Fuzzy neural systems for controlling sound localisation in stereophonic reproduction

P.-R. Chang
T.-H. Tan

Indexing terms: Fuzzy neural systems, Sound localisation, Stereophonic reproduction

Abstract: The paper presents a new fuzzy logic control (FLC) approach which leads to a stereophonic reproduction controller for localising an auditory image in the desired direction and at the expected distance. Since an auditory event is usually less precisely resolved than the physical sound space, the auditory event need not coincide with a physical sound source, and can occur at a position where nothing is visible to a listener. It turns out that controlling the auditory image is more difficult than the localisation of a sound image. Unlike the conventional sound image localisation approach, our fuzzy logic controller can take account of knowledge in human auditory perception. The ambiguous human auditory perception in conjunction with the spatial reverberation from the surrounding environment can be represented by a number of fuzzy-set values. From these fuzzy representations, the auditory image localisation controller characterises the function of how control outputs depend on control inputs as fuzzy implications or associations. Furthermore, the overall stereophonic reproduction controller can be realised by a 45-rule fuzzy associative memory (FAM) system. The performance of FLC-based auditory image localisation is verified in a number of experiments.

1 Introduction

Listening to music and other acoustic signals is to have a continuum of sound locations. This includes the direct signals from the locations of the sources and the indirect or reverberant signals in the surrounding environment. Nevertheless, the number of source locations is determined and limited by the number and location of the loudspeakers. In stereophonic reproduction of music recorded in an enclosed space, the directional and distance cues of the various recorded sound sources are, to some extent, preserved. This would give illusion of location in an illusory sound space. Sounds are commonly perceived as arriving from specific directions, usually coinciding with the physical location of the sound source. The illusory perception may also carry with it a strong impression of the acoustical setting of the sound event, which normally is related to the dimensions, locations and sound-reflecting properties of the structures surrounding the listener and the physiology of the brain. The objective sound, as in sound event, refers to a physical source of sound, while the objective auditory identifies a perception. Thus the perceived location of an auditory event usually coincides with the physical location of the source. Under certain circumstances, however, the two locations may differ slightly or even substantially. The difference is then attributed to other parameters having nothing to do with the physical direction of the sound waves impinging on the ears of the listener, such as subtle aspects of a complex sound event or the processing of the sound signals within the brain.

The intent of this paper is to give some focus to the problem of synthesising a sound source in an auditory space and then localising it at a specific position in concert halls. The localisation is the rule by which the location of an auditory event (e.g. its direction and distance) is related to a specific attribute of a sound event. The cues of the sound localisation come from the comparison of the sounds at the two ears and the analysis of the difference between them. The two major parameters used to characterise sounds arriving from different horizontal angles are interaural amplitude difference (IAD) and interaural time difference (ITD). For the conventional stereophonic reproduction system, the listener would perceive a single auditory event midway between the two loudspeakers when the loudspeakers are radiating coherent sounds with identical levels and timing. The phantom or auditory sound source results from the summing localisation with IAD and ITD which is the basis for the present system of two-channel stereophonic recording and reproduction [1–3]. The impressions of the auditory source movement between the loudspeakers can be convincingly demonstrated by tuning either interchannel time or amplitude differences. In other words, both the interchannel time and amplitude differences of the stereophonic system are perceived as ITD and IAD, respectively. Moreover, [1, 3] showed that the direction of an auditory image is a function of both the interchannel time and amplitude differences. From the psychophysical standpoint, this function becomes more complicated, since the human perception system that evaluates the signals presented to the ears and that

IEE Proceedings online no. 19982109

Paper first received 3rd June 1997 and in revised form 2nd April 1998

The authors are with the Department of Communication Engineering, National Chiao-Tung University, Hsin-Chu, Taiwan

determines the direction of the auditory events cannot be regarded as linear. Moreover, it is difficult to describe the nonlinear function by an accurate expression because the localisation blur is introduced to the auditory event [1]. Thus, the attempts at determining the appropriate settings of both the interchannel time and amplitude differences in order to manipulate the localisation of auditory events are unlikely to achieve the image size and positional precision associated with the events. Sakamoto *et al.* [4] proposed a technique to localise a sound image in a desired direction from a listener by manipulating the head-related acoustic transfer functions on IAD and ITD. Actually, IAD and ITD provide the auditory system only with information on whether a sound source is to the left or right of listener. The adjustment of both IAD and ITD cannot suffice to place the sound source at a desired distance. According to a hypothesis for the psychoacoustic mechanism of distance perception of Peter Craven [5, 6], the apparent distance of sounds is derived by ITD, the values of both channel amplitude gains, and the relative amplitude ratio of the early reflection or reverberant sound to direct sound. [5] showed that the artificial distance cue can be achieved by the adjustment of both channel gains and interchannel delay without doing the recomputation of the early reflections every time. This implies that these gains and time delay can control the distribution and amplitude of direct and reverberant signals between the loudspeakers to provide the angular and distance information [6]. However, their model did not include the human auditory perception knowledge, and would degrade the performance of localisation cue. To tackle this difficulty, this paper presents the fuzzy logic localisation control systems, which provide a systematic and efficient framework for incorporating with fuzzy linguistic information from human auditory perception [7–10]. Fuzzy logic control (FLC) is a model-free approach (i.e. it does not require a mathematical model of the auditory system in conjunction with the acoustic properties of the room acoustics). The essential part of the FLC is a set of linguistic control rules related by the dual concept of fuzzy implication and the compositional rule of inference. In essence, the FLC provides an algorithm which can convert the linguistic control strategy based on the acoustic properties of the illusory auditory space into an automatic localisation control strategy. Recently, some fuzzy logic chips were designed [11] to speed up the fuzzy implication and inference processes to achieve real-time localisation.

2 Fuzzy control for an auditory image localisation

In stereophonic reproduction systems, the position of an auditory image is a function of both the right/left channel gains and interchannel time difference. However, the function becomes an expression of unknown form when the sound reproduction system involves the human auditory perception. Controlling the localisation of the auditory image by traditional controllers becomes a difficult task because the controller usually requires an explicit mathematical model of how control outputs depend on control inputs. The math-model controllers represent system uncertainty with probability distributions. Probability models describe system behaviour with the first-order and second-order statistics. They usually describe unmodelled effects and measurement imperfection with additive noise processes. Mathematical state and measurement models make it difficult to add nonmathematical human auditory perception knowledge to the system. Fuzzy controllers differ from classical math-model controller. Fuzzy controllers do not require a mathematical model of how control outputs functionally depend on control inputs. Fuzzy controllers also differ in the type of uncertainty they represent and how they represent it. The fuzzy approach represents ambiguous human auditory perception as partial implications and fuzzy set descriptions—fuzzy associations.

Fig. 1 illustrates the architecture of a stereophonic reproduction system in conjunction with the fuzzy auditory image localisation controller. Two input variables θ and r exactly describe the desired auditory image position in the horizontal plane, where θ and r represent the angle and distance from the listener to the phantom sound source, respectively. The g_R, g_L and d specify the fuzzy controller output variables, which are the inputs applied to the audio system, where g_R, g_L and d denote right/left channel gains and interchannel time difference, respectively. The goal of the fuzzy controller was to make the listener perceive the resulting auditory image of the audio signal at a desired location.

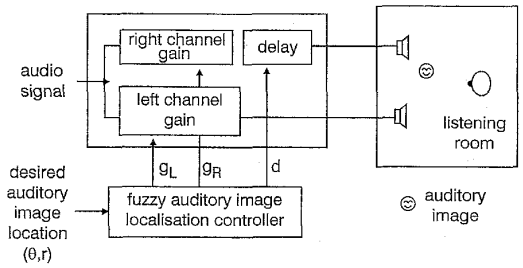

Fig. 1 *Block diagram of controlling location of auditory image*

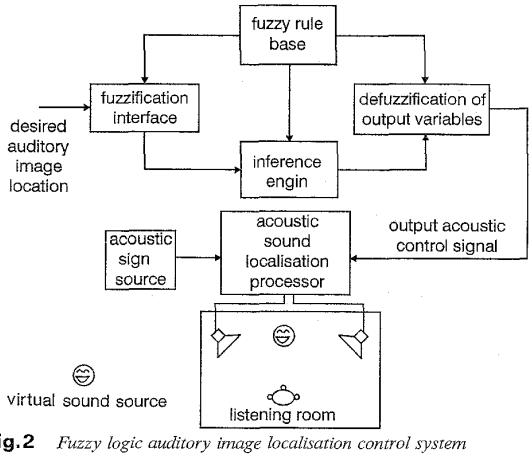

Fig. 2 *Fuzzy logic auditory image localisation control system*

2.1 Basic architecture of fuzzy logic control systems

Fig. 2 shows the basic configuration of an FLC which comprises four principal components: a fuzzification interface, a fuzzy rule base, an inference engine, and a defuzzification interface. The fuzzification interface converts the input values of the desired sound location into suitable linguistic values which may be viewed as

terms of fuzzy sets. The fuzzy rule base comprises a knowledge of the application domain and the attendant control goals. It consists of a fuzzy database and a linguistic (fuzzy) control rule base. The fuzzy database is used to define linguistic control rules and fuzzy data manipulation in an FLC. The control rule base characterises the control goals and control policy by means of a set of linguistic control rules. The inference engine is a decision-making logic mechanism of an FLC. It has the capability of simulating human auditory perception based on fuzzy concepts and of inferring fuzzy control actions employing fuzzy implication and the rules of inference in fuzzy logic. The defuzzification interface converts fuzzy control decisions into crisp nonfuzzy (i.e. physical) control signals. These control signals are applied to both the channel gains and time delay of the controller to achieve the expected auditory image location.

A fuzzy set A in a universe of discourse, U is characterised by a membership function m_A, which takes values in the interval [0, 1]; that is, $m_A : U \rightarrow [0,1]$. Thus, a fuzzy set A in U may be represented as a set of ordered pairs. Each pair consists of a generic element u and its grade of membership function; that is, $A = \{(u, m_A(u))|u \in U\}$. A linguistic variable is characterised by a quintuple $(x, T(x), U, G, \tilde{M})$ in which x is the name of the variable; $T(x)$ denotes the term set of x, that is, the set of names of linguistic values of x, with each value being a fuzzy variable denoted generically by x and ranging over a universe of discourse U which is associated with the base variable u; G is a syntactic rule for generating the name, X, of values of x; and M is a semantic rule for associating with each X its meaning, $\tilde{M}(X)$ which is a fuzzy subset of U. A particular X, that is a name generated by G, is called a term. It should be noted that the base variable u can also be vector valued. If x indicates the linguistic variable for the gain of right channel, then its term set $T(x)$ may be chosen as {zero (ZE), negative medium (NM), negative big (NB), negative very big (NV)}. In addition, g_R represents the base variable for the right channel gain in dB with its own universe of discourse $G_R = \{g_R| - 12\text{dB} \leq g_R \leq 0\text{dB}\}$. Thus M may assign a fuzzy set to the name of any term belonging to $T(x)$, for example, $\tilde{M}(NM) = \{(g_R, m_{NM}(g_R))|g_R \in G_R\}$ when X is NM, where $m_{NM}(g_R)$ is a triangular-shaped function shown in Fig. 6.

The fuzzification interface in Fig. 2 is a mapping from an input space to fuzzy sets in a certain input universe of discourse. So, for a specific value $u_i(t)$ at time instant t, it is mapped to the fuzzy set T_{xi}^1 with degree $m_{xi}^1(u_i(t))$ and to the fuzzy set T_{xi}^2 with degree $m_{xi}^2(u_i(t))$, and so on, where T_{xi}^j is the name of jth term or fuzzy-set value belonging to the term set $T(x_i)$. In the stereophonic sound localisation system, there are two input base variables (i.e. u_1 and u_2), and three output base variables, v_1, v_2, and v_3 correspond to θ, r, and g_L, g_R, d, respectively. Their corresponding term sets and membership functions will be determined in Section 3.

2.2 FAM system implementations of the inference engine and fuzzy rule base

The FAM system shown in Fig. 3 consists of a bank of fuzzy associative memory (FAM) rules or associations operating in parallel, and operating to degrees. The fuzzy system defines a mapping between an input fuzzy Cartesian product, $T(x_1) \times T(x_2) \times \ldots T(x_m)$, defined in the crisp product space $X_1 \times X_2 \times \ldots X_m$ and a single output term set $T(y)$. Thus a fuzzy system is a transformation $FS : T(x_1, x_2, \ldots x_m) = T(x_1), \times \ldots T(x_m) \rightarrow T(y)$. Usually, the number of FAM rules n in the system cannot be larger than $(p_1, \times p_2 \times \ldots \times p_m \times q)$, where p_i is number of the fuzzy-set values of $T(x_i)$, $1 \leq i \leq m$, and q is number of the fuzzy-set values of $T(y)$. Each FAM rule is a MISO (multi-input and single-output) set-level implication and denoted by $(T_{x1}^i, T_{x2}^i \ldots T_{xm}^i; T_y)$. It represents ambiguous expert knowledge on the learned input–output transformation. A FAM rule can also summarise the behaviour of human perception system. Each input fuzzy set $A(\in T(x_1) \times T(x_2) \ldots T(x_m))$ to the FAM system activates each stored FAM rule from the fuzzy rule base to a different degree. According to the correlation minimum inference discussed above, each FAM rule produces the output fuzzy set, $\hat{T}_y^i$, clipped at the firing strength, w_i determined by the input conditions, FAM rules and their associated membership functions (i.e. $M_{xj}^i(x_j)$, $1 \leq i \leq m$). $\hat{T}_y^i$ is called the partially activated version of T_y^i. The corresponding output fuzzy set T_y combines these partially activated fuzzy sets $\hat{T}_y^1$, $\hat{T}_y^2$, ... $\hat{T}_y^m$. T_y equals a weighted bounded sum of the partially activated sets:

$$T_y = \sum_{i=1}^{m} \hat{T}_y^i = \sum_{i=1}^{m} \min(w_i, T_y^i) \quad (1)$$

or equivalently:

$$M_y(\zeta) = \sum_{i=1}^{m} \min(w_i, M_y^i(\zeta)) \quad (2)$$

The partially activated output fuzzy sets $\hat{T}_y^i$ invoke the fuzzy version of the central limit theorem as the number of FAM rule increases. This tends to produce a symmetric, unimodal output fuzzy set T_y. The output fuzzy set T_y is then defuzzified to generate an exact numeral output by computing the fuzzy centroid of T_y with respective to the output universe of discourse Y.

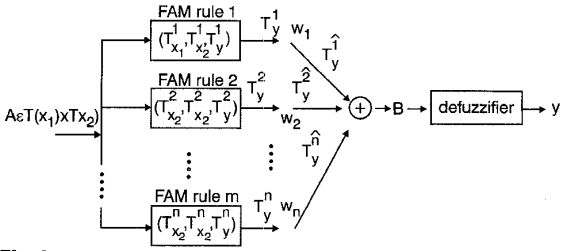

Fig. 3 *FAM system architecture*

2.3 FAM-rule generation by differential competitive learning

As mentioned above, the FAM system is the kernel of the fuzzy auditory image localisation controller. However, the user generally did not know how many FAM rules and their accurate expressions needed in performing the controller. Recently, several methods [8, 9] have been proposed to generate those FAM rules from numerical data. One of the promising methods which is called the competitive adaptive vector quantisation (AVQ) algorithm would be adopted to

our design methodology. The AVQ algorithm can adaptively estimate the unknown FAM rules from the input–output data of the audio system. More details of the competitive AVQ will be discussed in the next Section.

Suppose that the $p_i(= |T(x_i)|)$ fuzzy sets, $T^1_{x_i}$, $T^2_{x_i}$, ... $T^p_{x_i}$ quantise the ith input universe of discourse, X_i, where $|T(x_i)|$ denotes the number of terms of x_i (i.e. the fuzzy partition of input state linguistic variable x_i), $1 \leq i \leq m$, and the $q_j(= |T(y_j)|)$ fuzzy sets, $T^1_{y_j}$, ..., $T^{q_j}_{y_j}$ quantise the jth output universe of discourse, Y_j, $1 \leq j \leq l$. These quantising fuzzy sets may form a number of FAM cells which partition the input–output product space. In other words, this is called the fuzzy partition of input and output spaces.

The fuzzy Cartesian product $T^i_{x_i} \times T^i_{x_2} \times ... T^i_{x_m} \times T^j_{y_j}$ defines the FAM cell which corresponds to a possible MISO FAM rule $(T^i_{x1}, T^i_{x2}, ..., T^i_{xm}; T^j_{y_j})$. The number of all possible FAM cells is $(p_1 \times p_2 ... \times p_m \times q_j)$. Since the FAM cell may correspond to a FAM rule which is not active in the controller, the number of active FAM rules is usually less than $(p_1 \times p_2 \times ... \times p_m \times q_j)$. For a l-output system, the total number of all possible active FAM rules can not be larger than $(p_1 \times p_2 ... \times p_m \times q_1 \times ... q_l)$.

For simplicity, a simple two-input one-output case is chosen to emphasise and to clarify the ideas of generating the FAM rules by performing the competitive AVQ algorithm on a set of input–output data pair. Suppose that there is a stream of input–output data pairs generated from a product space $X_1 \times X_2 \times Y$, that is,

$$\left(x_1^{(1)}, x_2^{(1)}; y^{(1)}\right), \left(x_1^{(2)}, x_2^{(2)}; y^{(2)}\right), \cdots \quad (3)$$

where $x_1^{(i)}$, $x_2^{(i)}$ are input samples at sample time i, and $y^{(i)}$ is the output sample at sample time i.

The unsupervised competitive AVQ learning algorithm distributes the k synaptic quantisation vectors $\mathbf{m}_1, \mathbf{m}_2, ... \mathbf{m}_k$ in $X_1 \times X_2 \times Y$. Learning distributes them to different FAM cells. If there is at least one synaptic vector cluster around the centroid of a FAM cell, the FAM cell would correspond to an active FAM rule. The key idea is that cluster equals rule. Sometimes, the numbers of synaptic vectors clustered about two different centroidal FAM rules may also be difficult. Therefore, it is desirable to determine the most-frequent FAM rules or only the FAM rules with at least some minimum frequency in order to reduce the cost of implementalion. Suppose there are n FAM-rule centroids and $k > n$. Suppose k_i synaptic vectors around the ith centroid, and $k_1 + k_2 + ... + k_n = k$. The frequency of the ith FAM rule is defined by:

$$f_i = \frac{k_i}{k} \quad (4)$$

As a result, the number of quantisation vectors in each FAM cell measures the frequency of each possible FAM rule.

2.4 Stochastic competitive learning algorithms

Product-space clustering is a form of stochastic adaptive vector quantisation. Adaptive vector quantisation (AVQ) [8] systems adaptively quantise pattern clusters in $Z(= X_1 \times X_2 \times ... \times X_m \times Y) \subseteq R^N$, where $z = [x_1, x_2, ..., x_m, y] \in Z$ and $N = m + 1$. Stochastic competitive learning systems are neural AVQ systems. Neurons compete for the activation induced by randomly sampled patterns. The pattern space Z is quantised adaptively by the corresponding synaptic fan-in vectors. The k columns of the synaptic connection matrix $\mathbf{M}$ are specified by the k synaptic vectors $\mathbf{m}_j$. $\mathbf{M}$ interconnects the N inputs or linear neurons in the input neuronal field F_Z, to the k competing nonlinear neurons in the output field, F_O.

The AVQ system compares the current vector random sample $\mathbf{z}(t)$ in Euclidean distance to the k columns of the synaptic connection matrix $\mathbf{M}$, with the k synaptic vectors $\mathbf{m}_1(t) ... \mathbf{m}_k(t)$. If the jth synaptic vector $\mathbf{m}_j(t)$ is closest to $\mathbf{z}(t)$, the jth output neuron 'wins' the competition for activation at time t. The nearest or 'winning' synaptic vectors is updated by some scaled form of $\mathbf{z}(t) - \mathbf{m}_j(t)$. 'Losers' remain unchanged: $\mathbf{m}_i(t + 1) = \mathbf{m}_i(t)$. A cost-effective AVQ algorithm based on the differential competitive learning algorithm can be found in [10].

3 Estimation of the membership functions by pairwise comparison methods

Undoubtedly, the membership functions of the terms or fuzzy-set values associated with both the input and oulput lingiustic variables play an important role in FLC systems. For the fuzzification process, the evaluation of a subjective value from the crisp input value should be determined on the basis of the membership functions associated with the input linguistic variables. The crisp control signals to the plant are determined by performing the defuzzification process on the basis of the membership functions associated with the output linguistic variables. This means that the performance of an FLC system is greatly dependent on the accuracy of the estimation of those membership functions. To determine such membership functions, we first need to know how many fuzzy-set values or terms belong to the term set of each linguistic variable. Next, we may determine the shape, centre, and width of the membership function of each term for every linguistic variable. According to the characteristics of human auditory perception, it is assumed that the membership function of each term is a triangle-shaped function. The peak location of the triangle-shaped function corresponds to the centre of the membership function associated with a term. Hence, the determination of the centre of each term is equivalent to the problem of localising the peak of the triangle-shaped function. Since the membership function is triangle shaped, it allows the width of the function to be determined by the points resulted from intersecting with its adjacent terms. These intersection points can also be interpreted as the maximum ambiguity or fuzziness between the term and its adjacent terms whose degrees are all equal to 0.5.

The peak location of each term is estimated by making pairwise comparison of the elements of the universe of discouse. [12, 13] showed that the popular pairwise comparison is probably the best way to determine whether differences exist between two auditory events. Fig. 4 illustrates the configuration of the equipment layout for pairwise comparison experiment in a standard 3*3*4-m^3 listening room with two Roger 3/5 loudspeakers. A loudspeaker is mounted on the relative reference point. The other one is movable and can be moved to a point which is a candidate peak location. A test sound stimuli was presented in both loudspeakers. This test sound stimuli

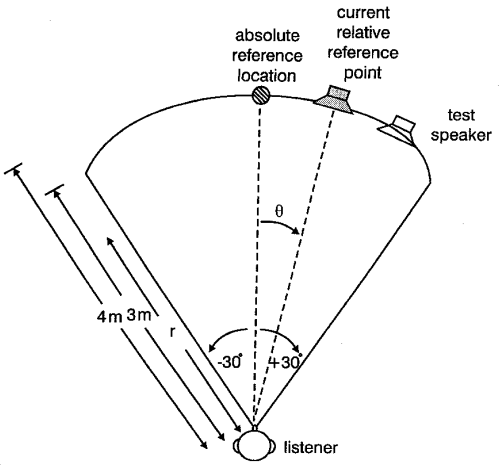

Fig. 4 *Experiment for determining membership functions by pairwise comparison method*

Table 1: Data for human perceptual direction angle discrimination

Student	LE	LC	CE	RC	RI
A	25	11	0	13	23
B	25	13	0	10	23
C	25	10	0	15	27
D	23	11	0	12	34
E	24	13	0	10	26
F	23	13	0	13	26
G	25	9	0	13	24
H	24	9	0	8	24
I	22	13	0	11	24
J	22	11	0	9	20
median	24	11	0	11.5	24

CE = centre, LC = left centre, LE = left, RC = right centre, RI = right,

is chosen as six-sec male vocal music : Track 6 'Impact 2' from Japan Audio Society Co., Ltd., GP-1086 which was generated by a Spectral sound workstation with 16-bit analogue input and output at sampling rate 44.1kHz. The polar coordinate (θ, r) can exactly describe the location or point in the horizontal plane with respect to the listener, where θ and r are the direction angle and distance of a sound source. The absolute reference point is set to be $(\theta_0, r_0) = (0°, 3m)$. At the begining, a loudspeaker is fixed in the current relative reference location coinciding with the absolute reference location. For determining the peak locations of terms of the linguistic variable 'direction angle θ' over the region of $-30° \leq \theta \leq 30°$, the movable loudspeaker which was initially mounted on the relative reference point should be moved clockwise (right direction) or counterclockwise (left direction) for every angle increment $\Delta\theta = 1°$ or $-1°$ and then compared with the sound from the loudspeaker at the relative reference point, where positive values of θ represent a clockwise rotation (right-direction), and negative values of θ represent counterclockwise rotations (left-direction). Notice that the distance of the movable loudspeaker remains at a normal distance $r_0 = 3m$ and the peak location of the initial term called CE (CEnter) is 0°, corresponding to the absolute reference point. Once the movable loudspeaker reaches the peak location of the adjacent term called the LC (left centre) for the left-side or RC (right centre) for the right side, the listener can clearly discriminate the difference between the sound directions of both loudspeakers. Experiments were conducted with ten listeners, all male, aged 21–25 years. Six listeners were graduate students, and four listeners were undergraduate students. The peak location data of both terms LC and RC are shown in the second column and the fourth column of Table 1, respectively. The actual peak locations of both terms can be estimated by a median filter which is a well-known robust estimator to location [14]. Let the n observation z_i, $1 \leq i \leq n$ be arranged on ascending order of magnitude and then written as:

$$z_{(1)} \leq z_{(2)} \leq z_{(3)} \leq \cdots \leq z_{(n)} \tag{5}$$

where $z_{(i)}$ is the so-called ith order statics. The median of z_i, $i = 1,2, ..., n$ is defined as:

$$median\left(\{z_i\}_{i=1}^{n}\right) = \begin{cases} z_{(\nu+1)} & \text{if } n = 2\nu + 1 \\ \frac{1}{2}\left(z_{(\nu)} + z_{(\nu+1)}\right) & \text{if } n = 2\nu \end{cases} \tag{6}$$

Hence, the peak locations of LC and RC are estimated as 11° and 11.5°, respectively. Next, the current relative reference point is now set to be either the estimated peak point (11°, $3m$) for LC or (11.5°, $3m$) for RC. Similarly, the peak location of the adjacent term with respect to the current reference term, which is either LC or RC, can be determined by the same technique. The peak locations of both the new adjacent terms called the LE (LEft) for left-side and RI (Right) for right-side are all identical to 24°.

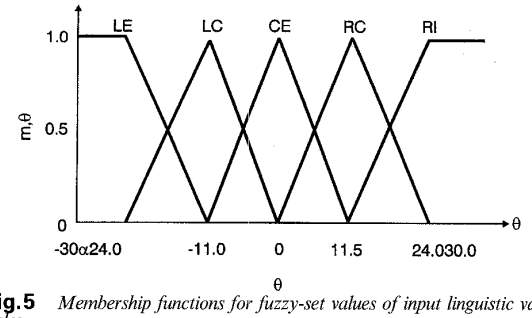

Fig. 5 *Membership functions for fuzzy-set values of input linguistic variables*
$T(0)$ = {LE, LC, CE, RC, R1}

As mentioned above, the width of membership functions of a term can be determined by finding the maximum ambiguity points between the term and its adjacent terms. For example, the maximum ambiguity point for the left-side of a test term is determined by discriminating the direction difference between the sound from the loudspeaker mounted on the peak point of the test term and the other sound from the loudspeaker which is initially placed at the peak point of the left adjacent term and then moved toward the test term for every angle increment $\Delta\theta = 1°$. If the movable loudspeaker reaches the maximum ambiguity point, the listener cannot distinguish the direction difference between them. It is found that the ambiguity point is always identical to the midpoint between the peak points of two terms. As a result, the membership functions of the term set, $T(\theta)$ = { LE,LC,CE,RC,RI }, of the linguistic variable, 'direction angle' θ, are illustrated in Fig. 5. Similarly, the term set of the distance,

r, over the region of, $2m \leq r \leq 4m$, can be obtained from the data shown in Table 2 and is identical to { NE(near), NO(normal), FA(far) }. The diagrammatic representation of those terms is shown in Fig. 6.

Table 2: Data for human perceptual distance discrimination

Student	NE	NO	FA
A	2.4	3	3.7
B	2.3	3	3.4
C	2.7	3	3.4
D	2.2	3	3.7
E	2.4	3	3.5
F	2.4	3	3.4
G	2.3	3	4.2
H	2.5	3	3.8
I	2.4	3	3.7
J	2.4	3	3.5
median	2.4	3	3.6

NE = near NO = normal FA = far

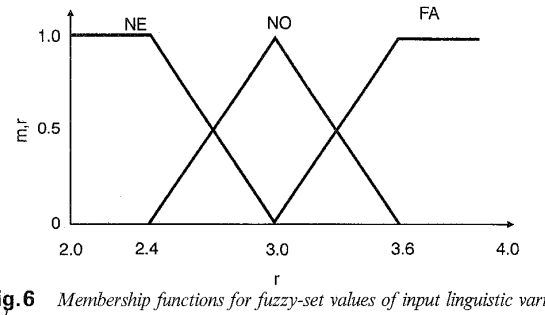

Fig. 6 *Membership functions for fuzzy-set values of input linguistic variables*
$T(0)$ = {NE (near), NO (normal), FA (far)}

The peak locations of the terms associated with the input variables, interchannel time difference d left channel gain g_L and right channel g_R, can be determined by placing two loudspeakers at the leftmost point (–30°, $3m$), and the rightmost point (30°, $3m$), and then tuning either the time difference or gain levels of both channels. Positive values of d mean that the right channel signal is faster than the left channel signal. Negative values of d represent that the left one is faster than the right one. The universe of discourse for d is $D = \{d | -40 \leq d \leq 40\}$, where the unit of d is 1/44 100 second. It is assumed that the maximum gain levels of both g_R and g_L are 0dB. The universe of discourse for both g_L and g_R are identical and given by $G_L = \{g_L | -12 \leq g_L \leq 0\}$ and $G_R = \{g_R | -12 \leq g_R \leq 0\}$, where the unit of both g_L and g_R is 1dB. The iterative procedure of determining the peak point of the current term of d is performed by changing the interchannel time difference and then comparing with the sound resulted from the peak point of the previous term, which has been already determined. Notice that the peak point of an initial term called the SM(small) is 0. If it reaches the peak point of current term, the listener will feel that there is a significant change between them. However, we found that it is difficult to find the maximum ambiguity point between a term and its adjacent term. Hence, we use a good rule of thumb that adjacent terms should overlap approximately 25% to determine the width of the associated membership functions [8]. From the above discussion, the diagrammatic representation of the term set, $T(d)$ = { NB(negative big), NM(negative medium), SM(small), PM(positive medium) PB(positive big)} is illustrated in Fig. 7. Similarly, the diagrammatic representation of the term set, $T(g_L)$ or $T(g_R)$ = {ZE(ZEro), NS(negative small), NM(negative medium), NB(negative big), NV(negative very big)} is shown in Fig. 8.

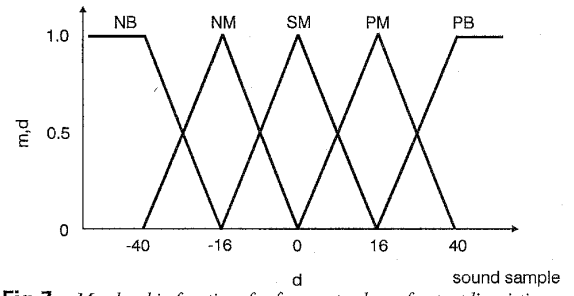

Fig. 7 *Membership functions for fuzzy-set values of output linguistic variables*
$T(d)$ = {NB (negative big), NM (negative medium), SM (small), PM (positive medium), PB (positive big)

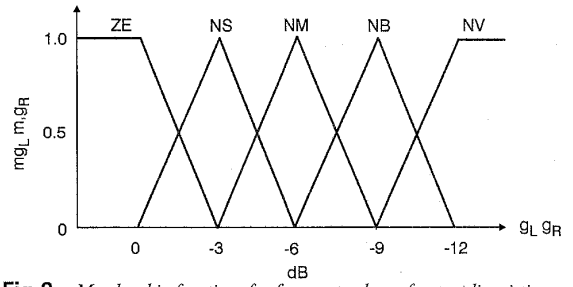

Fig. 8 *Membership functions for fuzzy-set values of output linguistic variables*
$T(g)_L$ {ZE (ZEro), NS (negative small), NM (negative medium), NB (negative big), NV (negative very big)}

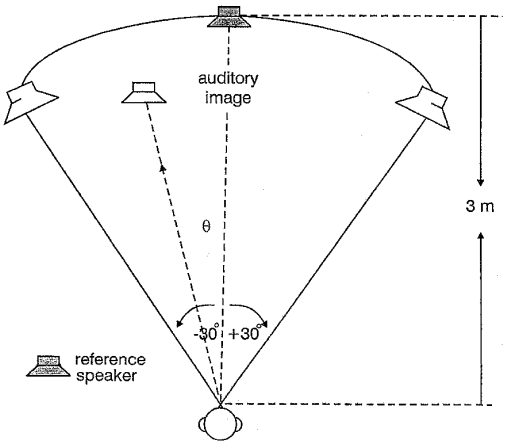

Fig. 9 *Experimental layout for FAM rule generation*

4 Product-space clustering to generate FAM rules of auditory image localisation systems

As discussed in Section 2.4, it is known that the differential competitive AVQ is able to estimate the unknown FAM rules from training data. Laboratory experiment illustrated in Fig. 9 was conducted to generate those training data. A LOGO loudspeaker is

mounted on the absolute reference point, (0°, 3*m*), and the other two Roger 3/5 loudspeakers are placed at the leftmost point, (–30°, 3*m*) and rightmost point (30°, 3*m*), respectively. The listener was instructed to indicate the direction and distance of an auditory image resulted from both the Roger 3/5 loudspeakers by comparing with the sound from the reference LOGO loudspeaker. This idea is based on the pointer methods [1] that are the most widely used in determining points of perceptual quality for spatial attributes of auditory events. However, the direction of an auditory event cannot be determined from what the subject indicates by pointing unless the relationship between the physically measurable direction of the pointer and the direction of the perceptual event corresponding to the pointer is known. A common method is to have the subject displace a movable sound source (loudspeaker) so that the auditory event appears in an agreed direction or at an agreed distance. Unfortunately, this method is not cost-effective for our laboratory. The modified method is that a listener is asked to point out the direction and distance of the auditory image by referring to the scale of both the direction angle and distance indicated on the floor and then comparing with a fixed reference sound source (LOGO loudspeaker). The procedure of generating the training data can be summarised as (i) select randomly and uniformly five gain levels from both the universes of discourse of g_R and g_L, (i.e. G_R and G_L), and five time differences from the universe of course of d, D, respectively. Thus, there are 125 (= 5 × 5 × 5) possible settings to the spectral sound workstation. By inputting the 6 s test sound stimuli into the workstation according to 125 settings, there would be 125 auditory image events, (ii) ten students are instructed to indicate the location of each auditory event and then compare with the fixed reference sound source after 1s. As a result, we will have 1250 training data for the generation of FAM rules.

The training vectors $[\theta, r, d, g_R, g_L]^T$ define points in a five-dimensional input–output product space $X_\theta \times X_r \times D \times G_R \times G_L$. θ had five fuzzy set values or terms: LE, LC, CE, RC, and RI. r had three terms: NE, NO, and FA. d had five terms: NB, NM, SM, PM, and PB. Both the g_R and g_L had identical five terms: ZE, NS, NM, NB, and NV. So there were 1875 (5 × 3 × 5 × 5 × 5) FAM cells. The space $X_\theta = \{-30° \leq \theta \leq 30°\}$ is divided into five almost-uniform intervals: [–30, –17.5], [–17.5, –5.5], [–5.5, 5.75], [5.75, 17.75], and [17.75, 30]. Each interval represents its associated fuzzy set value of five terms, LE, LC, CE, RC, and RI. This choice corresponded to the nonoverlapping intervals of the fuzzy membership function graph in Fig. 5. Similarly, the space $X_r = \{2 \leq r \leq 4\}$ can be divided into three almost-uniform intervals : [2, 2.7], [2.7, 3.3] and [3.3, 4] which corresponded, respectively, to NE, NO, and FA. $D = \{-40 \leq d \leq 40\}$ is divided into five uniform intervals: [–40, –24], [–24, –8], [–8, 8], [8, 24] and [24, 40], which corresponded, respectively, to NB, NM, SM, PM and PB. Both $G_R = \{-12 \leq g_R \leq 0\}$ and $G_L = \{-12 \leq g_L \leq 0\}$ are divided into same five nonuniform intervals: [–12, 10.5], [–10.5, –7.5], [–7.5, –4.5], [–4.5, –1.5], and [–1.5, 0].

We performed product-space clustering with the version of DCL discussed in Section 2.4 The number of synaptic vectors k may be chosen as the number of FAM cells and equals 1875. Thus, the dimension of the synaptic connection matrix **M** becomes a huge number, 5 × 1875. This will increase greatly the complexity of computing DCL algorithm. To reduce the complexity, one may eliminate the infeasible FAM rules by observing the feature of auditory events. The number of possible feasible FAM rules is estimated to be less than 50. Hence, we used 50 synaptic vectors of quantisation to estimate the FAM rules. The DCL algorithm classified each of the 1250 training input–output data vectors into one of the 50 FAM cells. We added a FAM rule to the FAM system if a DCL-trained synaptic vector fell into the FAM cell.

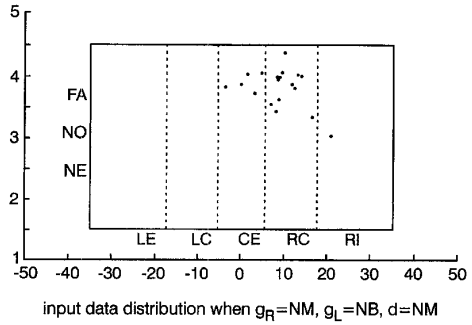

Fig. 10 *Input data distribution when gR = NM, gL = NB, and d = NM*

Fig. 10 shows the input sample distribution $[\theta, r]^T$ which is the two-dimensional projection of the five-dimensional input–output product space when d = NM, g_R = NM, and g_L = NB. Performing DCL on this input sample distribution, it yields a synaptic vector which fell into a projected FAM cell, [5.75, 17.75] × [3.3, 4], where [5.75, 17.75] and [3.3, 4] correspond to θ = RC and r = FA, respectively. This clustered FAM cell corresponds to an MIMO FAM rule, (RC, FA; NM, NM, NB). Since the three outputs of an MIMO rule are independent, this rule can be decomposed into three MISO FAM rules (RC, FA; NM) for output variable d, (RC, FA; NM) for g_R, and (RC, FA; NB) for g_L. For the five-dimensional product space, it is found that most FAM cells do not generate FAM rules. DCL distributed the 50 synaptic vectors to the most frequent 15 FAM cells. According to the above discussion, we have 45 MISO FAM rules which are represented by three FAM-bank matrices and their corresponding control surfaces shown in Figs. 11, 12, 13, 14, 15 and 16, respectively. The control surfaces are used to define the input–output transformation of a control system.

r \ θ	LE	LC	CE	RC	RI
FA	NV	NB	NM	NM	NM
NO	NB	NM	NS	NS	NS
NE	NM	NS	ZE	ZE	ZE

Fig. 11 *FAM bank for fuzzy auditory image localisation control system when output variable is g_R*

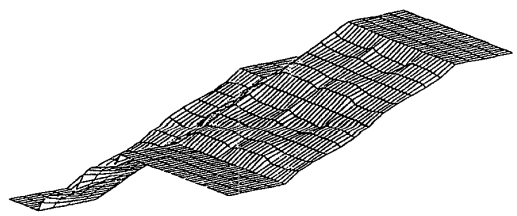

Fig. 12 *Corresponding control surface for conditions in Fig. 11*

r \ θ	LE	LC	CE	RC	RI
FA	NM	NM	NM	NB	NV
NO	NS	NS	NS	NM	NB
NE	ZE	ZE	ZE	NS	NM

Fig. 13 *FAM bank for fuzzy auditory image localisation control system when output variable is g_L*

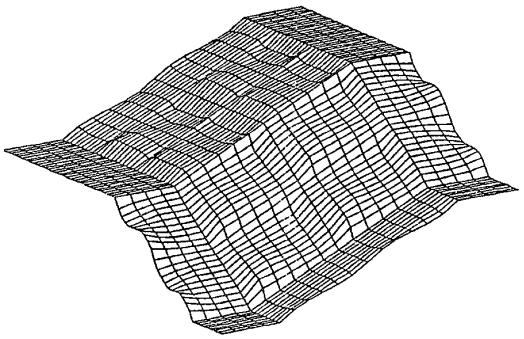

Fig. 14 *Corresponding control surface for conditions in Fig. 13*

r \ θ	LE	LC	CE	RC	RI
FA	NB	NM	ZE	PM	PB
NO	NB	NM	ZE	PM	PB
NE	NB	NM	ZE	PM	PB

Fig. 15 *FAM bank for fuzzy auditory image localisation control system when output variable is d*

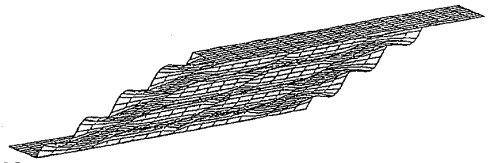

Fig. 16 *Corresponding control surface for conditions in Fig. 15*

5 Performance verification of FLC-based auditory image localisation

To evaluate the performance of controlling auditory localisation by the proposed 45-rule FLC system, two different desired input auditory locations (4°, 3.1*m*) (near to the forward axis and outside the reference ring) and (–22°, 2.7*m*) (near to the leftmost axis and inside the reference ring) are considered in our system. Fig. 17 shows an example of correlation-minimum inference for the four FAM rules of left channel gain g_L followed by centroid defuzzification of the combined output fuzzy sets when the desired input (4°, 3.1*m*) is applied to the FLC system. These four FAM rules are *R*1 : (CE, NO; NS), *R*2 : (CE, NE; NM), *R*3 : (RC, NO; NM), and *R*4 : (RC, NE; NB). The antecedent of each FAM rule cojoins θ and r fuzzy-set values. The scalar firing strength w_i of the ith FAM rule's consequent equals the minimum of both the antecedent conjuncts' values. For example, from Fig. 17, the FLC system activates the consequent fuzzy set NS of the 1st FAM rule (CE, NO; NS) to degree w_1 = min (0.6, 0.8) = 0.6. Moreover, applying w_1 to the consequent fuzzy set of rule 1 results in the shaded trapezoid shown in Fig. 17. Similarly, the firing strengths for FAM rules, R_2, *R*3, and *R*4 are w_2 = 0.1, w_3 = 0.3 and w_4 = 0.1, respectively. Applying each firing strength to the consequent fuzzy set of its associated rule would result in a similar trapezoid with different size. By superimposing the resulting memberships over each other and using the bounded sum operator, the membership function for the combined conclusion of these rules is found (as shown the lower right-hand side of Fig. 17). Furthermore, using the fuzzy centroid defuzzification, the defuzzified value for the conclusion is found as g_L = –4.94dB. Similarly, by applying the same correlation-minimum inference procedure to the right channel gain and interchannel time difference, g_R and d can be found as –3.75dB and 6 time units, respectively. For the desired input (–22°, 2.7*m*), it can be found that the crisp control signals g_L, g_K and d are –1.5dB, –7.5dB and –32 time units, respectively. A test sound stimuli is applied to the Spectral sound workstation according to the resulted crisp control signals. Ten students are instructed to indicate the direction and distance of the resulting auditory events according to the reference direction angle and distance labelled on the floor. To improve the accuracy of identifying both direction and distance, one may compare the estimated location with a fixed reference sound source. The resulted data for both cases are shown in Table 3. The deviation between the desired location (4°, 3.1*m*) and the mean average of those data, (7.8°, 2.75*m*) is 3.8° for θ and 0.35*m* for r. Similarly, the deviation for the other desired auditory location, (–22°, 2.7*m*) is 3° and 0.13*m*. Figs. 18 and 19 show a typical example of the experiments in the above fuzzy sound localisation.

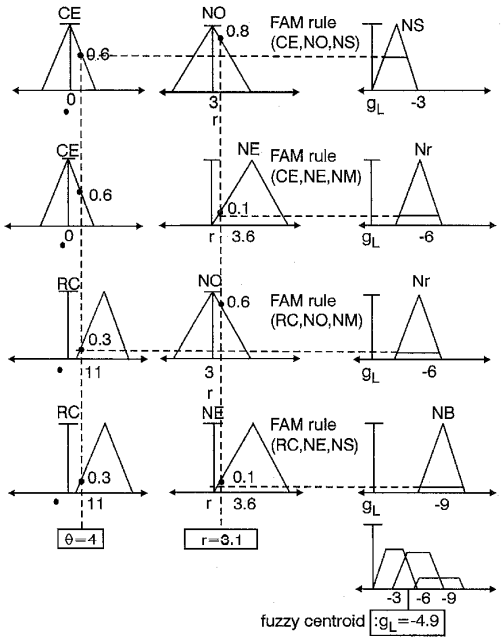

Fig. 17 *Correlation-minimum inference of four activated FAM rules of g_L followed by centroid defuzzification*

In normal hearing, the precision with which we are able to identify the direction and distance of sounds depends on a number of factors. The measure of precision is called localisation blur, the smallest

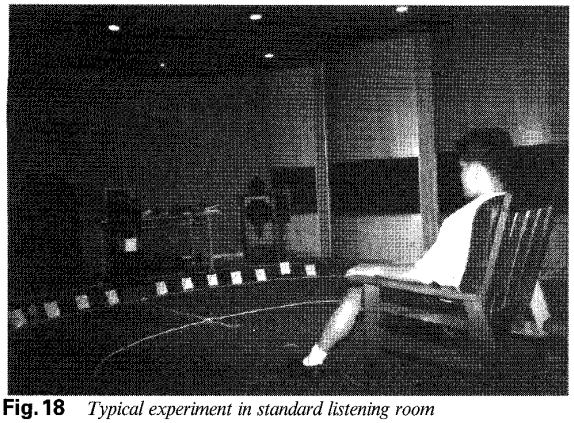

Fig. 18 *Typical experiment in standard listening room*

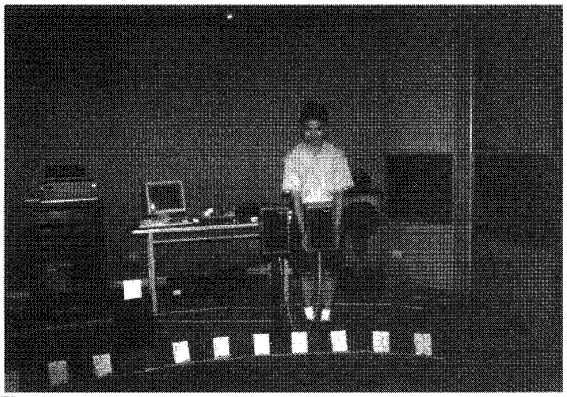

Fig. 19 *Typical experiment in standard listening room (continued)*

displacement of the sound event that produces a just-noticeable difference (JND) in the corresponding auditory event. The concept of localisation blur characterises the fact that auditory space (the perception) is less precisely resolved than the physical sound space [1, 2]. Hence, the deviations between the resulting auditory image locations and the expected sound locations are acceptable.

6 Conclusions

In this paper, we presented a novel fuzzy logic controller design methodology for constructing the auditory image localisation controllers in the stereophonic reproduction system. The fuzzy logic controller does not require an accurate mathematical model of the system under control and is capable of incorporating the human auditory perception into the controllers. We have shown that the human auditory perception knowledge can be represented by a bank of 45 FAM rules which is the essential part of FLC-based localisation system. These 45 FAM rules were generated by the DCL clustering technique on the basis of 1250 training data obtained from ten human listeners. The fuzzy-set values of each linguistic variable associated with the FAM rules were derived by applying the pairwise comparison method to auditory events. Thus, it results in three FAM banks and their corresponding fuzzy control surfaces in order to yield the appropriate settings of interchannel time delay and both channel gains of the reproduction system and achieve the desired auditory image localisation. Two experimental tests were conducted to demonstrate that the accuracy of the proposed 45-rule localisation controller is allowable in accordauce with the JND of human localisation blur.

Table 3: Data for the evaluation of the FLC-based auditory image localisation

Student	direction, deg.	distance, m
A	5	2.8
B	6	2.4
B	11	2.4
D	10	3.1
E	8	3.0
F	9	2.3
G	10	2.8
H	7	2.9
I	4	2.8
J	8	2.7
mean	7.8	2.75
desired	4	3.1
error	3.8	0.35
Student	direction, deg.	distance, m
A	−25	2.8
B	−23	2.6
B	−23	2.4
D	−27	2.4
E	−25	2.6
F	−24	2.2
G	−26	2.9
H	−27	2.6
I	−24	2.4
J	−26	2.8
mean	−25	2.6
desired	−22	2.57
error	3	0.13

7 References

1 BLAUERT, J.: 'Spatial hearing' (MIT., Cambridge, MA., 1983)
2 MAKOUS, J.C., and MIDDLEBROOKS, J.C.: 'Two-dimensional sound localization by human listeners', *J. Acoust. Soc. Am.*, 1990, **87**, pp. 2188–2200
3 MAN'KOVSKII, V.S.: 'Localization of the virtual sound source in two-channel stereophonic transmission', *Sov. Phys. Acoust.*, 1959, **2**, pp. 256–261
4 SAKAMOTO, N., GOTOH, T., KOGURE, T., and SHIMBO, M.: 'Controlling sound-image localization in stereophonic reproduction (part I)', *J. Audio Eng. Soc.*, 1981, **29**, pp. 794–799
5 GERZON, M.A.: 'The design of distance panpots'. 92nd AES convention, 1992, pp. 1–28
6 CHOWNING, J.M.: 'The simulation of moving sound sources', *J. Audio Eng. Soc.*, 1970, **19**, pp. 692–694
7 LEE, C.C.: 'Fuzzy logic in control systems: fuzzy logic controller : part I and II', *IEEE Trans.*, 1990, **SMC–20**, (2), pp. 404–435
8 KOSKO, B.: 'Neural networks and fuzzy systems' (Prentice Hall, Englewood Cliffs, NJ), pp. 315–316, 343
9 KONG, S.G., and KOSKO, B.: 'Adaptive fuzzy systems for backing up a truck-and-tailer', *IEEE Trans.*, 1992, **NN–3**, (2), pp. 211–223
10 KONG, S.G., and KOSKO, B.: 'Differential competitive learning for centroid estimation and phoneme recognition', *IEEE Trans.*, 1991, **NN–2**, pp. 118–124
11 TOGAI, M., and WATANABE, H.: 'Expert system on a chip: an engine for real-time approximate reasoning', *IEEE Expert*, 1986, **1**, (3), pp. 55–62
12 LIPSHITZ, S.P., and VANDERKOOY, J.: 'The great debate: subjective evaluation', *J. Audio Eng. Soc.*, 1981, **29**, pp. 482–491
13 CLARK, D.: 'High-resolution subjective testing using a double-blind comparator', *J. Audio Soc. Eng.*, 1982, **30**, pp. 330–338
14 ARCE, G.R., GALLAGHER, N.C., and NODES, T.A.: 'Median filter: theory for one- and two-dimensional filters', *in* HUANG, T.S. (Ed.): 'Advances in computer vision and image processing' (JAI Press, 1986)