribution to the design of a robust speech/music discrimination system. An effective approach based on defining in the signal analysis stage a new simple feature, called Warped LPC-based Spectral Centroid (WLPC-SC), and applying in the classification stage an Adaptive Network-Based Fuzzy Inference System (ANFIS) is proposed. The behavior of the WLPC-SC feature and the ANFIS classifier are assessed by comparison. ANFIS is compared to classical Statistical Pattern Recognition (SPR) classifiers, such as Gaussian model (GS), Gaussian Mixture Model (GMM) and k-Nearing Neighborhood (k-NN)-based classifiers. Other more complex classifiers, such as Support Vector Machines (SVM) and Radial Basis Function-based Neural Networks (RBF-

NN), are also considered.

2. SPEECH/MUSIC DISCRIMINATION

2.1. New Warped LPC-based feature

We propose the use of the centroid frequency each analysis window to discriminate between speech and music excerpts. Usually, speech signals has a low centroid frequency, which varies sharply at a voiced-unvoiced boundary. Instead, music signals show a quite changing behavior. There is no a specific pattern for such signals. We compute the centroid frequency by a one-pole lpc-filter. Geometrically, the lpc-filter minimizes the area between the frequency response of the filter and the energy espectrum of the signal. The one-pole frequency tells us where the lpc-filter is frequency-centered. Therefore, someway, the one-pole frequency informs us where most of the signal energy is frequency-localized.

However, the human auditory system is nonuniform in relation to the frequency. According to this statement, the Mel, the Bark and the ERB (Equivalent Rectangular-Bandwidth) scales [14] are defined for audio processing. For speech/music discrimination, it would be desirable to use a feature that works directly on some of these auditory scales, resulting in frequency-warped audio processing. The transformation from frequency to Bark scale is a well studied problem [14] [15]. Generally, the Bark scale is performed via the all-pass transformation defined by the substitution in the z domain

$$z = A_\rho(\zeta) \equiv \frac{\zeta + \rho}{1 + \zeta\rho} \tag{1}$$

which takes the unit circle in the z plane to the unit circle in the ζ plane, in such a way that, for $0 < \rho < 1$, low frequencies are stretched and high frequencies are compressed. Parameter ρ depends on the sampling frequency of the original signal [15]. Applying (1), the Bark scale values can be approximated from frequency positions as follows [14]:

$$b = 13arctan(0.76f(kHz)) + 3.5arctan(\frac{f(kHz)}{7.5})^2 \tag{2}$$

We propose the use of a one-pole warped-lpc filter based on this bilinear transformation to compute the WLPC-SC feature each analysis window.

The implementation of these filter can be downloaded from: *http://www.acoustics.hut.fi/software/warp* [14].

As can be seen in Fig. 1, the WLPC-SC feature shows clear differences between voiced and unvoiced phonemes due to the frequency-warped processing. Besides, these differences are bigger than in a drum-based music signal. The results in Fig. 1 suggest us that WLPC-SC could be a profitable low complexity feature to design a robust music/speech discriminator. It will be assessed in section 3.

In our system, an *analysis window* of 23 ms (1024 samples at 44100 Hz sampling rate) and a *texture window* of 1 s (43 analysis windows) are defined. Overlapping with a hop size of 512 samples is performed. Hence, the vector for describing the proposed feature consists of 86 values, which are updated each 1 s-length texture window. This large dimensional feature vector is difficult to be handled for classification tasks, giving rise to two main drawbacks: 1) too much computational cost, 2) possible too high misclassification rate. Therefore, it is required reducing the feature space to a few statistical values each 1 s-length texture window. Mean, variance and skewness of the feature vector are here computed.

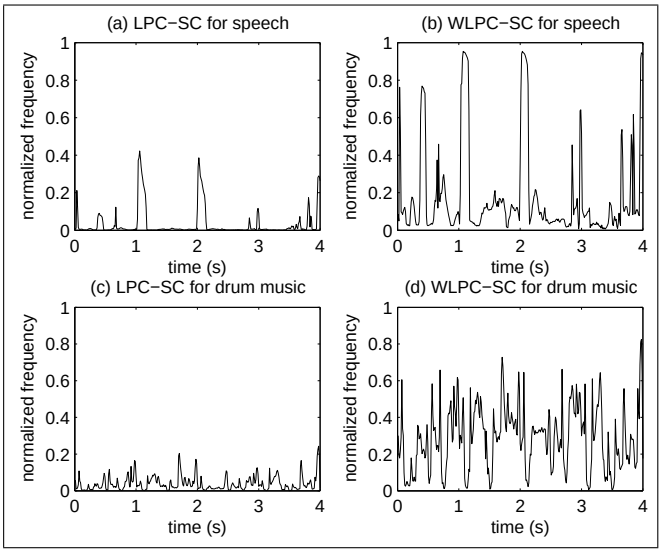

Figure 1: Example illustrating the values that LPC-SC and WLPC-SC takes for both speech and music signals.

2.2. Classification by ANFIS

ANFIS is a fuzzy inference system which is carrier out by means of adaptive networks. Using a hybrid learning procedure, ANFIS can construct an input-output mapping based on both human knowledge, in the form of fuzzy rules, and stipulated input-output data pairs.

2.2.1. Fuzzy If-Then Rules

Fuzzy rules are defined by their consequents and antecedents, which are associate to fuzzy concepts. In other words, fuzzy rules are expressions of the form IF A THEN B, where A and B are labels of fuzzy sets [16] characterized by appropriate membership functions. Due to their concise form, fuzzy rules are often employed to represent the imprecise modes of reasoning that play an essential role in the human ability to make decisions in an environment of uncertainly and imprecision.

A form of fuzzy rule which has fuzzy sets involved only in the premise part is described in [17]. A example with this kind of fuzzy rules that describes a simple fact is:

If velocity is high, then space = k(velocity)*

where *high* is in the premise part as a linguistic label characterized by an appropriate membership function. However, the consequent part is described by a non-fuzzy equation of the input variable, *velocity*. If the consequent is a linear function of the input variables, the fuzzy inference system is catalogued as one-order. If the consequent is a constant, the system is classified as zero-order.

2.2.2. Fuzzy Inference Systems

Fuzzy inference systems are also know as fuzzy rule-based systems. Basically a fuzzy inference system is composed of four

functional blocks (see figure 2):

- A *Knowledge base*, containing a number of fuzzy rules and the database, which defines the membership functions of the fuzzy set used in the fuzzy rules.
- A *Ingerence engine*, which performs the inference operations on the rules.
- A *Fuzzification interface*, which transforms the crisp inputs into degrees of match with linguistic values.
- A *Defuzzification interface*, which transforms the fuzzy results of the inference into a crisp output.

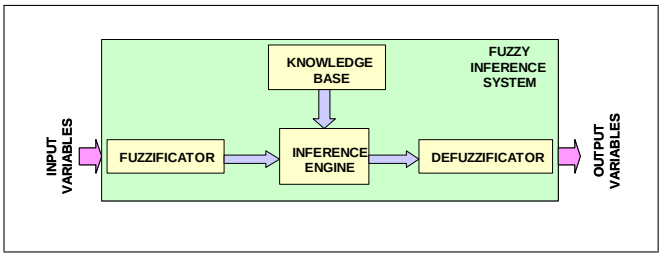

Figure 2: Fuzzy inference system

In addition to the functional blocks that compose a fuzzy inference system, two additional blocks are necessary, one at the input and another at the output. The first one (input block) allows variable magnitudes to be scaled in such way they are in the range [0,1] or [-1,1] (normalization). The second one (output block) performs the opposite operation (denormalization).

The basics of fuzzy rules and fuzzy inference systems are well known topics [16][18] [19].

2.2.3. Adaptive Networks

An adaptive network, as its name implies, is a network structure consisting of nodes and directional links through which the nodes are connected. Moreover, part or all nodes are adaptive, which means their outputs depend on the parameter/s pertaining to these nodes, and the learning rule specifies how these parameters should be changed to minimize a prescribed error measure.

An adaptive network (see figure 3) is a multilayer feedforward network in which each node performs a particular function on incoming signals as well as on a set of parameters pertaining to this node.

To reflect different adaptive capabilities, we use both circle and square nodes in an adaptive network. A square node (adaptive node) has parameters, while a circle node (fixed node) has no parameters.

Since the basic learning rule is based on the gradient method, which is notorious for its slowness and tendency to become trapped in local minima, we use a hybrid learning rule [20], which combines the gradient method and the least squares estimate (LSE) to identify the parameters and can speed up the learning process substantially.

The architecture and learning procedure of adaptive networks are well described in [21].

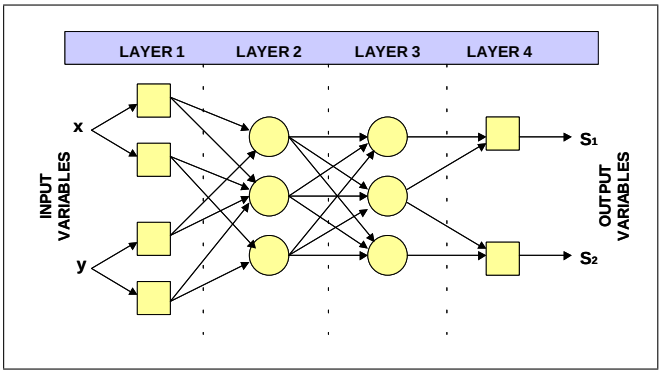

Figure 3: Adaptive network

2.2.4. ANFIS Architecture

For simplicity, we assume the fuzzy inference system under consideration has two inputs x and y and one output z. Suppose that the rule base contains two fuzzy if-then rules Takagi and Sugeno's type [17].

R1: **if** x *is* A_1 **and** y *is* B_1, **then** $f_1 = p_1x + q_1x + r_1$

R2: **if** x *is* A_2 **and** y *is* B_2, **then** $f_2 = p_2x + q_2x + r_2$

In this case, the type of reasoning is showed in figure 4, and the corresponding equivalente ANFIS architecture in figure 5.

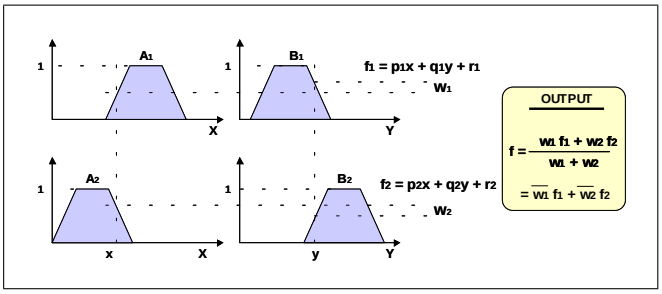

Figure 4: Fuzzy reasoning

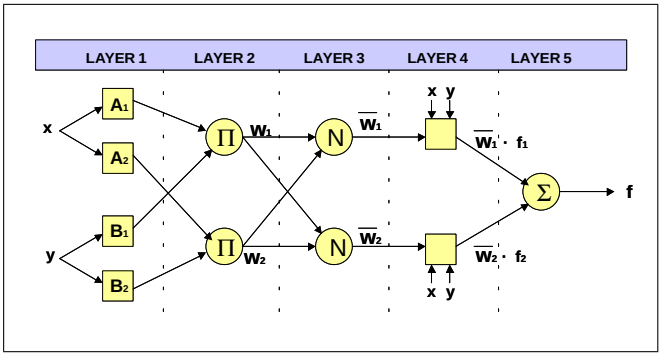

Figure 5: Equivalent ANFIS architecture

The node functions in the same layer are of the same function family as described bellow:

- *Layer 1*: Every node i is a square node with a node function $O_i^1 = \mu_{A_i}(x)$, where x is the input to node i and A_i is the

linguistic label (small, large, etc) associated with the node function. More concretely, O_i^1 is the membership function of A_i, and it specifies the degree to which the input x satisfies the quantifier A_i. Trapezoidal, triangular and bell-shaped function are membership functions commonly used.

- *Layer 2*: Every node in this layer is a circle node, labelled Π, which multiplies the incoming signals and send the product out:
  $w_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y), i = 1, 2.$
  Each node output represent the firing strength of a rule.
- *Layer 3*: Every node in this layer is a circle node labelled N. The i-th node calculates the ratio of the i-th rule's firing strength to the sum of all rules firing strengths:
  $\bar{w}_i = \frac{w_i}{w_1+w_2}, \quad i = 1, 2$
- *Layer 4*: Every node in this layer is a square node with a node function:
  $O_i^4 = \bar{w}_i \cdot f_i = \bar{w}_i \cdot (p_i x + q_i x + r_i)$
  where $\{p_i, q_i, r_i\}$ is the parameter set for each node in layer 4.
- *Layer 5*: The single node in this layer is a circle node, labelled E, that computes the overall output as the summation of all incoming signals:
  $O_i^5 = output = \sum_i \bar{w}_i \cdot f_i = \frac{\sum_i w_i \cdot f_i}{\sum_i w_i}$

Thus, we have constructed an adaptive network which is functionally equivalent to a fuzzy inference system. More information about ANFIS architecture can be found in [22].

3. EXPERIMENT EVALUATION

First of all, the audio test database is carefully prepared. The speech data come from news programs of radio and TV stations, as well as dialogs in movies, and the languages involve English, Spanish, French and German with different levels of noise, especially in news programs. The speakers involve male and female with different ages. The length of the whole speech data is about an hour. The music consists of songs and instrumental music. The songs cover as more styles as posible, such as rock, pop, folk and funky, and they are sung by male and female in English and Spanish. The instrumental music we have chosen covers different instruments (piano, violin, cello, pipe, clarinet) and styles (symphonic music, chamber music, jazz, electronic music). Some music pieces in movies are also included, which are played by multiple different instruments. The length of the whole music data is also about an hour.

Next, we intend to assess the speech/music discrimination ability of the proposed feature. To achieve such goal, comparison with the timbral features proposed in [11] is performed. The WLPC-SC feature is separately compared to all timbral texture features proposed in [11]. The vector for describing our psychoacoustic based feature consist of the mean, the variance and the skewness over each texture window.

The following specific features are used in [11] to represent timbral texture: Spectral Centroid (SC), Spectral Rolloff (SR), Spectral Flux (SF), Time Domain Zero Crossings (ZC), Mel Frequency Cepstral Coefficients (MFCC) and Low Energy (LE) feature [11]. The last one (LE) is the only feature that is based on the texture window rather than the analysis window. Note that WLPC-SC is also based on the analysis window. Table 1 shows the classification accuracy percentage results when WLPC-SC is compared to the timbral features.

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
SC	94.60	95.70	95.21
SR	94.25	94.37	94.27
SF	90.19	89.55	89.85
ZC	93.66	91.09	92.32
MFCC	96.30	96.92	96.70
LE	92.28	89.45	90.81
WLPC-SC	95.25	95.50	95.40

Table 1: Classification accuracy percentage. WLPC-SC vs. timbral features

The results in table 1 are obtained by using ANFIS as classifier, which is properly trained and adjusted. The fuzzy inference system is zero-order type, because we have considered a constant as the consequent part of the fuzzy if-then rules. We have used bell-shaped functions as membership functions, and three fuzzy sets (low, medium and high) for each input variable (mean, variance and skewness of the WLPC-SC feature computed each 1 s-length texture window). Fifty iterations have been performed for training the ANFIS system.

At the sight of the results in table 1, we can say that the proposed feature performs better than most of the timbral features in [11] for speech/music discrimination. The Spectral Centroid (SC) performs almost as well as the Warped LPC-based Spectral Centroid (WLPC-SC), while the Mel Frequency Cepstral Coefficients (MFCC) give a little better accuracy percentages. The good discrimination ability provided by the SC and MFCC features is achieved at the cost of a complexity increase regarding the WLPC-SC feature, which is much higher in the case of the MFCC feature.

Note that WLPC-SC does not require a DFT computation, while SC and MFCC need this computation. As shown in table 1, the proposed feature achieves high accuracy percentages while maintaining the complexity at a reduced degree.

The results in table 1 can be improved if transformation of the feature space is accomplished. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are very often used methods for feature space transformation, yielding LDA to higher accuracy percentages than PCA because it achieves a better class separation.

Finally, the behavior of ANFIS for speech/music discrimination is assessed by comparing it to other commonly used classifiers. Table 2 shows the classification accuracy percentages when ANFIS is compared to classical statistical pattern recognition classifiers, such as GS, GMM and k-NN based classifiers. More complex classifiers, such as SVM and RBF-NN, are also considered for comparison. The results have been obtained using the mean, variance and skewness of the WLPC-SC feature computed each 1 s-length texture window. No transformation of the feature space is accomplished, and the same audio database has been considered for testing the different classifiers.

The results in table 2 show the good behavior of ANFIS for speech/music discrimination, which implies that fuzzy-based classifiers can be an interesting alternative to classical SPR classifiers and even to more sophisticated classifiers, such as those based on SVM and RBF-NN.

CLASSIFIER	SPEECH (%)	MUSIC (%)	GLOBAL (%)
GS	87.15	95.27	91.38
GMM	94.67	92.73	93.63
k-NN	91.50	94.55	93.09
SVM	95.07	94.36	94.71
RBF-NN	94.50	94.06	94.20
ANFIS	95.25	95.50	95.40

Table 2: Classification accuracy percentage. ANFIS vs. other classifiers

As expected, classical SPR classifiers provided the worst results between all evaluated classifiers, being the GS classifier the worst ranked and the GMM one the best ranked between the SPR classifiers. Note that GMM and k-NN classifiers report very low differences. As also expected, SVM and RBF-NN behave very well for speech/music discrimination with a slight difference in favor of SVM. The highest classification accuracy percentages correspond to the ANFIS system. Slightly better results (up to 1% of improvement) can be obtained whether the number of iterations in the training stage of the system is increased (up to 300 iterations). Although ANFIS provides the best results between all the evaluated classifiers, it has an important drawback, which restricts its application scope. The complexity of the system exponentially grows with the number of input variables, becoming unfeasible when this number is high enough. However, further research has to be carried out in such direction.

4. CONCLUSIONS

The paper presents an effective approach based on an Adaptive Network-Based Fuzzy Inference System (ANFIS) for the classification stage required in a speech/music discrimination system. A new simple feature, called Warped LPC-based Spectal Centroid (WLPC-SC), is also proposed. To evaluate the performance of ANFIS for discriminating speech and music, comparison to classical SPR classifiers and more sophisticated classifiers, such as those based on SVM and RBF-NN, is performed. Each classifier is evaluated using different input parameters. The results reported are the best among the trials. ANFIS provided the best results between all the evaluated classifiers and, as expected, classical SPR classifiers provided the worst results. The classification accuracy percentage achieved by the ANFIS system is above 95% for a wide range of audio styles, which shows the good discriminating power of the proposed approach.

5. REFERENCES

[1] Saunders, J. "Real-time discrimination of broacast speech/music", *Proc. IEEE ICASSP'96*, Atlanta, USA, pp. 993-996, 1996.

[2] Scheirer, E. and Slaney, M. "Construction and evaluation of a robust multifeature speech/music discriminator", *Proc. IEEE ICASSP'97*, Munich, Germany, pp. 1331-1334, 1997.

[3] El-Maleh, K., Klein, M., Petrucci, G. and Kabal, P. "Speech/music discrimination for multimedia applications", *Proc. IEEE ICASSP'2000*, vol. 6, pp. 2445-2448, 2000.

[4] Harb, H. and Chen, L. "Robust speech music discrimination using spectrum's first order statistics and neural networks ", *Proc. IEEE Int. Symp. on Signal Processing and Its Applications*, vol. 2, pp. 125-128, 2003.

[5] Wang, W.Q., Gao, W., Ying, D.W. "A fast and robust speech/music discrimination approach", *Proc. 4th Pacific Rim Conference on Multimedia.*, vol. 3, pp. 1325-1329, 2003.

[6] ISO-IEC. "MPEG-4 Overview (ISO/IEC JTC1/SC29/WG11 N2995 Document", 1999.

[7] Tancerel, L., Ragot, S., Ruoppila, V.T. and Lefebvre, R. "Combined speech and audio coding by discrimination", *Proc. IEEE Workshop on Speech Coding*, pp. 17-20, 2000.

[8] Zhang, T. and Kuo, J. "Audio content analysis for online audiovisual data segmentation and classification", *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, 2001.

[9] Minami, K., Akutsu, A., Hamada, H. and Tonomura, Y. "Video handling with music and speech detection", *IEEE Multimedia*, vol. 5, no. 3 , pp. 17-25, 1998.

[10] Carey, M.J., Parris, E.S. and Lloyd-Thomas, H. "A comparison of features for speech, music discrimination", *Proc. IEEE ICASSP'99*, Phoenix, USA, pp. 1432-1435, 1999.

[11] Tzanetakis, G. and Cook, P. "Musical genre classification of audio signals", *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, 2002.

[12] Burred, J.J. and Lerch, A. "Hierarchical automatic audio signal classification", *Journal of the Audio Eng. Soc.*, vol. 52, pp. 724-739, 2004.

[13] Karneback, S. "Discrimination between speech and music based on a low frequency modulation feature", *European Conf. on Speech Comm. and Technology*, Alborg, Denmark, pp. 1891-1894, 2001.

[14] Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U.K. and Huopaniemi, J. "Frequency-warped signal Processing for audio applications", *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011-1031, November 2000.

[15] Smith III, J.O. and Abel, J.S. "Bark and ERB bilinear transforms", *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 697-708, November 1999.

[16] Zadeh, L.A. "Fuzzy sets", Information and Control, vol. 8, pp. 338-353, 1965.

[17] Takagi, T. and Sugeno, M. "Derivation of fuzzy control rules from human operator's control actions", *Proc. IFAC Symp. Fuzzy Information, Knowledge Representation and Decision Analysis*, pp. 55-60, 1983.

[18] Tsukamoto, Y. "An approach to fuzzy reasoning methods", Advances in Fuzzy Set THeory and Applications , pp. 137-149, 1979.

[19] Lee, C.C. "Fuzzy logic in control systems: fuzzy logic controller-Part I", IEEE Trans. System, Man and Cybern., vol. 20 , pp. 404-435, 1990.

[20] Jang, J.S.R. "Fuzzy modeling using generalized neural networks and kalman filter algorithm", Proc. Ninth Nat. Conf. Artificial Intelligence, pp 762-767, 1991.

[21] Jang, J.S.R. and Sun, C.T. "Functional equivalence between radial basis function networks and fuzzy inference systems", IEEE Trans. Neural Networks, vol. 4, pp. 156-159, 1993.

[22] Jang, J.S.R. "Adaptive-Network-based Fuzzy Inference Systems", IEEE Transactions on Systems, Manchines and Cybernetics, vol. 23, no. 3, pp. 665-685, 1993.