16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

VOICE ACTIVITY DETECTION USING PERIODIOC/APERIODIC COHERENCE FEATURES

Sofia Ben Jebara

Research Unit TECHTRA
Ecole Supérieure des Communications de Tunis, TUNISIA
Phone:+216 71 857 000, Fax: +216 71 856 829, email: sofia.benjebara@supcom.rnu.tn

ABSTRACT

This paper introduces novel features for Voice Activity Detection (VAD). They are based on the coherence function between the considered frame and its LPC residue, calculated for both periodic and aperiodic components. The development of these features was motivated by the possible distinction between the periodicity and the aperiodicity character of speech and noise frames. Two statistical based decision techniques are used, they are the Discriminant Analysis (DA) and Gaussian Mixture Models (GMM) based bayesian classifier. We tested the proposed VAD technique on TIMIT database. We obtain consistent improvement as compared to features without periodic and aperiodic decomposition. In addition, we obtain encouraging results in real environmental noise.

Index Terms: Voice Activity Detection, Periodic/Aperiodic Coherence based features, Discriminant Analysis, Gaussian Mixture Model classifier.

1. INTRODUCTION

The Voice Activity Detection (VAD) is a classification problem used to discriminate between speech frames and silence frames in an audio sequence. It plays a crucial role in many speech processing techniques such as speech enhancement, speech coding, speech recognition, voice over IP, mobile telephony, etc.

The VAD task becomes difficult in the presence of background noise which alters the characteristics of the speech waveform because of its high level or its characteristics which may be similar to that of speech (talking noise, street noise, etc).

In general, VAD consists on two parts: acoustic features extraction and decision module. The feature extraction reduces the input data dimensionality by representing the frame to be classified by a reduced number of parameters. The features nature depends on the application (Mel Frequency Cepstral Coefficients for speech recognition [1], powers in band-limited regions for UMTS variable rate speech coding [2], delta line spectral frequencies for G.729 speech coding [3], etc).

Another category of features which can be used for VAD is related to speech linear prediction. In fact, speech frames, well modeled by an auto-regresssive model can be described by the prediction coefficients and the LPC residual error (see for example [4]). In previous works, we developed features which manipulate both linear prediction and coherence technique. More precisely, the similarity between the residual prediction signal and the signal itself is exploited in the frequency domain [5], creating coherence features.

In this paper, such coherence features are improved thanks to Periodic/APeriodic decomposition (PAP). Our method decomposes observed frames into their periodic and aperiodic components and calculates coherence features of these components. The term ‘aperiodic component’ includes both environmental noise and speech aperiodic components. The term ‘periodic component’ includes the dominant harmonic part of speech.

The classification strategy can be either manual or machine learning. The manual strategy makes use of thresholds for each feature provided by the user. The machine learning algorithms are usually based on statistical methods such as Neural Networks and Hidden Markovian Models. In this work, we use Discriminant Analysis technique and bayesian approach based on Gaussian Mixture Models (GMM).

The paper is organized as follows. Section 2 gives an overview about coherence based features developed in previous works. Section 3 is devoted to the description of proposed features which are justified in detail in section 4. The decision strategies based on DA and GMM are presented in section 5. Experimental results with white Gaussian noise and real environments noises are given in section 6. Finally, concluding remarks are drawn in section 7.

2. COHERENCE BASED FEATURES OVERVIEW

2.1 Basic idea

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 $n(k)$:

$$x(k) = s(k) + n(k). \tag{1}$$

It is well known that speech can be modeled by an autoregressive process. Its prediction error is given by:

$$e_s(k) = s(k) - P^T(k)S(k-1), \tag{2}$$

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 $S(k-1) = [s(k-1), s(k-2), .., s(k-L_P)]^T$ is the past input vector. Classically, when considering the quasi-stationarity of speech, the predictor is calculated frame by frame using the classical Levinson-Durbin algorithm.

The linear prediction residue constitutes the excitation source [6]. It is a quasi-random white noise for unvoiced frames and a quite periodic signal for the voiced frames. The similarity between the considered speech frame and its prediction residue is then weak. On the other hand, during silence, the considered frame is noise and the similarity between noise and its prediction residue is huge (since the noise is not an autoregresive process).

For noisy speech signals, the prediction residue is composed of two terms:

$$e(k) = e_s(k) + \tilde{n}(k), \tag{3}$$

where $e_s(k)$ is the prediction residue of $s(k)$ and $\tilde{n}(k)$ is an additive noise related to $n(k)$. When $n(k)$ is white, $\tilde{n}(k) = n(k)$ and the properties about similarity between clean speech and its prediction residue are maintained for noisy speech.

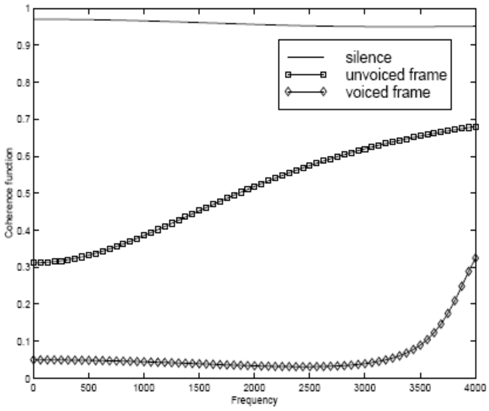

Figure 1: Averaged coherence function between noisy signal and its prediction residue .

According to this noting, the similarity between the considered speech frame and its prediction residue is a possible solution to propose features for voice activity detection. It is operated on the frequency domain by means of the coherence function.

2.2 Coherence function

The well known coherence function is usually used to measure the similarity between two signals [7]. It is developed in the frequency domain where short-time spectrum are computed using FFT. The data are segmented into frames of 16 ms and the coherence function of the m^{th} segment is defined as:

$$\mathbf{C}_{x,e}(m,f) = \frac{\mathbf{P}_{x,e}(m,f)}{\sqrt{\mathbf{P}_{x,x}(m,f)\mathbf{P}_{e,e}(m,f)}}, \qquad (4)$$

where $\mathbf{P}_{x,x}(m,f)$ and $\mathbf{P}_{e,e}(m,f)$ are spectral densities of m^{th} frame of signals $x(k)$ and $e(k)$ respectively and $\mathbf{P}_{x,e}(m,f)$ is the inter-signal spectral density.

Fig.1 represents the averaged coherence function obtained by averaging $\mathbf{C}_{x,e}(m,f)$ for a large number of voiced, unvoiced and silence frames. We notice that pure noise during silence intervals has coherence values around one for the whole frequency interval. The voiced frames have small coherence values (between zero and 0.3) whereas unvoiced frames have intermediate coherence values.

Fig.1 inspires us the following idea: instead of manipulating all frequency bins of coherence function for VAD descriptor, we simplify the procedure by considering their mean:

$$\mathbf{E}^m = \frac{1}{N} \sum_{f\in[0,F_e/2]} |\mathbf{C}_{x,e}(m,f)|, \qquad (5)$$

where F_e is the sampling frequency and N is frequency bins number during FFT calculation.

The ranges of values of $\mathbf{E}^m$ for speech and silence are well separated (large for silence and small for speech). Such descriptor permits to limit the number of features and hence reduces the calculus complexity .

3. PAP COHERENCE FEATURES

According to the fact that the descriptor $\mathbf{E}_m$ is large during silence and small during active speech, the VAD decision can simply be limited to thresholding. 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, etc).

In a previous work, we improved the decision part by using fuzzy logic based decision which is suitable for problems requiring approximate rather than exact solutions [8]. In this paper, we investigate an alternative solution to improve VAD descriptors by introducing the PAP decomposition.

3.1 PAP overview

We propose to refine the previous feature by taking into account more properties about speech components and their related prediction residues. In fact, we propose to decompose the speech observation into two sub-signals: the periodic and the aperiodic components. The periodic component is associated to the “deterministic” or “harmonic” part while the aperiodic component is associated to the “stochastic” or “random” or “noise” part of the observation. For noise-free speech, we separate harmonic speech parts from noise-like speech parts. In case of noisy speech, we separate harmonic speech parts from noise-like parts of both speech and background noise. In fact, in major cases, the background noise is randomly distributed and does not include harmonic parts. The justification of the use of PAP decomposition for VAD is detailed in section 4.

In this paper, we used the periodic/aperiodic decomposition proposed in [9]. A first approximation of the aperiodic component is obtained using the liftered cepstrum principle which estimates first the pitch location and then the inharmonic frequency regions. The aperiodic approximation is refined using an iterative algorithm based on successive Discrete Fourier Transforms and Inverse Discrete Fourier Transforms. After algorithm convergence, the periodic component is obtained by subtracting the reconstructed aperiodic component from the considered speech frame.

3.2 Features principle

The VAD features algorithm steps are the following (see Fig. 2).
- The whole speech sequence is decomposed into two sub-signals: the periodic sub-signal and the aperiodic sub-signal.
- Each sub-signal is segmented into frames of 16 ms duration. We denote $x_p^m(k)$ (resp. $x_{ap}^m(k)$) the k^{th} sample relative to m^{th} frame of periodic (resp. aperiodic) part and $e_p^m(k)$ (resp. $e_{ap}^m(k)$) the related prediction residues.
- The m^{th} frame coherence functions of aperiodic and aperiodic parts are defined as

$$\mathbf{C}_{x_j,e_j}(m,f) = \frac{\mathbf{\Gamma}_{x_j,e_j}(m,f)}{\sqrt{\mathbf{\Gamma}_{x_j,x_j}(m,f)\mathbf{\Gamma}_{e_j,e_j}(m,f)}}, \qquad (6)$$

where $j \in \{p, ap\}$ denotes the kind of sub-signal (periodic or aperiodic), $\mathbf{\Gamma}_{x_j,x_j}(m,f)$ and $\mathbf{\Gamma}_{e_j,e_j}(m,f)$ are spectral densities of signals $x_j^m(k)$ and $e_j^m(k)$ respectively. $\mathbf{\Gamma}_{x_j,e_j}(m,f)$ is the inter-signal spectral density between $x_j^m(k)$ and $e_j^m(k)$.
- The average of each coherence function in the whole frequency band is calculated and constitutes the VAD features.

$$\mathbf{CFF}_j^m = \frac{1}{N} \sum_{f\in[0,F_e/2]} |\mathbf{C}_{x_j,e_j}(m,f)|. \qquad (7)$$

Both $\mathbf{CFF}_p^m$ and $\mathbf{CFF}_{ap}^m$ constitute the set of parameters to be used for VAD.

4. JUSTIFICATION OF PROPOSED FEATURES

4.1 Particular case of ideal periodic and ideal aperiodic speech frames

Let’s consider two noisy speech frames characterized by the same coherence feature value α calculated on original speech

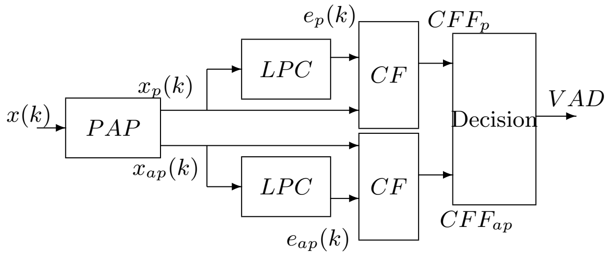

Figure 2: Block diagram of the proposed VAD.

without PAP. The first speech is an 'ideal periodic' frame (denoted s_p) composed of only periodic components and the second one can be either an 'ideal aperiodic' speech frame or a silence frame in presence of noise composed of only noise-like components (denoted s_{ap}).

They are both considered in the noisy context. The noisy frame is written $x^m(k) = s_p^m(k) + n^m(k)$ for ideal periodic frame and it is written $x^m(k) = s_{ap}^m(k)+n^m(k)$ for ideal aperiodic frame. $n^m(k)$ is the additive noise and we'll consider the case of white Gaussian noise, which corresponds to an ideal aperiodic signal.

After PAP decomposition, we obtain the following sub-signals and coherence features.

- Ideal periodic speech frame: The periodic sub-signal is exactly the considered speech frame $x_p^m(k) = s_p^m(k)$ whereas the aperiodic sub-signal is the considered noise $x_{ap}^m = n^m(k)$. The periodic LPC residue $e_p^m(k)$ is the same as the one obtained for clean speech $s_p^m(k)$. We show easily by coherence features calculus and comparison that $\mathbf{CFF}_p^m \leq \alpha$.
  The aperiodic LPC residue $e_{ap}^m(k)$ is the additive noise $n^m(k)$ and the coherence feature is close to one $\mathbf{CFF}_{ap}^m \approx 1$ [11].
- Ideal aperiodic speech frame: The periodic sub-signal is quasi-null $x_p^m(k) \approx 0$ and the aperiodic sub-signal is composed of the considered speech frame and the additive noise $x_{ap}^m(k) = s_{ap}^m(k) + n^m(k)$. The periodic LPC residue is then null $e_p^m(k) \approx 0$ and the coherence feature is close to one $\mathbf{CFF}_p^m = 1$. The aperiodic LPC residue is composed of the speech frame residue which is denoted $e_{s_{ap}}^m(k)$ and the additive noise $e_{ap}^m(k) = e_{s_{ap}}^m(k) + n^m(k)$. We show easily by coherence features calculus and comparison that $\mathbf{CFF}_{ap}^m \approx \alpha$ [11].

Hence, we can conclude that speech frames having the same coherence feature $\mathbf{E}^m$ can be differentiated according to their periodic and aperiodic coherence features. In fact, periodic frames have smaller values of periodic coherence feature than the initial coherence feature without PAP whereas its aperiodic coherence feature is close to one. By the other side, the aperiodic speech frames have the periodic coherence feature equal to one whereas the aperiodic coherence feature is closer the one obtained without PAP.

4.2 Case of real speech frames

In practice, a speech frame is a mixture of periodic and aperiodic components. Periodic parts are for example steady parts of vowels and voiced consonants, aperiodic parts are for example fluctuations included in vowels, stop, fricative and affricate consonants. We expect that periodic and aperiodic coherence features during silence are always close to one. For speech frames, aperiodic components coherence is different from the classic coherence feature. However, periodic components coherence depends on the frame type. For unvoiced frames, it is close to one. For voiced frames, the range of variation is large since the coherence depends strongly on voicing importance.

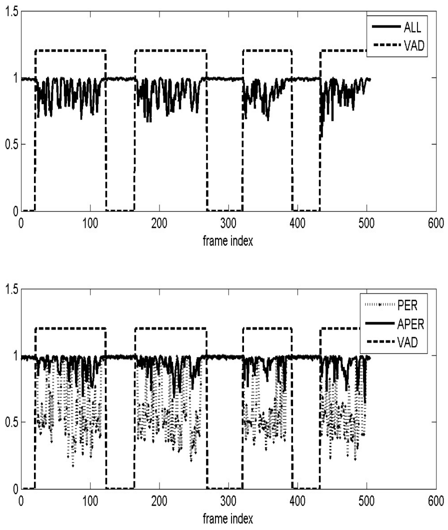

Figure 3: Evolution of $\mathbf{E}^m$, $\mathbf{CFF}_p^m$ and $\mathbf{CFF}_{ap}^m$ for speech and silence frames.

As an illustration of the previous analysis, Fig. 3 shows the evolution of coherence, periodic and aperiodic coherence features for a noisy speech sentence ($SNR = 10\,\text{dB}$) composed of intervals of silence and speech indicated by the real VAD (multiplied by 1.2). We notice that silence frames have all coherence features close to one. During speech frames, the coherence feature $\mathbf{E}^m$ ranges from 0.55 to 1. The aperiodic coherence feature $\mathbf{CFF}_{ap}^m$ ranges from 0.6 to 1 whereas the periodic coherence feature $\mathbf{CFF}_p^m$ ranges from 0.2 to 1. Such large range of variation will helps on frames classification.

5. VAD DECISION STRATEGY

There are many techniques for features classification. We are interested in statistical supervised techniques where a training data is used to construct decision functions. Then, performances are evaluated in test data. the data used in this work comes from the popular TIMIT database. We used 300000 speech frames pronounced by 438 male and 192 female speakers and 180000 silence frames. 60% of frames are used for training and 40% are used for test.

5.1 Discriminant Analysis for VAD

Discriminant Analysis (DA) is a parametric classification approach which uses a decision function that tries to maximize the distance between the centroids of each class of the training data and at the same time minimizes the distance of the data from the centroid of the class to which it belongs. It is named linear if the decision function is linear in the input data, quadratic if the decision function is quadratic,...

We used the training sequence to estimate the Discriminant Analysis classifier in noiseless case. Fig.4 shows the two classes in the plan (periodic-aperiodic) features. The regions for the two groups are well separated. We notice also that, as expected (section 4), silence periodic and aperiodic features are close to one. Speech frames periodic and aperiodic features take a wide range of values in the interval $[0, 1]$ and lie in the remaining part of the plan.

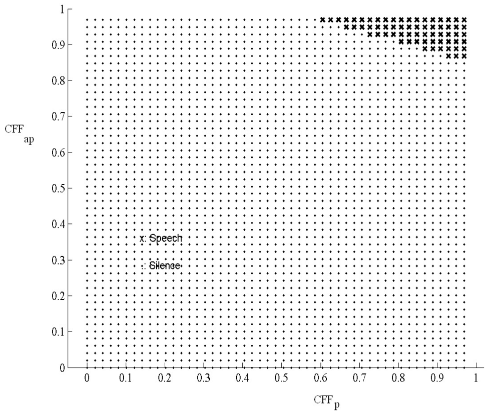

Figure 4: Speech and silence classes using Discriminant Analysis in noiseless case.

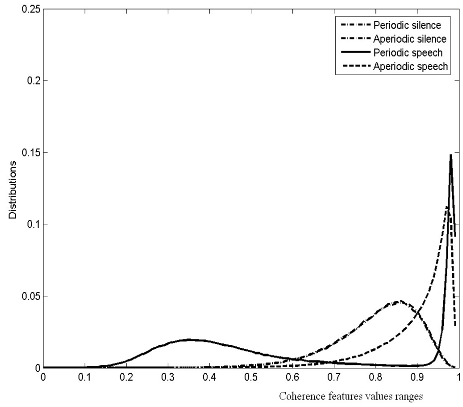

Figure 5: GMM distributions for the VAD classes.

5.2 Bayesian calssifier for VAD

We used bayesian classification based on probability theory. The posterior probabilities are then computed with the Bayes formula and one class is chosen if it has the highest posterior probability. We used the GMMBayes Matlab toolbox [10] which contain efficient classification functionality (training and classification) based on statistical theory (Bayesian inference) and Gaussian mixture model probability densities.

A Gaussian mixture is a weighted sum of Gaussian distributions whos model parameters are computed from the training data using Figueiredo-Jain algorithm which finds the "best" overall model directly using an iterative approach. The method is based on *Minimum Message Length* MML-like criterion which is directly implemented by a modification of the *Expectation-Maximization* algorithm (EM) [12].

Fig. 5 illustrates features histograms calculated for noisy training database ($SNR = 10\,\text{dB}$). It can be seen that
- Periodic and aperiodic silence histograms are merged into one histogram and occupy a quite large range around 0.85.
- Aperiodic speech histogram is concentrated around 0.95
- Periodic speech histogram occupy two separate regions (one large region around 0.35 and one small region with a peak atound 0.9.

6. EXPERIMENTAL RESULTS

To examine the validity of the proposed features, we conducted experiments using clean TIMIT database for the speech data. We added silence intervals between sentences and we added an artificial white Gaussian noise to simulate the noisy environment. We tested the following features.

- the coherence feature $\mathbf{E}^m$ calculated for original speech sequences (without PAP).
- 9 coherence features calculated for original speech sequences. Each coherence feature is obtained in a selected frequency band. Such features are developed for voiced/unvoiced/silence speech classification [13].
- Proposed periodic and aperiodic coherence features.

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_{sp} (resp. P_{si}): the probability of correct speech (resp. silence) decision. It is calculated as the ratio of correctly classified speech (resp. silence) frames to the total number of speech (resp. silence) frames.

We used the Discriminant Analysis and bayesian classification based on GMM modelization for the three kinds of features. Tab. 1 illustrates performances of VAD. The two cases of noiseless and noisy environments ($SNR = 10\,\text{dB}$) are tested. Tab. 1 permits the following interpretations.

- In noiseless case, the VAD is well precise (P_e varies from 3.34% to 1.69%). The GMM 9 bands gives better results in term of P_e. The PAP improves performances in term of probability of correct detection. DA is better for speech classification while GMM is better for silence classification.
- In noisy case, the error rate increases. The GMM_{PAP} is the best in term of P_e while DA_{PAP} is well suited for speech detection and DA_{9B} is well suited for silence detection.
- When comparing different features for the same classification technique (GMM or DA), we notice that each kind of features is suited for a selected criterion. However, the PAP decomposition improves performances in many cases.

Table 1: Performance in term of speech/silence classification with TIMIT database.

(%)	Technique	P_e	P_{sp}	P_{si}
Noiseless	DA_{glob}	3.34	98.52	95.57
	DA_{9B}	**1.86**	98.2	**96.6**
	DA_{PAP}	3.18	**99.19**	95.44
	GMM_{glob}	3.33	97.7	96.08
	GMM_{9B}	**1.69**	95.5	95.4
	GMM_{PAP}	2.75	**98.72**	**96.40**
$SNR = 10\,\text{dB}$	DA_{glob}	18.13	96.91	72.60
	DA_{9B}	23.7	76	**95.7**
	DA_{PAP}	**14.69**	**98.35**	77.28
	GMM_{glob}	17.72	**93.62**	75.29
	GMM_{9B}	17.27	70	**88**
	GMM_{PAP}	**13.66**	92.67	82.44

Furthermore, we analyze the influence of the amount of noise in VAD performances. Hence, for different values of SNR, we calculate the different features and test different classification techniques. Results are summarized in Tab. 2 for a long speech sequence of duration 1 minute. We notice the improvement of performances thanks to the use of PAP decomposition in features calculus.

Table 2: Performance in term of speech/silence classification for additive white Gaussian noise. Case of a long speech sequence.

(%)	Technique	P_e	P_{sp}	P_{si}
Noiseless	DA_{glob}	7.5	**97.87**	87.28
	DA_{PAP}	5.54	90.16	**98.9**
	GMM_{glob}	6.73	89.97	96.75
	GMM_{PAP}	**5.49**	91.42	97.76
$SNR = 20\,\text{dB}$	DA_{glob}	7.32	86.13	**99.59**
	DA_{PAP}	**5.84**	89.01	**99.59**
	GMM_{glob}	6.63	87.66	99.39
	GMM_{PAP}	**5.84**	**89.39**	99.19
$SNR = 10\,\text{dB}$	DA_{glob}	13.55	73.6	**100**
	DA_{PAP}	**10.89**	**79**	99.7
	GMM_{glob}	13.31	74.73	99.29
	GMM_{PAP}	**10.84**	**79.17**	99.7
$SNR = 0\,\text{dB}$	DA_{glob}	32.1	38.63	98.7
	DA_{PAP}	30.08	41.3	**100**
	GMM_{glob}	29.14	49.57	93.29
	GMM_{PAP}	**27.61**	**52.46**	93.39

We also propose to justify our approach for real environment noise, namely car noise, flat noise and babble noise. The car noise is correlated and present low frequency spectral characteristics, the flat noise looks like a white noise and the babble noise contains some tone components which can be viewed as harmonic components of speech. Tab. 3 illustrates the proposed algorithm performances. It confirms once again the usefulness of PAP decomposition to improve VAD performances (except in case of automobile noise). Furthermore, we remark that we detect more efficiently silence fra mes than speech frames. In fact, some speech frames such as unvoiced frames looks like noise and are not well selected.

Table 3: Performance in term of speech/silence classification for different kinds of noise $SNR = 10\,\text{dB}$.

(%)	Technique	P_e	P_{sp}	P_{si}
Babble	DA_{glob}	32.49	44.51	**91.77**
	DA_{PAP}	28.5	79.46	63.11
	GMM_{glob}	29.84	53.04	88.21
	GMM_{PAP}	**27.21**	**72.13**	73.48
Flat communications noise	DA_{glob}	23.05	55.3	99.7
	DA_{PAP}	9.85	81	99.7
	GMM_{glob}	17.91	66.25	98.78
	GMM_{PAP}	**9.4**	**81.97**	**99.9**
Automobile highway noise	DA_{glob}	22.21	60.5	**96.04**
	DA_{PAP}	35.28	41.47	89.23
	GMM_{glob}	**22.17**	**77.82**	77.85
	GMM_{PAP}	38.84	75.22	46.34

7. CONCLUSION

In this paper, we proposed a noise robust VAD method based on periodic and aperiodic coherence features and statistical decision techniques. The experiments confirmed that the proposed features perform better than those obtained without PAP decomposition. In the future, we will affine periodic and aperiodic coherence features by considering them in different frequency bands as it was done for VAD without PAP.

REFERENCES

[1] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," *Proc. of INTERSPEECH*, Lisboa, Portugal, 2005.

[2] ETSI Standard documentation, ETSI ES 202 050 V1.1.3, 2003.

[3] ITU-T Recommandation G729 Annex B., 1996.

[4] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Acoust. Speech and Signal Processing*, vol.9, pp. 217-231, March 2001.

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

[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] G. C. Carter, "Coherence and Time-Delay Estimation", *Proc. of IEEE*, Vol. 75, pp. 236-255, February 1987.

[8] S. Ben Jebara and T. Ben Amor, "On improving voice activity detection by fuzzy logic rules: case of coherence based features," *Proc. of the European Signal Processing Conference EUSIPCO*, Vienna, Austria, Sept. 2004.

[9] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE. Trans. Speech and Audio Processing,* vol. 6, no. 1, pp. 1–11, Jan. 1998.

[10] http://www.it.lut.fi/project/gmmbayes/.

[11] S. Ben Jebara, "On the use of PAP decomposition for voice activity detection," *internal report TECHRA-01*, January 2008.

[12] M. A. T Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no.3, pp. 381-396, March 2002.

[13] S. Ben Jebara, "Multi-band coherence features for voiced-voiceless-silence speech classification," *Proc. of Int. Conf. on Informatics & Communications Technologies from theory to applications ICTTA*, Damascus, Syria, 2005.