EFFECTIVE SPEECH/PAUSE DISCRIMINATION COMBINING NOISE SUPPRESSION AND FUZZY LOGIC RULES

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

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

ABSTRACT

This paper shows an effective speech/pause discrimination method combining spectral noise filtering and fuzzy logic rules. The fuzzy system is based on a Sugeno inference engine with membership functions defined as combination of two Gaussian functions. 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 fuzzy classifier consists of ten fuzzy rules defined in terms of the denoised subband signal-to-noise ratios (SNRs) and the zero crossing rate (ZCRs). An exhaustive analysis conducted on the Spanish SpeechDat-Car databases is conducted in order to assess the performance of the proposed method and to compare it to existing standard VAD methods. The results show improvements in detection accuracy over standard VADs and a representative set of recently reported VAD algorithms.

1. INTRODUCTION

The emerging wireless communication systems are demanding increasing levels of performance and speech processing systems working in noise adverse environments. These systems often benefits from using voice activity detectors (VADs) which are frequently used in such application scenarios for different purposes. Speech/non-speech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition [1, 2], discontinuous transmission [3, 4], real-time speech transmission on the Internet [5] or combined noise reduction and echo cancellation schemes in the context of telephony [6]. The speech/non-speech classification task is not as trivial as it appears, and most of the VAD algorithms fail when the level of background noise increases. During the last decade, numerous researchers have developed different strategies for detecting speech on a noisy signal [7, 8, 9, 10] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [11]. Most of them have focussed on the development of robust algorithms with special attention on the derivation and study of noise robust features and decision rules [12, 13, 14, 7]. The different approaches include those based on energy thresholds, pitch detection, spectrum analysis, zero-crossing rate, periodicity measure or combinations of different features.

Since its introduction in the late sixties [15], fuzzy logic marked the beginning of a new era in defining the behaviour of many systems by means of qualitative expressions in a more natural way than mathematical equations. Thus, an effective alternative to deal with the problem of voice activity detection is to use these methodologies. Beritelli [16] 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.

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

2. FUZZY LOGIC

Fuzzy logic [17] 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 those 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. The 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 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 op-

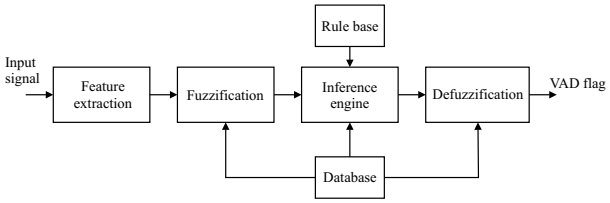


Fig. 1. Feature extraction

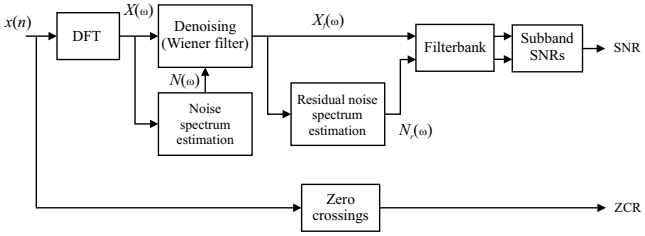


Fig. 2. Feature extraction

erators), 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 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 extraction process is shown in figure 2. The input signal $x(n)$ sampled at 8 kHz is decomposed into 25-ms overlapped frames with a 10-ms window shift. The feature vector consists of Zero Crossing Rates (ZCR) defined as:

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

and subband SNRs. For estimating the subband SNRs, the current frame consisting of *N*= 200 samples is zero padded to 256 samples and power spectral magnitude $X(\omega)$ is computed through the discrete Fourier transform (DFT). A denoising process based on a Wiener filter is applied to improve the performance of the VAD in high noise environments. Fig. 3 shows a block diagram of the denoising process. Thus, the noise spectrum is estimated during a short

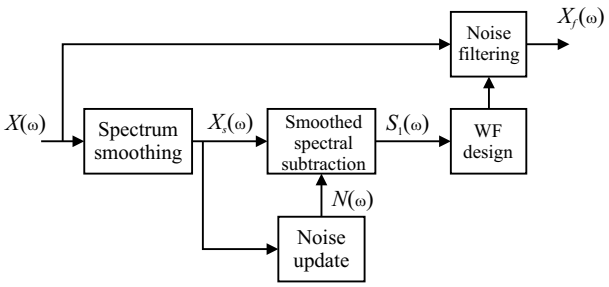


Fig. 3. Denoising stage previous to feature extraction

initialization period in order to design the optimum Wiener filter in the frequency domain. The denoising process is described as follows:

i) Spectrum smoothing. The power spectrum is averaged over two consecutive frames and two adjacent spectral bands.

ii) Noise estimation. The noise spectrum $N(\omega)$ is updated during non-speech periods by means of a 1^{st} order IIR filter on the smoothed spectrum $X_s(\omega)$, that is, $N(\omega) = \lambda N(\omega) + (1-\lambda)X_s(\omega)$ where $\lambda = 0.99$.

iii) WF design. First, the clean signal $S_1(\omega)$ is estimated by combining smoothing and spectral subtraction:

$$S_1(\omega) = \gamma X'(\omega) + (1-\gamma)\max(X_s(\omega) - N(\omega), 0) \quad (2)$$

where γ= 0.98. Then, the WF $H(\omega)$ is designed as:

$$H(\omega) = \frac{\eta(\omega)}{1+\eta(\omega)} \quad (3)$$

where:

$$\eta(\omega) = \max\left[\frac{S_1(\omega)}{N(\omega)}, \eta_{\min}\right] \quad (4)$$

and $\eta_{\min}$ is selected so that the filter H yields a 20 dB maximum attenuation. Note that, $X'(\omega) = H(\omega)X(\omega)$ is the spectrum of the cleaned speech signal, assumed to be zero at the beginning of the process and needed for designing the WF through Eqs. 2 to 4. The filter $H(\omega)$ is smoothed in order to eliminate rapid changes between neighbor frequencies that may often cause musical noise. Thus, the variance of the residual noise is reduced and consequently, the robustness when detecting non-speech is enhanced. Smoothing is performed by truncating the impulse response of the corresponding causal FIR filter to 17 taps using a Hanning window.

iv) Frequency domain filtering. The smoothed filter $H_s(\omega)$ is applied in the frequency domain to obtain the de-noised spectrum $X_f(\omega) = H_s(\omega)X(\omega)$.

Once the input signal has been denoised, the filterbank shown in figure 2 reduces the dimensionality of the feature vector to a representation including broadband spectral information suitable for detection. Thus, the signal and the residual noise is passed through a K-band filterbank which is defined by

$$E_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} X_f(\omega); \quad N_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} N_r(\omega)$$
$$\omega_k = \tfrac{\pi}{K}k \qquad k = 0, 1, ..., K-1 \quad (5)$$

and the subband SNRs are computed as

$$\mathrm{SNR}(k) = 20\log_{10}\left(\frac{E_B(k)}{N_B(k)}\right) \qquad k = 0, 1, ..., K-1 \quad (6)$$

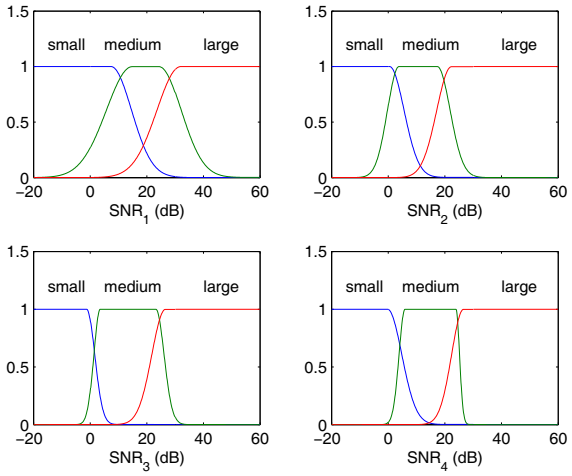


Fig. 4. Membership functions for subband SNRs

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

$$f(x;\mu_1,\sigma_1,\mu_2,\sigma_2) = f_1(x;\mu_1,\sigma_1) f_2(x;\mu_2,\sigma_2)$$
$$f_i(x;\mu_i,\sigma_i) = \begin{cases} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) & x \le \mu_i \\ 1 & x > \mu_i \end{cases} \quad (7)$$

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 curve. Figures 4 and 5 show the membership functions used for the problem addressed when four subband SNRs and the ZCR were used as inputs to the fuzzy logic VAD.

3.4. Rule base

The rule base consists of ten fuzzy rules which were trained using ANFIS [18]. 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 database [19]. 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.

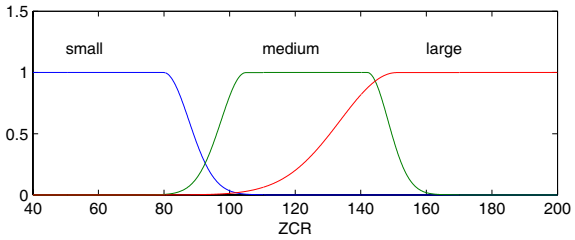


Fig. 5. Membership functions for ZCR

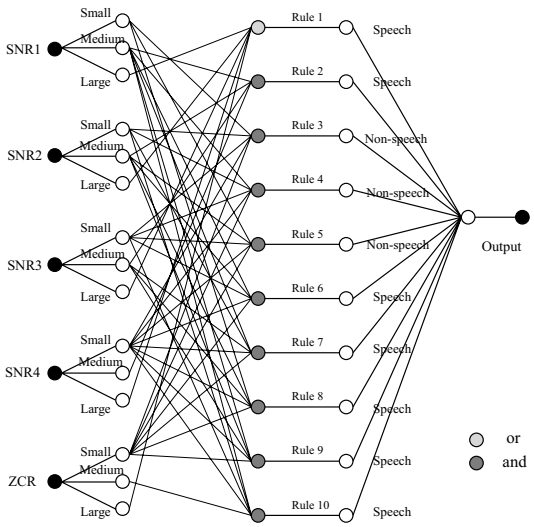
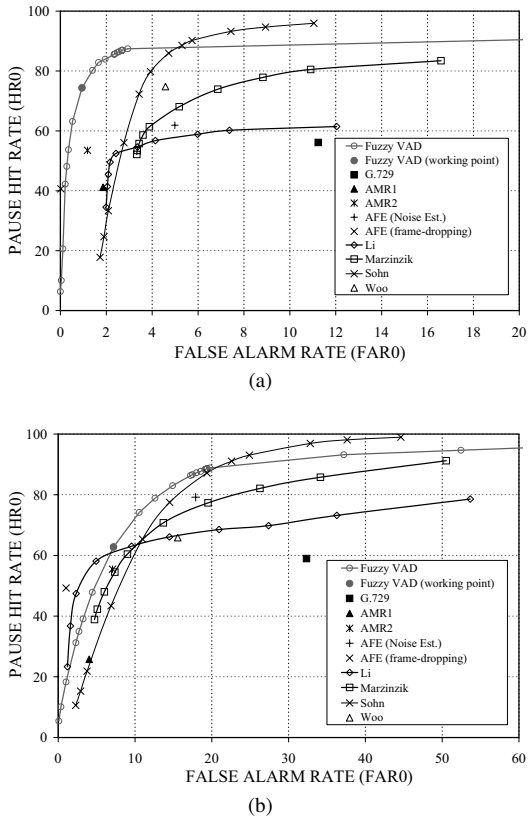


Fig. 6. Rule base

4. EXPERIMENTAL FRAMEWORK

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, a frequently used methodology to describe the VAD error rate. The Spanish SDC database [19] 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 7 shows the ROC curves of the proposed VAD and other frequently referred algorithms [12, 13, 14, 7] for recordings from the distant microphone in quiet and high noisy conditions. The working points of the ITU-T G.729, ETSI AMR and AFE VADs are also included. The results show



(a)

(b)

Fig. 7. Comparative results. *a*) Quiet ch1, *b*) High ch1.

improvements in detection accuracy over standard VADs and a representative set of recently reported VAD algorithms [12, 13, 14, 7].

5. CONCLUSIONS

This paper proposed an effective fuzzy logic voice activity detection algorithm. The VAD is based on a Sugeno inference engine with membership functions defined as combination of two Gaussian functions. 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. A comparison with the most representative standard VAD methods and recently reported algorithm was provided. The exhaustive analysis conducted on the Spanish SpeechDat-Car database showed relevant improvements when compared to G.729 and AMR VADs in speech/pause detection accuracy and other existing proposals for a representative set of noisy conditions.

6. REFERENCES

[1] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," *Speech Communitation*, , no. 3, pp. 261–276, 2003.

[2] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. of EUROSPEECH 2003*, Geneva, Switzerland, September 2003, pp. 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] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the Internet," in *IEEE International Conference on High-Speed Networks and Multimedia Communications*, 2002, pp. 46–50.

[6] F. Basbug, K. Swaminathan, and S. Nandkumar, "Noise reduction and echo cancellation front-end for speech codecs," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 1, pp. 1–13, 2003.

[7] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3, 1999.

[8] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278, 2001.

[9] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, 2003.

[10] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. of EUROSPEECH 2003*, Geneva, Switzerland, September 2003, pp. 501–504.

[11] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Communication*, vol. 16, pp. 245–254, 1995.

[12] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[13] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[14] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[15] L. A. Zadeh, "Fuzzy algorithm," *Information and Control*, vol. 12, pp. 94–102, 1968.

[16] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal of Selected Areas in Communications*, vol. 16, no. 9, pp. 1818–1829, 1998.

[17] J.M. Mendel, "Fuzzy logic systems for engineering: A tutorial," *Proceedings of the IEEE*, vol. 83, no. 3, pp. 345–377, 1995.

[18] J. S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[19] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proceedings of the II LREC Conference*, 2000.