NN and Hybrid Strategies for Speech Recognition in Romanian Language

Corneliu-Octavian Dumitru and Inge Gavat

Faculty of Electronics Telecommunications and Information Technology
Politehnica University Bucharest, 313 Splaiul Independetei, Bucharest, Romania
{odumitru, igavat}@lpsv.pub.ro

Abstract. In this paper we present results obtained with learning structures more "human likely" than the very effective and widely used hidden Markov model. Good results were obtained with simple artificial neural networks like the multilayer perceptron or the Kohonen maps. Hybrid structures have proven also their efficiency, the neuro-statistical hybrid applied enhancing the digit recognition rate of the initial HMM. Also fuzzy variants of the MLP and HMM gave good results in the tested tasks of vowel recognition.

1 Introduction

To make a first step on the way to bring near to HSR (Human Speech Recognition) the ASRU (Automatic Speech Recognition and Understanding) performance, it could be important to compare how speech recognition, as the receiving part of the verbal communication, is realised by machines and by humans [8]. Even though the process of verbal communication is a very natural one and thus seems to be a fairly easy for humans, there are several underlying operations that need to be carried out before the communication can be considered successful. The operations designed for ASR try to model what we know about speech and language, or what we assume about it. The models are often much simpler than the reality, and thus are imperfect.

In Fig. 1, a schematic overview of both speech recognition processes, the HSR and ASRU is shown [6]. The first operation in human communication comprises the hearing of the message. We first have to realise that somebody is talking to us and then we have to listen to what he is saying. In ASR, an equivalent process is done, by recording the message with a microphone.

Both systems, human and automatic, need to have some knowledge about the sounds that are used. If one of the talkers uses sounds the other talker does not know, they cannot understand each other. For this we need a vocabulary that is a set of words.

When humans process the message, they extract the meaning out of what was said. They can do so by inferring the meaning from the actual sequence of words that they recognised, because they can directly associate meanings to words and more important to word sequences.

The system however, searches for the word or word sequence that was most likely spoken, given acoustic signal. Even if this was successful, it is still far away from understanding what the meaning of this sequence of words is. Of course, approaches already exist that try to extract the meaning out of recognised word sequences.

Although the procedures in the human and the automatic systems seem to be very likely, the results are very different, the differences being pointed by Lippmann in his well–known study [11]. For complicated tasks, involving sentence recognition, the performance difference is not so surprising and can mainly be explained by the advantage constituted by the natural context use of humans [3]. For simple tasks, like vowel recognition for instance when the context is not advantaging humans, they are indeed better than the machine and that remains surprising [4].

Because of the good human performance in the speech recognition task, it could be interesting to mimic the most important human action in this process, namely the learning and to do it also in a more "human likely" maner by involving neuronal and fuzzy techniques.

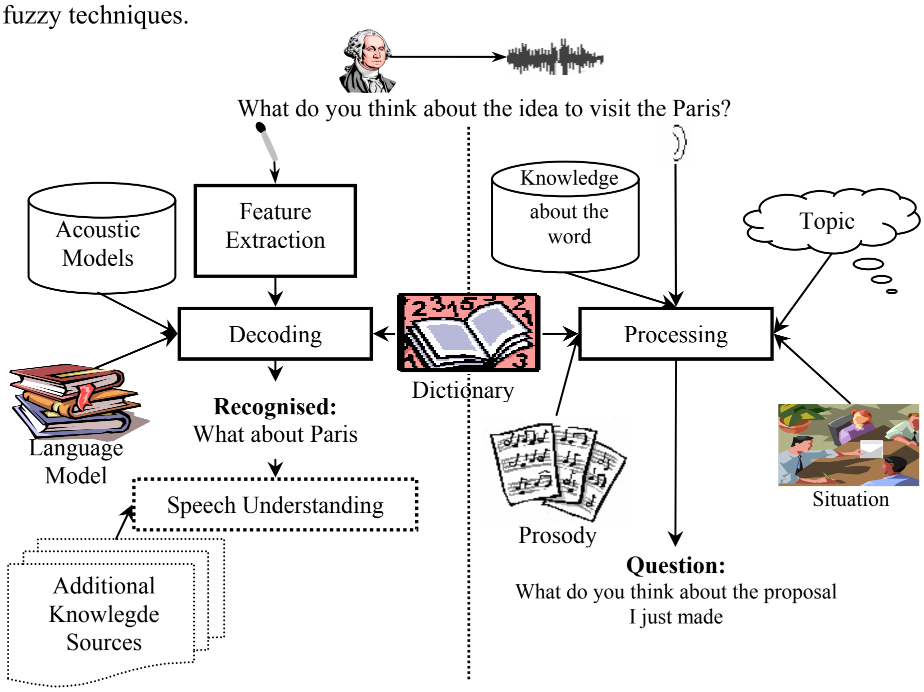

Fig. 1. Overview of a HSR system (right side) and an ASRU system (left side).

The paper has the following structure. Section 2 will describe the basic learning structures applied to build acoustical models. There are investigated classical neural strategies, like the multilayer perceptron (MLP) and the Kohonen maps (KM) but also hybrid strategies like a neuro-statistic model or fuzzy variants of a neural structure, namely the MLP and of a statistical structure, namely the HMM. The Section 3 presents a large variety of conditions into which can be carried out the basic experiments with our Automatic Speech Recognition System for Romanian Language (ASRS_RL) in tasks for vowel recognition and digit recognition. Section 4 concludes the paper.

2 Learning Strategies

Learning is the basic process for humans in acquiring knowledge and was successfully mimicked in technical systems. Artificial Neural Networks generating the neural strategies are a good example. They lead to good recognition performance and can also improve through hybridization the performance of the statistical method based on HMM.

2.1 Neural Strategies

The neural strategies can model very well an important characteristic in human learning, namely associativity: all inputs of the ANN concur to obtain the resulting output, and therefore the recognition performance is high. Due to the fixed number of inputs their flexibility in accommodating time sequences is to low for more complicated recognition tasks, like continuous speech recognition, but appropriate to recognize vowels or digits. Further we will discuss two fundamental artificial networks, namely the MLP and the KM.

Multilayer Perceptron (MLP)

MLP is the most common ANN architecture used for speech recognition. Typically, MLPs have a layered feed–forward architecture, with an input layer, one or more intermediate (hidden) layers, and one output layer. The structure without hidden layer is called Boolean network and is a simple perceptron [13].

Each layer computes a set of linear discriminative functions, followed by a non-linear function, which is often a sigmoid function.

The numbers of neurons in the hidden layer was experimentally determined, trying to achieve an optimum between the following two opposite requirements: (a) lower computing volume and more rapid process of convergence in the learning period; (b) better performances from the correct classification of input patterns percentage.

In the learning phase are determined the optimum values [14] for weights connecting the pairs of neurons from the adjoint layers in the input–output direction using the Back-Propagation algorithm.

Kohonen Maps

Kohonen maps are competitive neural networks with topological character. The setting up of the winner neurons at output is done with keeping the topological relations between the input vectors.

That is the reason for which this neural network is successfully used in pattern recognition [9]. In the learning phase the structures are trained and the weights of the networks are established in two steps: (a) the determination of the winner neurons; (b) the adaptation of the weights for the winner neurons and for the neurons existing in a certain neighborhood. In this step important are: (a) the neighborhood dimension $r(t)$ decreasing during the learning; (b) the learning rate $\eta(t)$ following in our experiments one of the laws [5]:

$$\eta(t) = t^{-1} \text{ or } \eta(t) = t^{-1/2} \qquad (1)$$

2.2 Hybrid Strategies

Hybrid strategies can improve the performance of the emerging ones. In order to make they more “human likely”, we have applied a neuro-statistical hybrid, adding to a HMM a MLP as *a posteriori* probability estimator and realizing fuzzy variants of a MLP and a HMM.

Neuro-statistic Hybrid (HMM –MLP)

The HMM-based speech recognition methods make use of a probability estimator, in order to approximate emission probabilities $p(x_n/q_k)$, where x_n represents the observed data feature, and q_k is the hypothesized HMM state. These probabilities are used by the basic HMM equations, and because the HMM is based on a strict formalism, when the HMM is modified, there is a great risk of losing the theoretical foundations or the efficiency of the training and recognition algorithms. Fortunately, a proper use of the MLPs can lead to obtain probabilities that are related with the HMM emission probabilities [10].

In particular, MLPs can be trained to produce the *a posteriori* probability $p(x_n/q_k)$, that is, the *a posteriori* probability of the HMM state given the acoustic data, when each MLP output is associated with a specific HMM state. Many authors have shown that the outputs of an ANN used as described above can be interpreted as estimates of *a posteriori* probabilities of output classes conditioned by the input, so we will not insist on this matter, but we will mention an important condition, useful for finding an acceptable connectionist probability estimator: the system must contain enough parameters to be trained to a good approximation of the mapping function between the input and the output classes [14].

Thus, the *a posteriori* probabilities that are estimated by MLPs can be converted in emission probabilities by applying Bayes' rule (2) to the MLP outputs:

$$\frac{p(x_n / q_k)}{p(x_n)} = \frac{p(q_k / x_n)}{p(q_k)} \tag{2}$$

That is, the emission probabilities are obtained by dividing the *a posteriori* estimations from the MLP outputs by estimations of the frequencies of each class, while the scaling factor $p(x_n)$ is considered a constant for all classes, and will not modify the classification.

This was the idea that leads to hybrid neuro-statistical methods, that is, hybrid MLP-HMM methods, applied for solving the speech recognition problem.

Fuzzy Variants

Human judgement is rarely a binary one, and therefore binary logic even if very simple, is not the best solution to model human acting in speech classification tasks. It seems that fuzzy logic, able to a nuanced, shaded processing is much more suitable and indicated to be used in machine performing such tasks. In this subsection of our paper we will introduce fuzzy logic on two ways: realizing a fuzzification of the input parameters, like in the fuzzy – MLP, or introducing instead the probabilistic similarity measure applied in the usual HMM the fuzzy similarity measure, like in the fuzzy (generalized) HMM.

Fuzzy-MLP

Introducing a fuzzy processing of the input features of the MLP is a solution to improve the MLP performances [16].

First, the input values are described through a combination of 3 membership values in the linguistic property sets: low, medium and high. For doing this the π membership function is used:

$$\pi(r,c,\lambda) = \begin{cases} 2(1- \| r-c \| / \lambda)^2 & \mathit{for}\ 0 \leq \| r-c \| \leq \lambda/2 \\ 1-2(\| r-c \| / \lambda)^2 & \mathit{for}\ 0 \leq \| r-c \| \leq \lambda/2 \\ 0 & \mathit{otherwise} \end{cases} \tag{3}$$

where: λ>0 is the radius of the π function with c as the central point, || . || denotes the Euclidian norm.

For each component F_{ji} of the input vector F_j the parameters of the π membership function for each linguistic property: low (l), medium (m) and high (h) are computed using the relations:

$$\begin{gathered} \lambda_{m(Fij)} = (F_{ji\max} - F_{ji\min})/2 \\ c_{m(Fij)} = F_{ji\min} + \lambda_{m(Fij)} \\ \lambda_{1(Fji)}=(c_{m(Fji)}-F_{ji\min})/f_{dn} \\ c_{1(Fji)} = c_{m(Fji)} - 0.5\lambda_{1(Fji)} \\ \lambda_{h(Fji)} = (F_{ji\max} - c_{m(Fji)})/f_{dn} \\ c_{h(Fji)} = c_{m(Fji)} + 0.5\lambda_{h(Fji)} \end{gathered} \tag{4}$$

where $F_{ji\max}$, $F_{ji\min}$ denote the upper and lower bounds of the observed range of feature and F_{ji} and f_{dd} is a parameter controlling the extent of overlapping.

After this, the structure of the fuzzy neural network, like the classical one, is composed from a hidden layer and an output layer.

The output vector is defined as the fuzzy class membership values. The membership value of the training $F_i = (F_{i1}\, F_{i2}\, ...F_{in})^t$ to class C_k is computed using:

$$\mu_k(F_i) = 1/(1 + z_{ik} / f_d))^{f_c} \tag{5}$$

where: f_d, f_c are constants controlling the amount of fuzziness in the class-membership set, z_{ik} is the weighted distance between the input vector F_i and the mean $O_k = (O_{k1}\, O_{k1}\, ...O_{k1})^t$ of the k-th class, defined as:

$$z_{ik} = \sqrt{\sum_{j=1} [(F_{ji} - O_{kj}) / v_{kj}]^2} \tag{6}$$

where: v_{kj} is the standard deviation of the j-th vectors' component from the C_k class.

In the training stage, the Back-Propagation algorithm is used to determine the weights which minimized the mean square error (*mse*) between the real output d_j and the desired one y_j :

$$mse = \sum_{\substack{j=1 \\ F \in train}}^{n} \left(\sum (d_j - y_j)^2\right) \tag{7}$$

During training, the learning rate is gradually decreased in discrete steps {1, 0.5, 0.3, 0.1, 0.05, 0.03, 0.01}, until the network converges to a minimum error solution.

Fuzzy-HMM

The generalized model $\overline{\lambda} = (\overline{A}, \overline{B}, \overline{\pi})$ can be characterized by the same parameters [7] like the classical, well known model. The major difference in the fuzzy variant, is the interpretation of the probability densities for the classical HMM, as fuzzy densities. On this way the probabilistic similarity measure applied in the classical HMM is replaced by a more suitable fuzzy similarity measure.

The succession of parameter vectors, called the observation sequence, *O*, produces the state sequence *S* of the model, and, visiting for example at the moment *t+1* the state $q_{t+1} = S_j$, the symbol b_j is generated. The corresponding symbol fuzzy density $b_j(O_t)$ measures the grade of certainty of the statement that we observed O_t given that we are visiting state S_j .To perform classification tasks, the fuzzy similarity measure must be calculated. Based on the fuzzy forward and backward variables, a fuzzy Viterbi algorithm is proposed in [12] for the case of the Choquet integral with respect to a fuzzy measure and multiplication as intersection operator.

The fuzzy formulation of the forward variable *α*, bring an important relaxation in the assumption of statistical independence.

The joint measure $\overline{\alpha}_{\Omega_y}\left(\{O_1,...,O_t\} \times \{y_j\}\right)$ can be written as a combination of two measures defined on $O_1, O_2, ..., O_t$ and on the states respectively, no assumption about the decomposition of this measure being necessary, where $Y = \{y_1, y_2, ..., y_N\}$ represent the states at time *t+1* (*Ω* is the space of observation vectors).

For the standard HMM, the joint measure $P(O_1, O_2, ..., O_t, q_{t+1} = S_j)$ can be written as the product $P(O_1, O_2, ..., O_t) \cdot P(q_{t+1} = S_j)$, so that two assumptions of statistical independence must be made: the observation at time *t+1*, O_{t+1}, is independent of the previous observations $O_1, O_2, ..., O_t$ and the states at time *t+1* are independent of the same observations, $O_1, O_2, ..., O_t$.

These conditions find a poor match in case of speech signals and therefore we hope in improvements due to the relaxation permitted by the fuzzy measure.

Training of the generalized model can be performed with the re-estimation formulas also done in [12] for the Choquet integral. For each model we have trained with the reestimation formulas the corresponding generalized models, GHMMs, with 3-5 states, analog to the classical case.

After the training, we have calculated the fuzzy measure $\overline{P}\left(O / \overline{\lambda}\right)$, with the fuzzy Viterby algorithm and made the decisions for recognition in the same manner like for the classical HMM: the correct decision corresponds to the model for which the calculated measure has a maximum.

3 Experimental Results

The experiment results are made by ASRS_RL, with multiple options for a large variety of speech recognition experiments [1].

In the next two sub-sections are applied the learning strategies presented in Section 2 for vowel and digit recognition and the obtained performance is evaluated.

The used databases (for vowel and digit) are sampled by 16 kHz, quantified with 16 bits, and recorded in a laboratory environment.

Vowel Recognition

The learning strategies applied in our recognition experiments are: the Kohonen maps, the MLP, the fuzzy-MLP and the fuzzy-HMM.

In the First Experiment, the vowel recognition rate using the VDRL database (Vowel Database for Romanian Language) and the MFCC coefficients (in form of 12 mel-frequency cepstral coefficients) was determined; the results obtained with MLP and HMM as learning strategies are comparatively presented in Table 1.

Table 1. Vowel recognition rate in the case of training with MS and testing with MS and FS.

Vowel	MLP		HMM	
	MS	FS	MS	FS
a	100.00%	80.71%	100.00%	82.28%
e	85.81%	43.25%	94.82%	50.67%
i	85.71%	85.71%	95.15%	92.41%
o	90.90%	51.33%	97.00%	52.78%
u	88.88%	71.42%	94.83%	77.85%
Mean	*91.26%*	*66.48%*	*96.36%*	*71.20%*

The database VDRL contained speech data from 19 speakers (9 males and 10 females) each reading the same 5 vowels (*a, e, i, o, u*). The database was organized as follows: one database for male speakers (MS), one database for female speakers (FS). In booth cases one male speaker (MS) and one female speaker (FS) was excluded from the training database and used their data for the testing.

The experimented MLP is a two-layer perceptron trained with Back-Propagation algorithm, having in the output layer 5 nodes corresponding to the 5 vowels to be classified and 100 nodes in the hidden layer (experimentally chosen).

The number of the input nodes is equal to the number of features (12 MFCC).

The HMMs chosen for comparison are Bakis (or left-right) structures with five states and for each vowel one model is created [15].

In Table 1, are displayed only the results in the case of training MS and testing with MS and FS. Similarly results were obtained for the training with FS [2].

In the Second Experiment, the vowels were described by three formant frequencies and the error rates obtained with different learning strategies are given in Table 2.

The database for formants contains 500 formant vectors, 100 for each vowel for the training and 250 formant vectors, 50 for each vowel for the testing.

The learning structures applied [5], [14] for these investigations are:

(1) KM with the input layer with 3 neurons, corresponding to the three formant frequencies and three variants for the output layer: unidimensional with 25 neurons, bidimensional with 5×5 neurons, and toroidal with 25 neurons.

(2) MLP with 3 layers organized as it follows: (a) the input layer with 3 neurons, corresponding to the three formant frequencies; (b) the hidden layer with 0 (Boolean network) or 4 neurons, (c) the output layer with 5 neurons corresponding each to a processed class (in our case the vowels *a, e, i, o, u*).

(3) Fuzzy-MLP.

Table 2. Error rates (%) in for different learning strategies for the case of format.

Vowel	KM 1-dim	KM 2-dim	KM toroidal	Boolean	MLP	Fuzzy MLP
a	2.60%	1.20%	2.00%	4.00%	0.00%	1.50%
e	3.20%	2.40%	2.40%	6.00%	2.50%	1.00%
i	2.20%	1.60%	1.20%	4.00%	1.00%	0.50%
o	1.80%	1.20%	1.20%	10.00%	1.00%	1.00%
u	2.20%	2.10%	1.80%	10.00%	1.00%	0.00%
Mean	*2.40%*	*1.70%*	*1.72%*	*6.80%*	*1.10%*	*0.80%*

In the Third Experiment the parameterization is realized with the mel-cepstral coefficients and the first and second order differences of these coefficients deduced from homomorfic filtering. The error rates obtained in the vowel recognition tests are given in Table 3, comparatively for the generalized HMM (fuzzy-HMM) and the classical HMM.

Table 3. Error rates (%) for generalized and for classical HMMs.

Vowel	Fuzzy - HMM	Classical HMM
a	5.10%	6.90%
e	2.40%	4.80%
i	3.80%	7.30%
o	2.50%	5.90%
u	0.70%	3.90%
Mean	*2.90%*	*5.78%*

Digit Recognition

In the second sub-section the performances obtained in digit recognition are evaluate with the hybrid strategies (HMM– MLP) for unenrolled and enrolled speaker.

The DDRL database (Digit Database for Romanian Language) contained speech data from 9 speakers (6 males and 3 females) each speakers reading 9 digits (*unu, doi, trei, patru, cinci, şase, şapte, opt, nouă*). We excluded two MS and one FS from the database and used them for the testing.

The digit parameters were extracted by cepstral analysis, in form of 12 mel-frequency cepstral coefficients (MFCC). The hybrid system (HMM-MLP) consists of 9 hybrid models corresponding to 9 digits. Each hybrid model is made of 5 states, each state being associated with one output node of the MLP. The MLP has one hidden layer (100 nodes experimentally chossen), and the input layer consisting of 12 nodes.

We compare the two kinds of tests (using DDRL database): first, with enrolled speakers, which mean that the speakers were involved both in training and testing, and second, with unenrolled speakers were the testing speakers are not involved in the training.

The results obtained for hybrid strategies are compared with other learning strategies (hidden Markov models, Support Vector Machine) and the performance being appreciated by their recognition rate and by their generalization capacity [5]. The word recognition rate (WRR) is reported in Table 4.

Table 4. The WRR (%) for different learning strategies.

Learning strategies	Enrolled speakers	Unrolled speakers
HMM-MLP	98.50%	98.30%
HMM	98.00%	97.50%
SVM	97.70%	91.70%

4 Conclusions

This paper reports a study focussed on the learning strategies applied in speech recognition for vowel and digit recognition.

(1) For vowel recognition in Romanian language the following conclusion can be reported:

a) The recognition rates in the case of MLP are higher than in the case of HMM. A possible explanation can be the fact that the model training is discriminative, while in the case of HMM the training is not discriminative, which represents a disadvantage of HMM utilization.

b) The KM 2-dimensional structures and the toroidal have the same performance, weaker is the performance of the 1-dimensional structure. The best balanced situation corresponds to the 2-dimensional map 5x5, in which all neurons are associated to a vowel to be recognized.

The performance obtained in the case of the Boolean network is unacceptable, but the MLP acts well.

Using fuzzy-MLP structure it is an improvement with a mean value of 0.30% comparative with the non-fuzzy structure.

c) A mean decreasing of nearly 3% is realized in the error rate by adopting the fuzzy-HMM instead of the probabilistic one.

(2) For digit recognition in Romanian language using our database (DDRL) the following observation can be reported:

a) Chosen this approach, which combine the HMM with MLP into a hybrid system is a very goad solution because the results are higher the results obtained for HMM and SVM.
b) It is to seen that SVM performances are slightly after that of the HMM, but is really promising, taking into account that the HMM has the benefit of a so long refinement time.

References

1. Dumitru, C.O.: Modele neurale si statistice pentru recunoasterea vorbirii. Ph.D. thesis, Bucharest (2006).
2. Dumitru, C.O., Gavat, I.: Vowel, Digit and Continuous Speech Recognition Based on Statistical, Neural and Hybrid Modelling by Using ASRS_RL. EUROCON 2007, Warsaw, Poland (2007), 856-863.
3. Gavat, I., & all: Elemente de sinteza si recunoasterea vorbirii. Ed. Printech, Bucharest (2000).
4. Gavat, I., Dumitru, C.O., Costache, G.: Speech Signal Variance Reduction by Means of Learning Systems. MEDINF 2003, Craiova-Romania (2003), 68-69.
5. Gavat, I, Dumitru, O., Iancu, C., Costache, G.: Learning Strategies in Speech Recognition. The 47th International Symposium ELMAR 2005, Zadar-Croatia (2005), 237-240.
6. Goronzy, S.: Robust Adaptation to Non-Native Accents in Automatic Speech Recognition. Springer - Verlag Berlin Heidelberg, Germany (2002).
7. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing – A Guide to Theory, Algorithm, and System Development. Prentice Hall (2001).
8. Juanhg, B.H., Furui, S.: Automatic Recognition and Understanding of Spoken Language–A First Step Toward Natural Human–Machine Communication. Proc. IEEE, Vol.88, No.8, (2000), 1142-1165.
9. Kohonen T.: Adaptive, Associative and Self-Organizing Function in Neural Computing. Artificial Neural Networks, IEEE Press, Piscataway- NJ (1992), 42-51.
10. Lippmann, R. and Singer, E.: Hybrid neural network/HMM approaches to word spotting. Proc. ICASSP '93. Minneapolis (1993), 565-568.
11. Lippmann, R.P.: Human and Machine Performance in Speech Recognition Tasks. Speech Communications, Vol.22, No.1 (1997), 1-15.
12. Mahomed, M. and Gader, P.: Generalized hidden Markov models. IEEE Transactions on Fuzzy Systems. (2000), 67-93.
13. Morgan, D.P., Scotfield, C.L.: Neural Networks. Prentice Hall, New York (1992).
14. Valsan, Z., Gavat, I., Sabac, B., Cula, O., Grigore, O., Militaru, D., Dumitru, C.O.: Statistical and Hybrid Methods for Speech Recognition in Romanian. International Journal of Speech Technology, Vol.5, No.3 (2002), 259-268.
15. Young, S.J.: The general use of tying in phoneme-based HMM speech recognizers. Proc. ICASSP'92, Vol.1, San Francisco (1992), 569-572.
16. Wang, Z. and Klirr, G.: Fuzzy Measure Theory. New York: Plenum (1992).