# ON IMPROVING VOICE ACTIVITY DETECTION BY FUZZY LOGIC RULES : CASE OF COHERENCE BASED FEATURES

*Sofia Ben Jebara and Taha Ben Amor*

Département de Mathématiques Appliquées, Signal et Communications,
Ecole Supérieure des Communications de Tunis
Cité Technologique des Communications de Tunis,
Route de Raoued 3.5 Km, Cité El Ghazala, 2083, Ariana, Tunisia
phone: +216 71 857 000, fax: +216 71 856 829, email: sofia.benjebara@supcom.rnu.tn

## ABSTRACT

In this paper, we investigate the use of fuzzy logic for Voice Activity Detection (VAD). The feature extraction part is based on coherence measure between the noisy speech and its prediction residue. The decision part uses fuzzy logic rules instead of classical thresholding tools. Different fuzzy logic models are developed in order to track noise characteristics. The performances of the algorithm are compared to that of ITU-T G.729B VAD and UMTS 3G TS 26.094 VAD in various conditions. The results show that the proposed algorithm has globally better performances than G.729B and presents moderate improvement when compared to UMTS 3G TS 26.094 VAD.

## 1. INTRODUCTION

Voice Activity Detection (VAD) is an important part in many speech communication systems. In fact, since speech in nature contains a lot of silence, localization of non-speech activity permits to enhance systems capacity: significant reduction in the bit rate, considerable bandwidth reduction, channel sharing with other information, lower power consumption in portable equipment,...

However, VAD in noisy environments is a challenging task. For example, in wireless communications, an environmental background noise with high level and which looks like speech (restaurant, train, shopping,...) is commonly present, rendering VAD difficult and some times not reliable : if some speech frames are detected as noise, intelligibility is seriously damaged (speech clipping) while if noise is detected as speech, advantages of silence detection are loosed.

Many different algorithms, tailored to the application, have been proposed (ITU-T recommendation G.729 annex B [1] in multimedia communications, ETSI recommendations for mobile systems [2], ...). They are based on signal spectral features estimation and heuristic decision rules. While different kind of features are investigated in the literature (spectral features, high order statistics, periodicity measure,...), few works deal with decision rule improvement. Recently, an effort to optimize it by applying fuzzy logic rules has been made (see for example [3].

Recently, we have proposed a VAD, which is based on the coherence measure between the noisy signal and its prediction residue [4]. The decision rule was based on threshoding and is applied in different frequency bands, according to the speech characteristics. Its simplicity of implementation and the small number of parameters to adjust can distinguish the proposed algorithm. While the algorithm performances in term of probability of error are noticeable for white-like noises even for low SNR, it dramatically fails for colored speech-like noises [5].

Complementary to the previous work, the main contribution in this paper consists on the use fuzzy logic tool for better VAD : the hard threshoding rule is discarded and is replaced by a soft fuzzy logic decision rule. It has the advantage of allowing a gradual, continuous transition rather than a sharp change between two values.

The paper is organized as follows : in section 2, we present the feature extraction module. Section 3 is devoted to a detailed description of the use of fuzzy logic for VAD. Section 4 illustrates the fuzzy model obtained for Voice Activity Detection for different kinds of noises. Section 5 gives some experimental results, validating the proposed approach.

## 2. VAD FEATURE EXTRACTION

### 2.1 Feature calculus

A noisy speech signal $x(k)$ is composed of a clean signal $s(k)$ which should be active speech or silence and an additive noise $n(k)$. A linear prediction of the noisy speech, which approximates a sample as a linear combination of past samples, is applied. The prediction error of the speech signal $x(k)$ is given by :

$$e(k) = x(k) - P^T(k)X(k-1), \tag{1}$$

Where $P(k) = [p_1(k), p_2(k), .., p_{L_P}(k)]^T$ is the predictor, $L_P$ is the predictor taps number and $X(k-1) = [x(k-1), x(k-2), .., x(k-L_P)]^T$ is the past input vector. The prediction coefficients $\{p_i(k)\}, i = 1, .., L_P$ are calculated using either short-term Levinson-Durbin algorithm or adaptive Normalized Least Mean Square algorithm [6].

The data are segmented into frames of 16 ms duration. $(.)_m(k)$ denotes the $k^{th}$ sample relative to $m^{th}$ frame of signal $(.)$ The coherence function between noisy speech and its residue of prediction is developed in the frequency domain. The $m^{th}$ frame coherence function is defined as :

$$\mathbf{C}_{x,e}(m,f) = \frac{\mathbf{\Gamma}_{x,e}(m,f)}{\sqrt{\mathbf{\Gamma}_{x,x}(m,f)\mathbf{\Gamma}_{e,e}(m,f)}}, \tag{2}$$

where $\mathbf{\Gamma}_{x,x}(m,f)$ and $\mathbf{\Gamma}_{e,e}(m,f)$ are spectral densities of signals $x_m(k)$ and $e_m(k)$ respectively. $\mathbf{\Gamma}_{x,e}(m,f)$ is the inter-signal spectral density between $x_m(k)$ and $e_m(k)$. The whole frequency band is split into different frequency bands $B_i$ $(i = 1, .., M)$ and we calculate the average of the coherence function in each frequency band $B_i$ :

$$\mathbf{E}_m^{B_i} = \sum_{f \in B_i} |\mathbf{C}_{x,e}(m,f)|. \tag{3}$$

The whole set of $\mathbf{E}_m^{B_i}$ constitutes the set of parameters to be used for voice activity detection.

### 2.2 VAD idea

The basic idea of feature extraction consists in measuring the similarity between the noisy speech and its residual of prediction by means of coherence in frequency domain. In case of white or weakly correlated noise, the following characteristics are observed [4] :

- the coherence between speech signal (auto-regressive model) and its prediction residue is weak,
- the coherence between noise signal during silence and its prediction residue is close to one.

The VAD decision can simply be limited to thresholding process in each band. Then, a logic combination between thresholded outputs in each band is applied, leading to the final VAD decision . However, when speech is affected by some real noises (such as restaurant, talking,...), the classification by thresholding concept completely fails. This fact is due to noise characteristics (colored/white, stationary or not, speech-like or not, comfortable or not,...)

## 3. DESCRIPTION OF FUZZY LOGIC CONTROLER

Instead of a decision rule based on thresholding, we propose to use a fuzzy logic based decision. In fact, the fuzzy logic is suitable for problems requiring approximate rather than exact solutions : it has the advantage of allowing a gradual continuous transition rather than a sharp change between two values [3].

### 3.1 Fuzzy model definition

Firstly, fuzzy membership functions, defined by both a range of values and a degree of membership, are developed for the collection of input parameters $\mathbf{E}_m^{B_i}$. We choose to quantize inputs into three intervals (small, medium and large) which are described by gaussian functions characterized by their center $C_i$ and their variance $D_i$. The output of the fuzzy system is a continuous parameter, varying in a range between 0 and 1, which indicates the degree of membership in the activity class and the complement in the non-activity class. It is then quantized into two intervals (small and large).

### 3.2 Fuzzy model rules

Once membership functions have been defined for inputs and output variables, some fuzzy rules are created to define the fuzzy controller. Each rule relates fuzzy input variables to output variable by means of NOT (complement), OR (maximum) and AND (minimum) operators. Furthermore, for each decision rule, a weighting factor $w_j$ ($j = 1, N$ rules) is necessary to counterbalance the importance of such rule. These rules are obtained during a training phase where several speech sequences are marked manually as belonging to the activity and non-activity classes. These rules are memorized in the fuzzy model which will be used for final decision.

We note that all parameters are optimized using genetic algorithms because of their robustness to search for the global solution with several variables [7].

### 3.3 Fuzzy logic decision module

During VAD decision, the fuzzy logic controller evaluates inputs and applies them to the rule base. From the fuzzy output, the system must generate a usable output, which is a boolean flag indicating presence of speech or silence. We use centroid calculation defuzzification method whose output is compared to a fixed experimentally chosen threshold (by minimizing the total error during training phase).

## 4. FUZZY MODEL ILLUSTRATION

During proposed VAD conception scheme, we have considered 6 types of noise : car, talking, street, white Gaussian, restaurant, and exhibition [8]. Due to lack of space, only three kind of noises are illustrated in this section.

As features, we considered the following four frequency bands to develop the fuzzy logic model : $B_1 = [0\,, 500Hz]$ , $B_2 = [500Hz\,, 1kHz]$, $B_3 = [1Khz\,, 2kHz]$ and $B_4 = [2kHz\,, 4Khz]$.

### 4.1 Fuzzy logic model illustration

Figure 1 summarizes all the steps of fuzzy logic controller. In rows, we have rules and in columns, we have the membership functions of the system inputs and the inference rule for the system output. A weighted average for all the active rules determines the output by centroid calculation.

This figure is developed in case of talking noise. We note that only bands $B_3$ and $B_4$ are considered because of the small dynamic of variation $\mathbf{E}_m^{B_1}$ and $\mathbf{E}_m^{B_2}$ in bands $B_1$ and $B_2$ [8].

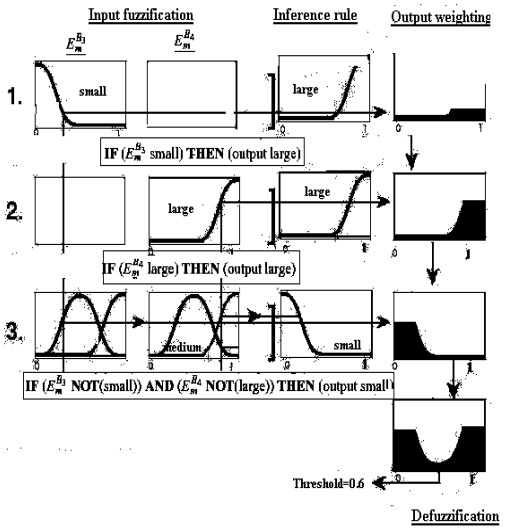


Figure 1: Decision process using fuzzy logic in case of talking noise.

### 4.2 Fuzzy logic model according to the kind of noise

Figure 2 illustrates membership functions obtained for white Gaussian (a) and restaurant (b) noises. Their relative rules are described in Table 1. These two illustrations permit the following interpretations :

- In case of white Gaussian noise, bands $B_1$ and $B_3$ have the main importance. In fact, for rule 1, any projected value in their membership function will have greater membership degree and their coherence values are linked using OR (maximum) operator. In case of rule 2, any projected

value in their membership function will have small membership degree and their coherence values are linked using AND (minimum) operator. We also note that weighting factor of rule 2 is bigger than the one of rule 1, which favors rule 2.

- In case of restaurant noise, bands $B_1$, $B_2$ and $B_3$ are used. Band $B_3$ is the most important and the first rule have priority.
- We note the few number of rules (2 or 3) and the limited number of considered bands in each case. This fact permits to reduce considerably the complexity of implementation.

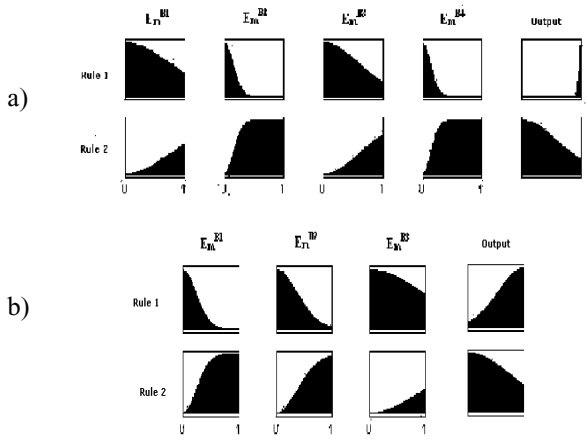


Figure 2: Membership functions of the fuzzy inputs and the fuzzy output for two noises: white Gaussian (a) and restaurant (b).

## 5. EXPERIMENTAL RESULTS

### 5.1 Speech database

We used 137 seconds of speech sampled at 8 kHz, quantized at 8 bits per sample and spoken by 8 females and 10 males. The database was subdivided into a learning database and a testing database. We decided manually on the nature of all database frames (speech activity or silence); the purpose is to determine VAD parameters during training and to evaluate algorithm performances during tests. Training sequences are recorded in different noisy environments whose *SNR* varies from 0 to 20 dB. We determine 6 classes of noise whose fuzzy logic model is optimized according to noisy data. We tested the VAD performances for 8 kind of noises : the 6 mentioned above and two others (subway and train) for which the fuzzy logic model is chosen from the 6 determined classes.

### 5.2 Performances criteria

To evaluate the effectiveness of the proposed approach, the probabilities of correct and false detection are computed. We denote:

- $P_e$ : the probability of false decision. It is calculated as the ratio of incorrectly classified frames to the total number of frames.
- $P_s$ : the probability of correct speech decision. It is calculated as the ratio of correctly classified speech frames to the total number of speech frames.
- $P_n$ : the probability of correct pause decision. It is calculated as the ratio of correctly classified pause frames to the total number of pause frames.

### 5.3 Performances evaluation

Table 2 illustrates performances of the proposed VAD in term of probability of correct and false decision. Different kinds of noise with different levels are tested (very high $SNR = 0\,\text{dB}$, high $SNR = 5\,\text{dB}$ and moderate $SNR = 10\,\text{dB}$). The proposed algorithm is compared to G.729 VAD [1] and to UMTS 3G TS 26.094 VAD for Adapative Multirate Codec (AMR) [2]. Table 2 leads to the following interpretations :

- In term of probability of error, the proposed VAD outperforms other algorithms in case of gaussian, talking, car, and subway noises. It is similar to G.729 in case of exhibition noise, slightly worse than AMR VAD in case of street noise (where G.729 fails). The proposed VAD is therefore worse in case of restaurant and train noises. This fact is due to dominant periodic components in noise (such as machine transitions in train). The limitations can be avoided by increasing the number of the classes and the number of features (more than 4 bands to track noise characteristics).
- In term of probability of speech detection, the AMR VAD outperforms in majority of cases. However, the proposed VAD is better than G.729 VAD. In fact, G.729 is conceived to work in multimedia systems where quasi-stationary noises appear whereas AMR VAD is conceived to work in mobile communications where all kind of noises are possible. The proposed VAD gives a good compromise, when dealing with communication environment.
- In term of probability of noise detection, each algorithm seems to be suitable for some kind of noises.

## 6. CONCLUSION

The purpose of this paper is to exploit fuzzy logic properties in order to improve the VAD decision. To this end, we used the coherence between the noisy signal and its prediction residue as features. It is measured in different frequency bands. Furthermore, in order to take into account noise characteristics, we developed different fuzzy models; each one is suitable for one kind of noise. The simulation results show that the proposed algorithm has globally better performance than G.729B and presents moderate improvement when compared to UMTS 3G TS 26.094 VAD. Further work for algorithm improvement includes increasing the number of noise classes and the number of considered frequency bands. This work can also be extended to refined audio classification.

## REFERENCES


[1] A. Benyassine, E. Shlomot and H. Su, "ITU-T recommendation G.729, annex B, A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64-72, September 1997.

[2] ETSI recommendations for VAD, "GSM 06.32, GSM 06.42, GSM 06.82, UMTS 3G TS 26.094."

[3] F. Beritelli, S. Casale and A. Cavallaro, "A Robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications (JSAC), Special Issue on Signal Processing for Wireless Communications*, vol. 16, no. 9, December 1998.

[4] S. Ben Jebara, "Coherence-based voice activity detector," *IEE Electronics Letters*, vol. 38, no. 22, pp. 1393-1397, October. 2002.

[5] S. Ben Jebara, "The multi-band LPC coherence for voice activity detection," *Proc. of the WSEAS Int. Conf. in Signal Processing, Robotics and Automation ISPRA*, Cadiz, Spain, June 2003.

[6] N. S. Jayant and P. Noll, "Digital coding of waveforms: principles and applications to speech and video coding", *Englewood Cliffs, NJ:Prentice-Hall*, 1984.

[7] K. S. Tang, K. F. Man, S. Kwong and Q. He, "Genetic algorithms and their applications," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 22-37, November 1996.

[8] T. Ben Amor, "Détection de l'activité vocale en environnement bruité," *B. S. Report*, SUP'COM, Tunisia, June 2003 (in French).


Table 1: Fuzzy rules for two noises : white Gaussian (a) and restaurant (b).

| Rule | Weighting | Threshold |
|---|---|---|
| **White Gaussian** | | |
| **(1) IF** ($E_m^{B_1}$ small) **OR** ($E_m^{B_2}$ small) **OR** (($E_m^{B_3}$ small) **OR** (($E_m^{B_4}$ small) **THEN** (output large) | 0.227 | 0.8 |
| **(2) IF** ($E_m^{B_1}$ **NOT**(small)) **AND** ($E_m^{B_2}$ (**NOT**(small)) **AND** (($E_m^{B_3}$ (**NOT**(small)) **AND**(($E_m^{B_4}$ **NOT**(small)) **THEN** (output small) | 0.813 | 0.8 |
| **Restaurant** | | |
| **(1) IF** ($E_m^{B_1}$ small) **OR** ($E_m^{B_2}$ small) **OR** (($E_m^{B_3}$ small) **THEN** (output large) | 0.548 | 0.6 |
| **(2) IF** ($E_m^{B_1}$ **NOT**(small)) **AND** ($E_m^{B_2}$ **NOT**(small)) ET (($E_m^{B_3}$ **NOT**(small)) **THEN** (output small) | 0.021 | 0.6 |

Table 2: Proposed VAD performances for different classes of noise and for different values of *SNR* (dB).

| | | $P_s$(%) | | | $P_n$(%) | | | $P_e$(%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *SNR* | AMR | G.729 | coh | AMR | G.729 | coh | AMR | G.729 | coh |
| Gaussian | 10 | 91 | 78 | **93** | 88 | **95** | 93 | 10 | 11 | **6** |
| | 5 | **94** | 72 | 86 | 83 | 88 | **95** | 12 | 17 | **7** |
| | 0 | 74 | 64 | **77** | 63 | 81 | **96** | 35 | 24 | **10** |
| Talking | 10 | **82** | 49 | 52 | 57 | **83** | 67 | 33 | 48 | **23** |
| | 5 | **52** | 49 | 48 | 61 | 78 | **89** | 41 | 50 | **24** |
| | 0 | 49 | 48 | **56** | 61 | 62 | **95** | 43 | 51 | **23** |
| Car | 10 | **83** | 72 | 74 | 86 | **91** | 86 | **15** | 16 | 17 |
| | 5 | 59 | **76** | 69 | **93** | 90 | 86 | 19 | **18** | **18** |
| | 0 | 52 | **67** | 54 | **93** | 86 | 87 | 22 | **20** | **20** |
| Exhibition | 10 | **93** | 62 | 65 | 50 | **89** | 79 | 33 | **24** | 26 |
| | 5 | **92** | 60 | 55 | 50 | **86** | 83 | 34 | **25** | 26 |
| | 0 | 44 | 59 | 47 | 93 | 84 | 85 | **26** | 27 | 27 |
| Subway | 10 | **97** | 64 | 74 | 74 | **85** | 79 | 16 | 24 | **15** |
| | 5 | **98** | 61 | 67 | 66 | 81 | **90** | 21 | 29 | **17** |
| | 0 | **98** | 59 | 59 | 68 | 78 | **91** | 20 | 30 | **19** |
| Street | 10 | **90** | 51 | 76 | **86** | 82 | 82 | **12** | 35 | 18 |
| | 5 | **90** | 47 | 74 | 82 | 79 | **86** | **14** | 39 | 17 |
| | 0 | **84** | 30 | 68 | 81 | 75 | **88** | 17 | 43 | **14** |
| Restaurant | 10 | **93** | 52 | 64 | 83 | **86** | 77 | **13** | 38 | 26 |
| | 5 | **92** | 51 | 63 | **83** | 81 | 81 | **13** | 43 | 23 |
| | 0 | **82** | 47 | 57 | 83 | 77 | **85** | **17** | 47 | 23 |
| Train | 10 | **94** | 64 | 79 | 88 | **94** | 46 | **10** | 22 | 41 |
| | 5 | **93** | 45 | 65 | **88** | 75 | 58 | **10** | 43 | 38 |
| | 0 | **92** | 45 | 54 | **88** | 64 | 65 | **11** | 45 | 38 |