Holonic Multi-agent System Model for Fuzzy Automatic Speech / Speaker Recognition

J.J. Valencia-Jiménez and Antonio Fernández-Caballero

Departmento de Sistemas Informáticos, Universidad de Castilla-La Mancha
Escuela Politécnica Superior de Albacete, 02071-Albacete, Spain
caballer@dsi.uclm.es

Abstract. An automatic speech / speaker recognition (ASSR) system has to adapt to possible changes of speaker and environment conditions, and act as close as possible to the way a human recognizes speeches / speakers. This kind of very complex system has to deal with speech signals, looking for the integration of different information sources; and this is precisely the reason to use fuzzy logic. The main objective of this paper is the description of a robust, intelligent and adaptive system, modeled as a multi-agent system (MAS), forming a recursive hierarchy of MAS denominated holonic MAS.

Keywords: Holony, Multi-agent systems, Fuzziness, Speech recognition, Speaker recognition.

1 Introduction

Automatic speaker and/or speech recognition (ASSR) is a challenge per excellence for the Turing Test [17]. No computational efforts, throw as fast and exact results as a person does. The essential question is the integration of information that comes from diverse sources of knowledge (acoustic, phonetic, phonologic, lexical, syntactic, semantic and pragmatic). These sources incorporate great doses of uncertainty and errors due to the noise corruption of the input data. Also it is negatively influenced by the inherently ambiguous nature of natural language in the different types of knowledge that are integrated.

Therefore, the option to use fuzzy logic is very attractive [19][20]. The proposal is to quantify with words, more diffuse in its meaning, instead of with numbers. The process of symbolic inference proper of fuzzy logic rests on the denominated "decision trees" structures. Their nodes are linguistic variables that represent intermediate concepts; the connections between nodes represent the set of rules that relate the connected concepts. The membership functions express the probability that a value belongs to one or several linguistic variables. Recent is also the paradigm that possibly more affects the relations between vision and auditory decoding. We are introducing the Fuzzy Logical Model of Perception, developed throughout the last years by Massaro [10]. The fact that this paradigm is applied to decoding of the sensorial information presented to a person, leads to the application of its principles to the processing of the speech

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 73–82, 2008.

signal with the knowledge sources previously mentioned. Another paradigm with ample development in the field of Artificial Intelligence is the use of neuronal networks. In fact, neuronal networks are so intimately bound to fuzzy logic. For a long time architectures and hybrid systems denominated *neurofuzzy* have been developed, taking advantage of the properties and flexibility of both paradigms.

The main objective of our proposal is the design of a speech recognition system independent of the speaker, and the identification of the speaker independently of the speech. The system has to be robust and intelligent, and able to adapt to the environmental circumstances of noise and the characteristics of the speaker. The system also has to deal with the difficult analysis of the pronunciation of the speaker. For it, due to its versatility, the option of using the holonic multi-agent systems (MAS) [18] paradigm is considered. The remainder of this paper introduces the holonic multi-agent architecture, its bases and what are the more interesting reasons to use it, as well as a preliminary design of the fuzzy recognition system, using Prometheus [12] MAS design methodology.

2 Holonic Multi-agent Architecture

A holonic architecture is based on the model of distributed systems architectures. It is a solution based on the theory of complex adaptable systems, and the name comes from the combination of the Greek word " holos" (everything) with the suffix "on", that conceives the idea of particle or part of something. What impelled Koestler [8] Koestler to propose the concept of holon were mainly two observations: (1) First, it is easier to construct a complex system when it is composed of intermediate elements - or subsystems - that are stable. Also, complex systems, like biological organisms or societies, always are structured as a stable hierarchy of subsystems in multiple levels. These, as well, are recursively divided in subsystems of an inferior order. Structurally it is not a simple aggregation of elementary parts, and functionally it is not a global behavior like a superposition of behaviors of elementary units. Therefore a real "synergy" exists. (2) Second, when hierarchies are analyzed, it is discovered that, although it is easy to identify "sub-all" and "parts" of the "all", the "all" and the "parts" do not exist in an absolute sense, because there is a double nature of "sub-all/part".

Holonic systems are modeled in terms of components (holons) that possess their own identity and at the same time belong of a greater set. This superior set is known as holarchy [13]. Holons are self-contained with respect to their subordinated parts and are simultaneously dependent parts when they are observed from superior hierarchic levels. Therefore, the word holarchy denotes hierarchic organizations of holons with a recursive structure. Holarchy guarantees stability, predictability and global optimization of the hierarchic control. Simultaneously it provides flexibility and adaptability [3][4][5][7], since, in an independent way, each holon is able to adapt to events and to cooperate with others holons. This could directly be compared to a distributed system with cooperative nodes. Nevertheless, the key characteristic introduced is that holons cooperate in dynamic hierarchies. They reorganize every certain time or when there have been

significant changes to reach a global objective. This description of adaptable systems agrees with the one of the MAS. Indeed, a holon could be an agent, a group of them or a complete independent multi-agent subsystem. Therefore, it may be stated that the holon introduces the concept of "recursive agent". And holarchy could consider the hierarchy of agents in that system for a given moment. Therefore, given the interchangeable characteristics of holarchy and multi-agent architecture, it seems reasonable to consider the design of a complex speech/speaker recognition system as the design of a holonic MAS, with all the advantages in performance, efficiency and scalability that it entails.

3 Fuzzy Speech / Speaker Recognition

Fuzzy logic applied to speech recognition or to speaker identification offers a great advantage. It may be included in any of the main approaches to the problem - phonetic-acoustic models and pattern matching - in almost any level of the different processing steps [1]. Fuzzy logic offers the possibilities of assimilating or of surpassing uncertainty, and in a similar way humans do, of offering different results, according to the knowledge level (phonetic, lexical, syntactic, semantic), without blocking the recognition process. A key question is the representation of the spectrogram of sounds, alophones, produced in speech. The fuzzification process consists of the transformation of the analogical signal spectrum into a fuzzy description.

The multidimensional representation influences directly in the amount of fuzzy variables that characterize the different properties of the sounds to recognize - the phonemes. Basing on the International Phonetic Alphabet (IPA) chart [6], the different features or dimensions for a sound can be inferred. A sound in this alphabet is fitted and its phoneme is determined. Also, by means of a combination of phonemes, the diphons (the union of two phonemes including the transition information from one to another), or the syllables (the set of two or the more phonemes around a phoneme representing a vowel, which is the common structure most used by the humans when processing spoken language) can be obtained. In these phonetic combinations it is necessary to consider the borders and overlapping of the phonetic units and the prosody. The previous and posterior phonemes give rise to co-articulation phenomena (energy and frequencies are transferred to the adjacent units). This causes an enormous computational explosion.

Of course, it is necessary to consider that there are numerous cases, in very homophonous languages, in which several different phonemes share sound features. In fuzzy set theory this is translated in that the intersection of the sets that identify their features is not the empty set. For almost any given phoneme the values in its different features are not exclusive, but for some central values of that phoneme's dimension some peripheral values exist. Also the opposite case can occur. Several phonemes of the international alphabet really correspond with the same one in a certain language. Therefore, a word could have several phonetic transcriptions in IPA, and, detecting any of them indifferently would be

correct in that language. This can also be established by means of clustering algorithms to determine optimal patterns for each phonetic unit of the vocabulary and several values as optimal segmentation thresholds.

The process of speech and speaker recognition is made up of a series of not totally sequential stages [14]. They are rather functional phases, corresponding to phonetic, lexical, syntactic and semantic processing. In an initial stage of preprocessing or digitalization a pre-emphasis and filtering of the input sound signal is performed by means of a set of filters. Its fundamental task is to heighten the signal in a non-homogenous way, weighing and better discriminating in those frequency bands of the auditory spectrum that contribute more information to the recognition. Also the effects of atmospheric noise are diminished, increasing this way the signal to noise ratio (SNR). The result of this stage is a non-homogenous spectrogram, with the different values of intensity or sound amplitude for the different frequencies, throughout the time the pronunciation lasts. Next, the phases of frame blocking are carried out, where the signal is divided into frames of N samples, and windowing, where the discontinuities of the signal in the beginning and at the end of each frame are minimized.

Also end point detection (EDP) of the word, the silence, is necessary [16]. More specifically, the separation between the phonetic units has to be detected for a later segmentation and classification. Generally, the beginning and the end of a word is detected in speech, processing the samples of the already filtered input wave and compressing the useful information of the used phonetic unit. Usually this is based on the analysis of the linear prediction of the mean square error. The output is a vector of the samples of the phonetic unit. It is a difficult process, since in natural language pronunciation usually there are no pauses between these phonetic units. Much more if the units are smaller than words. This characteristic of the fluid speech, where the sound of a phoneme is influenced by the adjacent phonemes (mainly by the previous one), is called the "co-articulation" effect.

It is faster to perform the extraction and comparison of patterns in a fuzzy way when the numerical values have previously been transformed into fuzzy values. In some representations a progressive scale of colors is used - from black to white - that correspond to the linguistic variables that express the different intensities. Along with their corresponding membership functions, they are applied to each division of the frequencies range where the signal is sampled, providing a fuzzy value of the quantification of the signal intensity in that frequencies. In other cases quantifying linguistic variables and its corresponding membership functions are used. For instance, there is *nothing, very little, little, enough*, etc. We have also to consider that usually the duration of the phonetic unit, or phoneme, syllable or word, is identified with quantifiers such as *very short, short, medium, long* and *very long*. Therefore, the application of the "Dynamic Time Warping" algorithm [11] (also in a fuzzy way) is made in a rather economic way in terms of consumed resources and time, since much numerical precision is not necessary. A certain freedom of the values between the recognized pattern and the phonetic unit tests is allowed, fundamentally when a superior phonetic unit is constructed from more elementary ones. This occurs, for example, in the words composed

of syllables, where the different lengths of the elementary units and their viable lengths have to be combined.

The comparison among the vectors is realized by means of inference rules, where the linguistic number and value of the antecedents are based on the fuzzy measures available in the own structure of the pattern. For that reason there also partial patterns or fuzzy patterns, since according to the human mechanism of analysis and speaker/speech recognition, the different data are combined or integrated in different ways. Sometimes, they are even omitted to deal with uncertainty and the result is a series of fuzzy decisions that are weighed to give the final decision [21]. The use of fuzzy logic in the comparison phase eases to separate the classes represented by several phonetic units. This segmentation eliminates ambiguity and more efficient decisions are generated to compare pattern templates with more dimensions and features in a flexible way. In addition, to determine a sequence of phonetic units (for example, syllables) with the intention to construct a superior unit (for example, a word) dynamic programming techniques are used. The computational load is considerably increased when trying to find the optimal paths by means of backtracking. By means of fuzzification and defuzzification this can be simplified enormously, since the different calculations of the costs of dynamic programming are more complicated than the different methods of calculation of centroids of defuzzification surfaces. Thus the individual values of the simpler phonetic units can be integrated directly to form the superior unit.

A very important part of an automatic patterns recognition and identification system is the initial training algorithm. In later phases the learning algorithm, automatic or supervised, must optimize the recognition of the patterns of the speech phonetic units as well as the phonetic patterns that identify the speaker. Without fuzzy logic, the test patterns are usually constructed using a clustering algorithm. It will be adaptive in case the independence of the speaker is looked for. Test patterns or templates are obtained from a superior number of training pronunciations for each individual class of phonetic unit. Nevertheless, genetic algorithms are also used to optimize the representation of the patterns. In the case of using fuzzy logic, the templates with partial and fuzzy representation are more adaptable than a representation with numeric parameters. In addition, learning using logical fuzzy is also based on the construction and modification of the inference rules from the observed data. A very effective method is ANFIS (Adaptive Neuro-Fuzzy Inference System) [2] where different genetic algorithms and different parameters to implement different learning models are applied [15].

4 Design of the Holonic Fuzzy Recognition System

In the preliminary design of the holonic multi-agent system the Prometheus methodology has been used [12]. The main advantage in the analysis and design of holonic systems with respect to multi-agent systems is that a holon allows differentiating the roles in a more compact way and obtaining affine behaviors from the very beginning. Thus, for example, the analysis of the characteristics

of a wave can be made in the temporal or frequency domains. The temporal analysis is flexible, fast and simple, but less precise than in the frequency domain. Therefore, holons could in parallel be dedicated in temporal and frequency analysis.

As soon as the sound signal is captured by a microphone, the following main goals have to be performed:

1. To extract the main features or dimensions those identify the speaker. The aim is to construct a partial or fuzzy pattern with a minimum number of features able to identify the speaker as rapidly as possible.
2. To identify the speaker by means of a fuzzy pattern-matching process of the previous partial patterns. If the identification is not possible, it is necessary to extract all possible features of the unknown speaker with the purpose of registering him as a new one.
3. To extract the main features or dimensions of the speech. This process is optimized by the previous identification of the speaker.
4. To recognize what is spoken by means of a fuzzy pattern-matching process of the previous partial patterns. If it is not possible to determine it, then all possible features or dimensions of the unknown speech are extracted to repeat fuzzy pattern-matching with more pattern features.
5. Learning. The positive results are used to obtain the different combinations of fuzzy dimensions of the patterns by means of inverse analysis and genetic algorithms. This way speech recognition and speaker identification is improved. These results produce the modification of the inference rules base optimized to each speaker. We are looking for a most self-learning system by selecting and adapting the learning algorithms.

The main advantage of the holonic system is without a doubt the structure of holarchy. The system can dedicate holons to obtain each dimension of the phonemes. Thus, we are in front of a multi-dimensional or hyper-dimensional phonetic analysis. Later, in a successive manner, when navigating through the holarchy, in superior levels the possible combinations of the results of inferior levels are obtained. Thus the effects of co-articulation are obtained with the diphons and later with the syllables. There is really a process of construction of superior phonetic and lexical units. In the holarchy the different phonological language rules can be expressed, besides expressing the integration of the different knowledge sources of the spoken signal, phonetic, lexical, syntactic, semantic and pragmatic. In addition, in the particular case of phonetic and lexical analysis, fuzzy patterns allow the flexibility of the recognition when rising in the holarchy level. This is possible by exploring the most coherent options, or even by treating the word or lexical unit like a whole in which the humans interpolate phonetic information if the initial and final syllables are correctly pronounced. A degree of finer analysis is also allowed to distinguish between phonetically similar words, constructing directed phonetic graphs in a fuzzy way. These connect the different phonemes in a single direction. Thus the temporary sequential nature of the phonemes is represented, reflecting the uncertainty due to the different

alternatives for a pronunciation. Fuzzy values of certainty (the conditional fuzzy probability) can be introduced according to the available dictionaries or corpus.

That distributed parallelism of holarchy allows the definition and resolution of a goal of the system in a holonic way as an integration of subgoals. This way, the most complex processes, such as learning and training, can be made in a parallel and independent way. And this even at the time when recognizing words of different languages and different interlocutors, where the training and learning processes for a given language and interlocutor are optimized. This is also the case when using the IPA language between the different languages. Genetic learning and training algorithms allow "mutations" by means of flexible fuzzy rules of the representations or patterns, which portray co-articulation phenomena between phonetic neighbors - the subtle differences of pronunciation between two related phonemes that share phonetic features.

A preliminary analysis of the system with Prometheus methodology throws the hierarchy of initial goals that must be fulfilled (as shown in Fig. 1). From a first moment the hierarchy of goals could be used as bases of the holarchy, assigning a goal to each holon. When developing each goal in subgoals, when responding to the questions "how?" and "why?", holons are added recursively, forming themselves a holarchy to fulfil the subgoals. The main goals at this initial hierarchy are speaker and speech recognized (*Recognized Speaker/Speech*). In order to arrive to this state they have to fulfill each one of the following goals:

- To digitize the input signal (*Digitize Signal*). This is the pre-processing phase where the signal passes through different filters to heighten certain frequencies ranges, to eliminate noise and to decompose an analogical signal in digital values.
- To detect the main parameters of speaker and speech (*Detect Speech/Speaker Main Parameters*). Here the most decisive parameters for the recognition are calculated.
- To construct the fuzzy or partial pattern for speaker and speech (*Build Speaker/Speech Main Fuzzy Pattern*). The results of the previous goal are fuzzified and the fuzzy feature vector is constructed.

The previous goals form the initial phase of the recognition, the digitalization and the construction of the pattern. The next primary objectives are *Recognize Speaker Pattern)* and *Recognize Speech Pattern*. Both recognition goals must fulfill a sub-hierarchy of goals: *Recognize Word Pattern, Recognize Syllable Pattern, Recognize Diphon Pattern* and *Recognize Phoneme Pattern*.

In this hierarchy a superior goal is made up of several subgoals of the inferior level. There is the composition from an inferior phonetic unit to a superior one - or to a lexical unit, a word with meaning. All these goals, each one at its level and with its patterns, are related to the goal in which the comparison of the fuzzy patterns is made (*Fuzzy Pattern Matching*). For the sake of legibility of the diagram it has been related to both main goals (*Recognized Speaker/Speech*) and to the superior goal of the sub-hierarchy of word recognition (*Recognize Word Pattern*). The comparison of speech and speaker patterns is assumed to be similar. The difference is in the inference rules base that determines the

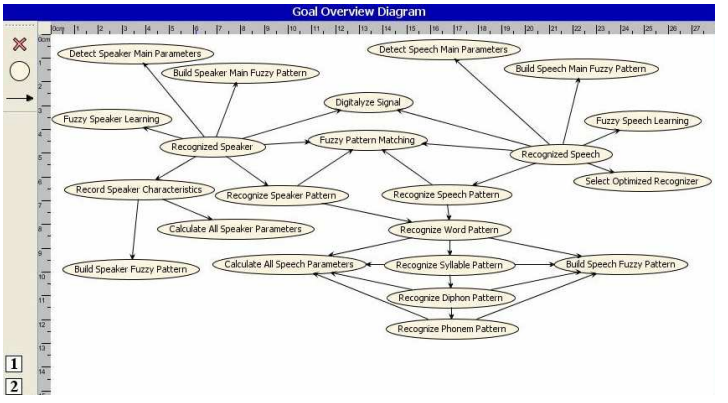

Fig. 1. Initial goals hierarchy of the holonic fuzzy recognition system

values of the features, the combination of the antecedents and the data base of the reference patterns.

In case the speaker has still not been identified, the goal to register the features of the speaker (Record Speaker Characteristics) has to be fulfilled. For it, all that speaker's possible parameters have to be calculated (*Calculate All Speaker Parameters*) and their fuzzy patterns (*Build Speaker Fuzzy Pattern*) must be constructed. In a similar form, when some part of the speech has not been recognized, then all the possible features (*Calculate All Speech Parameters*) have to be calculated and the fuzzy patterns (*Build Speech Fuzzy Pattern*) have to be construct at any phonetic unit level. Finally, the goals of fuzzy learning for speech and speaker (*Fuzzy Speaker/Speech Learning*) optimize the inference rules bases to fulfill the goal to select an optimal recognizer of the speech based on the speaker (*Select Optimized Recognizer*).

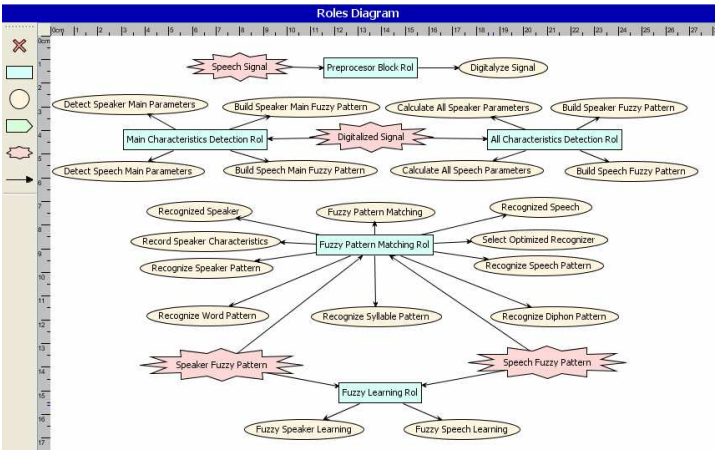

Fig. 2. Roles or functionalities diagram of the holonic fuzzy recognition system

Another interesting diagram is the roles or functionalities diagram (Fig. 2). The roles receive input information or percepts and give rise to actions, fulfilling the system goals. There are four basic percepts: (1) the input signal (*Speech Signal*), (2) the (*Digitized Signal*), (3) the fuzzy pattern of the speaker (*Speaker Fuzzy Pattern*), and, (4) the fuzzy pattern of the speech (*Speech Fuzzy Pattern*). The percepts relate the different basic roles to each other. Role pre-processor (*Preprocessor Block Role*) provides as result the digitized signal that relates the two functionalities in charge of computing the parameters or features (*Main Characteristics Detection Role, All Characteristics Detection Role*) of the fuzzy patterns of speaker and speech. The percepts are used by the role of comparing fuzzy patterns (*Fuzzy Pattern Matching Role*) and fuzzy learning (*Fuzzy Learning Role*).

5 Conclusions and Future Work

The interest to use holonic multi-agent systems comes from the desire to integrate, in an intelligent manner, systems already implemented that provide a very useful function in a given field. The paradigm of the holonic multi-agent systems can extend the use of smaller or specific holonic systems so that they take part of a constellation of stable systems that fulfill a hierarchy of goals. We considered that the paradigm of holonic multi-agent architecture is a robust, independent and scalable way to approach the construction of hierarchies of stable systems. This is the case of speech recognition and speaker identification systems, which must adapt and even learn to identify the speaker and recognize its speech, forming a dynamic hierarchic structure, oriented to the fulfilment of the goals.

The proposed system is part of an upper system of integral surveillance for facilities [9][18][13], able to integrate information from many heterogeneous platforms with multiple sensor types. In this ongoing research project, the microphones are used to extract sound information, which may be noise or human speech. The noise can be analyzed and useful information can be extracted about what occurs in the environment. However, human speech supplies very useful information in a surveillance system.

Acknowledgements

This work is supported in part by the Spanish Ministerio de Educación y Ciencia TIN2004-07661-C02-02 and TIN2007-67586-C02-02 grants, and the Junta de Comunidades de Castilla-La Mancha PBI06-0099 grant.

References

1. Beritelli, F., Casale, S., Cavallaro, A.: A robust voice activity detector for wireless communications using soft computing. IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing for Wireless Communications 16(9), 1818–1829 (1998)

2. Jang, J.S.R.: ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man & Cybernetics 23, 665–685 (1993)
3. Holland, J.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975)
4. Honma, N., Abe, K., Sato, M., Takeda, H.: Adaptive evolution of holon networks by an autonomous decentralized method. Applied Mathematics and Computation 91(1), 43–61 (1998)
5. Huang, B., Gou, H., Liu, W., Xie, M.: A framework for virtual enterprise control with the holonic manufacturing paradigm. Computers in Industry 49(3), 299–310 (2002)
6. International Phonetic Association home page (2005), http://www.arts.gla.ac.uk/IPA/index.html
7. Jarvis, J., Rönnquist, R., McFarlane, D., Jain, L.: A team-based holonic approach to robotic assembly cell control. Journal of Network and Computer Applications 29(2–3), 160–176 (2005)
8. Koestler, A.: The Ghost in the Machine. Arkana Books (1971)
9. López, M.T., Fernández-Caballero, A., Fernández, M.A., Mira, J., Delgado, A.E.: Visual surveillance by dynamic visual attention method. Pattern Recognition 39(11), 2194–2211 (2006)
10. Massaro, D.: Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. The MIT Press, Cambridge (1998)
11. Myers, C.S., Rabiner, L.R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal 60(7), 1389–1409 (1981)
12. Padgham, L., Winikoff, M.: Developing Intelligent Agent Systems: A Practical Guide. Wiley, Chichester (2004)
13. Pavón, J., Gómez-Sanz, J., Fernández-Caballero, A., Valencia-Jiménez, J.J.: Development of intelligent multisensor surveillance systems with agents. Robotics and Autonomous Systems 55(12), 892–903 (2007)
14. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs (1993)
15. Russo, M.: FuGeNeSys: A genetic neural system for fuzzy modeling. IEEE Transactions on Fuzzy Systems 6(3), 373–388 (1998)
16. Tsao, C., Gray, R.M.: An endpoint detector for LPC speech using residual error look-ahead for vector quantization applications. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 97–100 (1984)
17. Turing, A.M.: Computing Machinery and Intelligence. Mind 49, 433–460 (1950)
18. Valencia-Jiménez, J.J., Fernández-Caballero, A.: Holonic multi-agent systems to integrate independent multi-sensor platforms in complex surveillance. In: Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance, p. 49 (2006)
19. Zadeh, L.: From computing with numbers, to computing with words: A new paradigm. International Journal on Applied Mathematics 12(3), 307–324 (2002)
20. Zadeh, L.: Fuzzy logic, neural networks and soft computing. Communications of the ACM 37(3), 77–84 (1994)
21. Zadeh, L.: Outline of a new approach to the analysis of complex systems and decisión processes. IEEE Transactions on Systems, Man and Cybernetics, 28–44 (1973)