Fuzzy Logic Speech/Non-speech Discrimination for Noise Robust Speech Processing

R. Culebras, J. Ramírez, J.M. Górriz, and J.C. Segura

Dept. of Signal Theory, Networking and Communications
University of Granada, Spain

Abstract. This paper shows a fuzzy logic speech/non-speech discrimination method for improving the performance of speech processing systems working in noise environments. The fuzzy system is based on a Sugeno inference engine with membership functions defined as combination of two Gaussian functions. The rule base consists of ten fuzzy if then statements defined in terms of the denoised subband signal-to-noise ratios (SNRs) and the zero crossing rates (ZCRs). Its operation is optimized by means of a hybrid training algorithm combining the least-squares method and the backpropagation gradient descent method for training membership function parameters. The experiments conducted on the Spanish SpeechDat-Car database shows that the proposed method yields clear improvements over a set of standardized VADs for discontinuous transmission (DTX) and distributed speech recognition (DSR) and also over recently published VAD methods.

1 Introduction

The deployment of new wireless speech communication services finds a serious implementation barrier in the harmful effect of the acoustic noise present in the operating environment. These systems often benefits from voice activity detectors (VADs) which are frequently used in such application scenarios for different purposes.

Detecting the presence of speech in a noisy signal is a problem affecting numerous applications including robust speech recognition [1, 2], discontinuous transmission (DTX) in voice communications [3, 4] or real-time speech transmission on the Internet [5]. The classification task is not as trivial as it appears, and most of the VAD algorithms often fail in high noise conditions. During the last decade, numerous researchers have developed different strategies for detecting speech in a noisy signal [6, 7, 8, 9] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [10, 11, 12, 13].

Since its introduction in the late sixties [14], fuzzy logic has enabled defining the behavior of many systems by means of qualitative expressions in a more natural way than mathematical equations. Thus, an effective approach for speech/non-speech discrimination in noisy environments is to use these techniques that enables describing the decision rule by means of if-then clauses which

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part I, LNCS 3991, pp. 395–402, 2006.

are selected based on the knowledge of the problem. Beritelli [15] showed a robust VAD with a pattern matching process consisting of a set of six fuzzy rules. However, no specific optimization was performed at the signal level since the system operated on feature vectors defined by the popular ITU-T G.729 speech coding standard [4]. This paper shows an effective VAD based on fuzzy logic rules for low-delay speech communications. The proposed method combines a noise robust speech processing feature extraction process together with a trained fuzzy logic pattern matching module for classification. The experiments conducted on speech databases shows that the proposed method yields improvements over different standardized VADs for DTX and distributed speech recognition (DSR) and other VAD methods.

2 Fuzzy Logic

Fuzzy logic [16] is much closer in spirit to human thinking and natural language than traditional logic systems. Basically, it provides an effective means of capturing the approximate, inexact nature of the real world. Viewed in perspective, the essential part of a fuzzy logic system is a set of linguistic rules related by the dual concepts of fuzzy implication and the compositional rule of inference.

Fuzzy logic consists of a mapping between an input space and an output space by means of a list of if-then statements called rules. These rules are useful because they refer to variables and the adjectives that describe these variables. The mapping is performed in the fuzzy inference stage, a method that interprets the values in the input vector and, based on some set of rules, assigns values to the output.

Fuzzy logic starts with the concept of a fuzzy set. A fuzzy set F defined on a discourse universe U is characterized by a membership function $\mu_F(x)$ which takes values in the interval [0, 1]. A fuzzy set is a generalization of a crisp set. A membership function provides the degree of similarity of an element in U to the fuzzy set. A fuzzy set F in U may be represented as a set of ordered pairs of a generic element x and its grade of membership function: $F = \{(x, \mu_F(x))|x \in U\}$.

The concept of linguistic variable was first proposed by Zadeh who considered them as variables whose values are not numbers but words or sentences in a natural or artificial language. A membership function $\mu_F(x)$ is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The most commonly used shapes for membership functions are triangular, trapezoidal, piecewise linear and Gaussian. Membership functions were chosen by the user arbitrarily in the past, based on the user's experience. Now, membership functions are commonly designed using optimization procedures. The number of membership functions improves the resolution at the cost of greater computational complexity. They normally overlap expressing the degree of membership of a value to different attributes.

Fuzzy sets and fuzzy operators are the subjects and verbs of fuzzy logic. Fuzzy logic rules based on if-then statements are used to formulate the conditional statements that comprise fuzzy logic. A single fuzzy if-then rule assumes the form if x is F then y is G where F and G are linguistic values defined by fuzzy

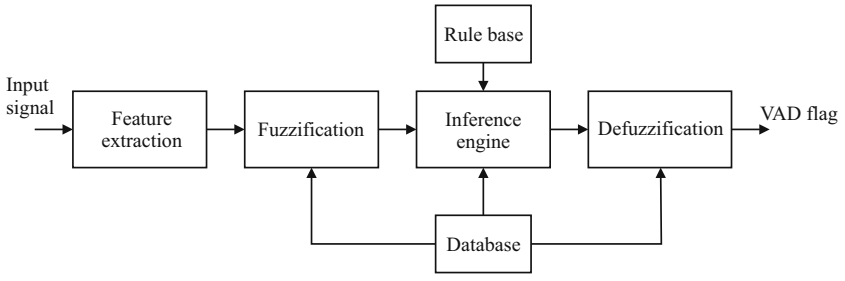


Fig. 1. Feature extraction

sets. The if-part of the rule is called the antecedent or premise, while the then-part of the rule is called the consequent or conclusion. Interpreting an if-then rule involves distinct parts: *i*) evaluating the antecedent (which involves fuzzifying the input and applying any necessary fuzzy operators), and *ii*) applying that result to the consequent.

3 Voice Activity Detection

Figure 1 shows the basic configuration of a fuzzy logic VAD which comprises five principal components: *i*) the feature extraction process that prepares discriminative speech feature for the fuzzy logic classifier, *ii*) the fuzzification interface performs a scale mapping, that transfers the range of values into the corresponding universe of discourse and performs the function of fuzzification, that converts input data into suitable linguistic variables viewed as labels of fuzzy sets, *iii*) the knowledge base comprises a knowledge of the application domain and the objective of the VAD. It consists of a database, which provides necessary definitions which are used to define linguistic VAD rules and a linguistic (fuzzy) rule base, which characterizes the VAD goal by means of a set of linguistic rules and the user experience, *iv*) the decision making logic is the kernel of the fuzzy logic VAD. It has the capability of simulating human decision making based on fuzzy concepts and of inferring actions employing fuzzy implication and the inference rules, and *v*) the defuzzification interface performs a scale mapping, which converts the range of output values into the corresponding universe of discourse, and defuzzification, which yields a nonfuzzy VAD flag.

3.1 Feature Extraction

The feature vector consists of the Zero Crossing Rates (ZCR) defined by:

$$\mathrm{ZCR} = \frac{\sum_{n=1}^{N-1} |\operatorname{sign}(x(n)) - \operatorname{sign}(x(n-1))|}{2} \tag{1}$$

and the K-band signal-to-noise ratios (SNR) that are calculated after a previous denoising process by means of a uniformly distributed filter bank on the discrete Fourier spectrum of the denoised signal.

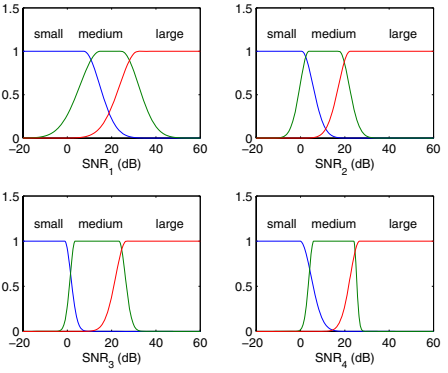


Fig. 2. Membership functions for subband SNRs

3.2 Inference Engine

A Sugeno inference engine was preferred over Mamdani's method since: *i*) it is computationally efficient, *ii*) it works well with linear techniques, *iii*) it works well with optimization and adaptive techniques, *iv*) it has guaranteed continuity of the output surface and *v*) it is well-suited to mathematical analysis. Once the inputs have been fuzzified, we know the degree to which each part of the antecedent has been satisfied for each rule. The input to the fuzzy operator is two or more membership values from fuzzified input variables. Any number of well-defined methods can fill in for the AND operation or the OR operation. We have used the product for AND, the maximum for OR and the weighted average as the defuzzification method. Finally, the output of the system is compared to a fixed threshold η. If the output is greater than η, the current frame is classified as speech (VAD flag= 1) otherwise it is classified as non-speech or silence (VAD flag= 0). We will show later that modifying η enables the selection of the VAD working point depending on the application requirements.

3.3 Membership Function Definition

The initial definition of the membership functions is based on the expert knowledge and the observation of experimental data. After the initialization, a training algorithm updates the system in order to obtain a better definition of the membership functions. Two-sided Gaussian membership functions were selected. They are defined as a combination of Gaussian functions

$$\begin{aligned} f(x;\mu_1,\sigma_1,\mu_2,\sigma_2) &= f_1(x;\mu_1,\sigma_1) f_2(x;\mu_2,\sigma_2) \\ f_i(x;\mu_i,\sigma_i) &= \begin{cases} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) & x \leq \mu_i \\ 1 & x > \mu_i \end{cases} \end{aligned} \tag{2}$$

where the first function specified by σ_1 and μ_1, determines the shape of the leftmost curve while the second function determines the shape of the rightmost

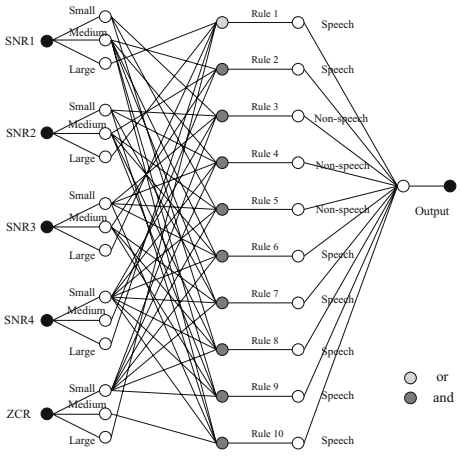


Fig. 3. Rule base

curve. Figure 2 shows the membership functions for the $K = 4$ subband SNRs that are used together with the ZCR as inputs of the fuzzy logic VAD.

3.4 Rule Base

The rule base consists of ten fuzzy rules which were trained using ANFIS [17]. It applies a combination of the least-squares method and the backpropagation gradient descent method for training membership function parameters to emulate a given training data set. An study of the better conditions for the training processed was carried out using utterances of the Spanish SpeechDat-Car (SDC) database [18]. This database contains 4914 recordings using close-talking (channel 0) and distant microphones (channel 1) from more than 160 speakers. The files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions. Four different training sets were used: *i*) quiet ch1, *ii*) low ch1 *iii*) high ch1, and *iv*) a combination of utterances from the three previous subsets. Training with data from the three categories yielded the best results in speech/pause discrimination.

Figure 4 shows the operation of the VAD on an utterance of the SpeechDat-Car database recorded with the close talking and distant microphones. It shows the effectiveness of the fuzzy logic pattern matching for discriminating between speech and non-speech even in a high noise environments.

4 Experiments

This section analyzes the proposed VAD and compares its performance to other algorithms used as a reference. The analysis is based on the ROC curves [19], a frequently used methodology to describe the VAD error rate. The Spanish

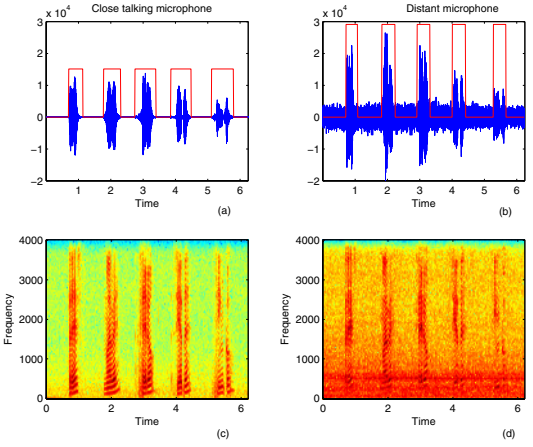
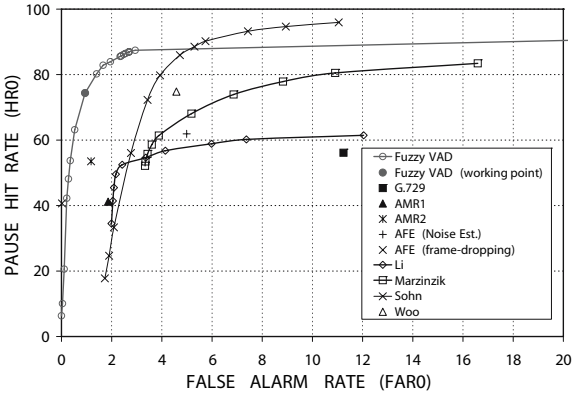


Fig. 4. Operation of the fuzzy VAD on an utterance of the Spanish SpeechDat-Car database. VAD decision for utterances recorded with close talking (a) and distant microphones (b) and associated spectrograms.

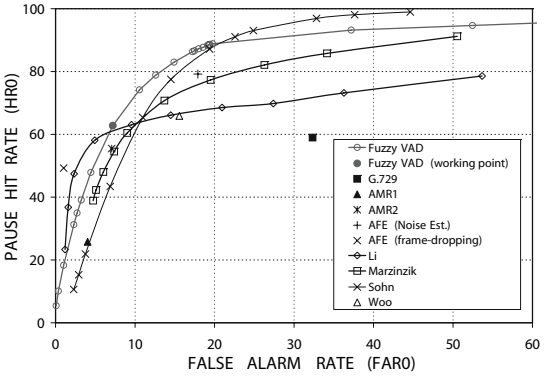
SDC database [18] was used. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined for each noisy condition being the actual speech frames and actual speech pauses determined by hand-labelling the database on the close-talking microphone. Figure 5 shows the ROC curves of the proposed VAD and other frequently referred algorithms [20, 21, 19, 6] for recordings from the distant microphone in quiet and high noisy conditions. The working points of the ITU-T G.729, ETSI AMR and ETSI AFE VADs are also included. The results show improvements in speech detection accuracy over standard VADs and a representative set of recently reported VAD algorithms [20, 21, 19, 6].

5 Conclusion

This paper showed the effectiveness of fuzzy logic concepts for robust speech/non-speech discrimination. The fuzzy system is based on a Sugeno inference engine with membership functions defined as combination of two Gaussian functions. The rule base consists of ten fuzzy if then statements defined in terms of the denoised subband signal-to-noise ratios (SNRs) and the zero crossing rates (ZCRs). Its operation is optimized by means of a hybrid training algorithm combining the least-squares method and the backpropagation gradient descent method for training membership function parameters. With these and other innovations the proposed algorithms unveils significant improvements over ITU-T G.729, ETSI AMR and ETSI AFE standards as well as over VADs that define the decision rule in terms of averaged subband speech features.



(a)



(b)

Fig. 5. Comparative results. *a*) Quiet ch1, *b*) High ch1.

Acknowledgements

This work has been funded by the European Commission (HIWIRE, IST No. 507943) and the Spanish MEC project TEC2004-03829/FEDER.

References

1. Karray, L., Martin, A.: Towards improving speech detection robustness for speech recognition in adverse environments. Speech Communitation (2003) 261–276
2. Ramírez, J., Segura, J.C., Benítez, M.C., de la Torre, A., Rubio, A.: A new adaptive long-term spectral estimation voice activity detector. In: Proc. of EUROSPEECH 2003, Geneva, Switzerland (2003) 3041–3044
3. ETSI: Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels. ETSI EN 301 708 Recommendation (1999)

4. ITU: A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70. ITU-T Recommendation G.729-Annex B (1996)
5. Sangwan, A., Chiranth, M.C., Jamadagni, H.S., Sah, R., Prasad, R.V., Gaurav, V.: VAD techniques for real-time speech transmission on the Internet. In: IEEE International Conference on High-Speed Networks and Multimedia Communications. (2002) 46–50
6. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Processing Letters **16** (1999) 1–3
7. Cho, Y.D., Kondoz, A.: Analysis and improvement of a statistical model-based voice activity detector. IEEE Signal Processing Letters **8** (2001) 276–278
8. Gazor, S., Zhang, W.: A soft voice activity detector based on a Laplacian-Gaussian model. IEEE Transactions on Speech and Audio Processing **11** (2003) 498–505
9. Armani, L., Matassoni, M., Omologo, M., Svaizer, P.: Use of a CSP-based voice activity detector for distant-talking ASR. In: Proc. of EUROSPEECH 2003, Geneva, Switzerland (2003) 501–504
10. Bouquin-Jeannes, R.L., Faucon, G.: Study of a voice activity detector and its influence on a noise reduction system. Speech Communication **16** (1995) 245–254
11. Ramírez, J., Segura, J.C., Benítez, C., de la Torre, A., Rubio, A.: An effective subband osf-based vad with noise reduction for robust speech recognition. IEEE Trans. on Speech and Audio Processing **13** (2005) 1119–1129
12. Ramírez, J., Segura, J.C., Benítez, C., García, L., Rubio, A.: Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Processing Letters **12** (2005) 689–692
13. Górriz, J., Ramírez, J., Segura, J., Puntonet, C.: Improved MO-LRT VAD based on bispectra gaussian model. Electronics Letters **41** (2005) 877–879
14. Zadeh, L.A.: Fuzzy algorithm. Information and Control **12** (1968) 94–102
15. Beritelli, F., Casale, S., Cavallaro, A.: A robust voice activity detector for wireless communications using soft computing. IEEE Journal of Selected Areas in Communications **16** (1998) 1818–1829
16. Mendel, J.: Fuzzy logic systems for engineering: A tutorial. Proceedings of the IEEE **83** (1995) 345–377
17. Jang, J.S.R.: ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man, and Cybernetics **23** (1993) 665–685
18. Moreno, A., Borge, L., Christoph, D., Gael, R., Khalid, C., Stephan, E., Jeffrey, A.: SpeechDat-Car: A Large Speech Database for Automotive Environments. In: Proceedings of the II LREC Conference. (2000)
19. Marzinzik, M., Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Transactions on Speech and Audio Processing **10** (2002) 341–351
20. Woo, K., Yang, T., Park, K., Lee, C.: Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters **36** (2000) 180–181
21. Li, Q., Zheng, J., Tsai, A., Zhou, Q.: Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing **10** (2002) 146–157