

Voice Activity Detector Based on Enhanced Cumulant of LPC Residual and On-line EM Algorithm

David Cournapeau¹, Tatsuya Kawahara¹, Kenji Mase^{2,3} and Tomoji Toriyama³

1: Graduate School of Informatics, Kyoto University, Japan

2: Graduate School of Information Science, Nagoya University, Japan

3: Media Information Science Lab, ATR, Japan

Abstract

This paper addresses the problem of segmenting audio data recorded with embedded devices for the purpose of intelligent sensing in the context of multi-modal interactions. We propose a real-time method for robust speech detection in natural, noisy environments. It is based on a fusion of high order statistics of the LPC residual and autocorrelation, and adopts an on-line version of Expectation Maximization algorithm for the classification. Experimental evaluations show that the proposed method provides better detection performance under different types of natural noises, working robustly against other voices in the context of multi-speaker interactive situations. As the proposed method is based on features which have a low computational cost, and has a small latency, it is suitable for real-time tracking applications.

Index Terms: voice activity detection, high order statistics, on-line EM

1. Introduction

The problem of detecting voice in audio sources, also called VAD (voice activity detection), is a classical problem in speech processing. It is a common front-end in most tasks involving speech processing. It is for example used as a front-end to automatic speech recognition (see [1]); having a VAD algorithm robust to noisy environments is crucial for recognition performance [2]; it is also used in speech coders, such as GSM 729. More recently, VAD has been used as a feature for conversational scene analysis [3] and multi-modal recognition of group actions or meetings ([4], [5]).

The VAD algorithm presented here was developed for those multi-modal applications; we intend to use it on wearable capture systems, which capture audio and video data to help the users with contextual information (see [6] for a presentation of the system). Segmenting audio for this purpose is significantly different from traditional tasks where VAD is studied. First, in the context of audio coding or speech recognition, the assumption of high proportion of speech can be made, and missing speech sections has a higher cost than detecting non-speech as speech. In other words, for those applications, the VAD algorithm is usually pretty conservative when detecting voice; a low False Rejection Rate (FRR) is preferred to a low False Alarm Rate (FAR). In our case, we are more interested by having a precise idea about when does the speaker takes its turn, while keeping a small FAR.

Generally, finding speech in noisy environments is a difficult task: simple techniques based on an adaptive threshold on simple features such as energy or zero crossing-rate work well in high SNR environments, but fall short in other cases. The situations

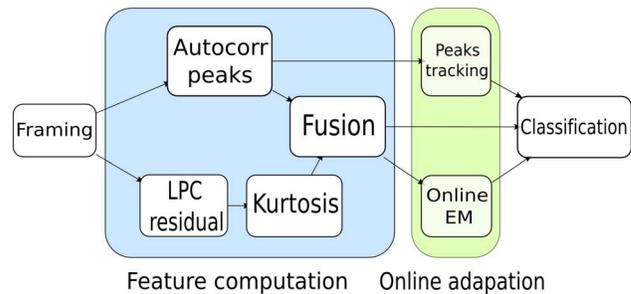


Figure 1: Overview of the proposed method

we are interested in are very adverse conditions for VAD: environmental noise changes in time, and the volume of the user utterances is not fixed because users may change the microphone position and direction. Moreover, as those situations inherently involve several speakers, the algorithm must be insensitive to other speaker voices. The proposed approach uses a feature based on high order statistics to be robust against background voices of moderate levels, enhanced by the use of autocorrelation, and adopts an on-line version of EM algorithm to adapt the voice decision step to environment and speaker variations.

The overview of the proposed method is depicted in Figure 1, and the details are explained in the rest of the paper as follows: section 2 briefly describes the task and data used in our work; in section 3, the new feature based on LPC residual and autocorrelation is introduced, and in section 4, the use of on-line EM algorithm for classification is described; the proposed method is evaluated in comparison to conventional methods in section 5.

2. Task and database

We are building an interaction corpus, which consists of a large amount of data of human communications. The goal of this corpus is to support the users using contextual, automatically extracted information ([6]), gathered with wearable sets and sensors in a smart room. VAD is one of the features used to sense the interactions and control other devices.

In this paper, we use a portion of this interactive corpus to develop and test an effective VAD algorithm. The data were recorded in the following conditions: people were wearing the embedded device equipped with a microphone in a room with

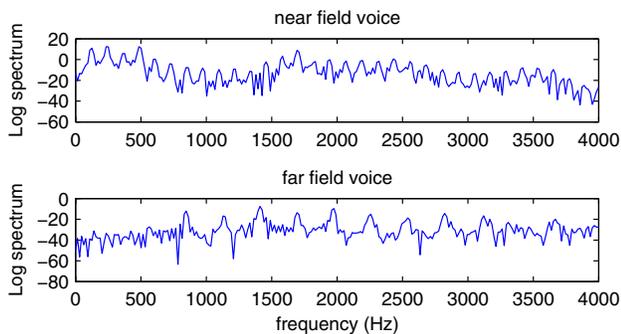


Figure 2: Example of log spectrum for near field speech (top) and far field speech (bottom)

other people. They were visitors to the lab during an openhouse. The data contain several kind of noises (air conditioning, other people, cars running on the street, etc.). The test database contains around 45 minutes of audio data, split into around 30 files of the same length, each file containing different speaker (thus different microphone position), different gender, different language (mainly Japanese, but also English), different sparsity and different SNR (between 10 dB and 25 dB).

3. Proposed feature

3.1. LPC residual and high order statistics

To be robust against background voices, we cannot directly use standard features related to pitch such as autocorrelation or zero-crossing. The most obvious feature would be energy, since the energy of recorded signal is directly dependent on the distance between the source of the sound and the microphone. But then we have to cope with the normalization problem for real-time applications [2]. Another characteristic observed on distanced speech is the loss of low frequency harmonics, and thus far-field speech has less harmonics than near-field speech, as shown for example in Figure 2

Instead of trying to capture directly the number of harmonics in the spectrum, we use the cumulants, also called high order statistics (HOS). Cumulants of a random signal X are derived from the logarithm of the moment generating function $\Phi(t)$ of X . More exactly, the cumulant of order n is defined as the n^{th} coefficient of the Taylor expansion series of the logarithm of $\Phi(t)$:

$$\log \Phi(t) = \log \mathbb{E}[e^{tX}] \quad (1)$$

$$= \sum_{n=0}^{\infty} k_n \frac{t^n}{n!} \quad (2)$$

We primarily considered normalized kurtosis (cumulant of order 4 normalized by the variance) and normalized skewness (cumulant of order 3 normalized by the variance) of the LPC residual which increase with respect to the number of harmonics when the signal can be well approximated by a sinusoidal model, as proved in [7]. Also, all cumulants of order strictly bigger than 2 are 0 for Gaussian distributed signals. This property makes them robust against some kinds of noises such as wideband noises. As there is an explicit relationship between the cumulant of order n

and the moments of order n and below, we can easily compute their values ¹. Both kurtosis (noted k) and skewness (noted s) were considered, either separately or together; the kurtosis was found to be more effective than skewness for our use.

3.2. Enhancing cumulant estimators

Use of cumulants involves several problems for VAD. Because of the weak convergence properties of standard estimators for skewness and kurtosis, the estimated values can be quite different from the true value; they are also quite sensitive to outliers, problem aggravated in the case of normalized estimators. This is a problem for certain kinds of noise, such as transient noises (noises well located in time). To enhance the behaviour of the estimators, we have to incorporate another feature whose behaviour does not change HOS distribution in the case of speech, but is insensitive to transient noises. In this study, we propose to combine HOS with normalized autocorrelation.

Autocorrelation is a good cue to indicate pitch, and is fairly robust to transient noises; for those reasons, it has often been used for VAD (for example in [8]). To improve robustness to energy variation of the signal, we use the normalized autocorrelation $a[k]$ for a frame $x[t]$, given by the following formula:

$$a[k] = \frac{\sum_{n=k-1}^N x[n]x[n-k]}{\left(\sum_{n=0}^{N-1} x[n]^2\right)^{\frac{1}{2}}} \quad (3)$$

For periodic signals of period T samples, the autocorrelation will have maxima at multiple of T lags. We detect a peak if its value is strictly bigger than its nearest neighbors on both sides. Because of the normalization process, though, peaks can appear for low energy noise which have a sharp spectrum (an example of such noise is motor noise); also, it is near useless by itself to discriminate between the main speaker's speech and background voices. However, in this study, the motive to use autocorrelation is that its peaks have low amplitude for transient noises, which are the most problematic noises when using HOS.

We combine autocorrelation's peak amplitude m and kurtosis of the LPC residual k as follows:

$$f = m \log(1 + k) \quad (4)$$

The logarithm is used to compensate high values of kurtosis in the case of really strong voiced frames; it also gives a more Gaussian-like behaviour to the feature, which is important for our classification method (see section 4). In the case of voice, both kurtosis and autocorrelation peak should have a high value; for distanced voice, the low kurtosis should compensate for the high autocorrelation value, and for transient noises, near 0 autocorrelation should compensate for the high kurtosis value. Figure 3 shows the behaviour of this feature for a small example. This extract contains mostly speech in the second half, and transient noises can be seen around 10 second. The enhancement of the proposed feature on standard kurtosis is apparent. Also, the relatively loud background speech noise in the first seconds, which present high autocorrelation peaks, is effectively suppressed by the low kurtosis values.

¹more exactly their estimated value

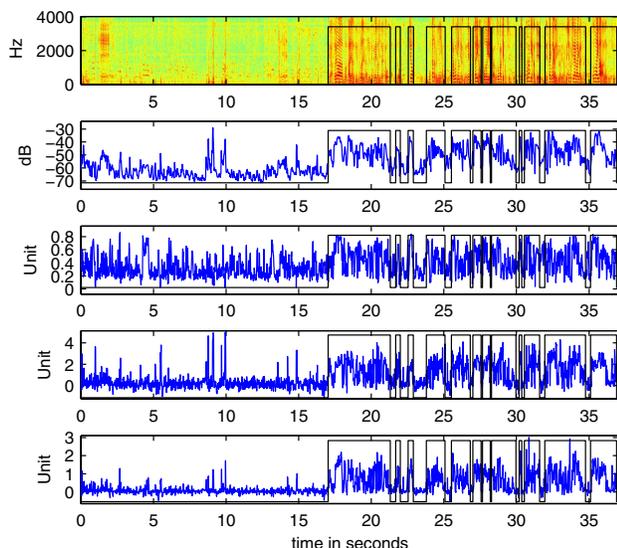


Figure 3: Audio example with spectrogram (top), energy (2^{nd}), main autocorrelation peak amplitude (3^{st}), "log kurtosis" (4^{th}) and the proposed feature (bottom). The black boxes are speech segments of the main speaker, hand-labeled.

4. Tracking the feature on-line

4.1. On-line EM algorithm

To demonstrate the effectiveness of the proposed feature in a straightforward manner, we adopt a naive Bayes classifier: each class c_i (main speaker's speech / other) is modeled as a Gaussian of mean μ_i and variance σ_i , with a prior $P(c_i) = w_i$; this is a simple binary Gaussian mixture model. The Gaussian parameters were estimated by the Expectation Maximization algorithm. The classification using the standard EM algorithm tested on the whole signal for each audio file gives satisfactory results, but is obviously not suited for real-time applications.

This section presents a simple adaptation of the batch EM algorithm for the on-line case. On-line versions of the EM algorithms have been proposed by several authors, for example [9]. The basic idea is simple: instead of running on the whole feature signal, the internal state of the EM algorithm (that is, the parameters of the E step) is updated at each frame, taking into account both the current feature vector and past feature vectors. The learning rate $\lambda(n)$, where n is the frame index, influences the convergence, and [10] gives the conditions on λ to obtain convergence.

We have to use some values for the EM algorithm's initial state; using random data may lead to problems such as one weight tending toward 0. Two simple solutions were implemented: one is to use some initial data (below one second of signal is enough); the other is to use random data generated from a Gaussian mixture model using prior parameters. Both methods gave similar results. Another issue is how to handle the case where the condition number of a covariance matrix approaches zero. In the on-line case, we use a regularization scheme, as presented in [9].

Table 1: Frame error rates for the proposed algorithm (top), using on-line EM on kurtosis only (2^{nd}), on-line EM on energy (3^{rd}) and using offline EM on the proposed feature (bottom)

	FAR	FRR	GER
Proposed algorithm	7.8 %	13.0 %	9.5 %
Using energy	15.8 %	10.6 %	13.3 %
Using kurtosis only	19.0 %	13.8 %	16.3 %
Offline EM	8.0 %	12.0 %	9.5 %

4.2. Tracking in correlogram

For computing the peaks of the autocorrelation for our feature, we also track them in real-time. We implement a tracking algorithm which simply builds tracks of peak positions between the current and the former frame; at frame t , for each peak candidate, we use the peak of frame $t + 1$ which minimizes the distance between frame t and frame $t + 1$. The parameters are quite 'loose', as the actual classification will be done by the on-line EM anyway.

5. Experimental evaluation

5.1. Evaluation measure

We use the most traditional metric: frame-level classification error, that is

- False Rejection Rate (FRR), defined as the ratio between the number of missed speech frames and the total number of speech frames
- False Alarm Rate (FAR), defined as the ratio between the number of incorrectly detected speech frames and the total number of non-speech frames
- Global Error Rate (GER), defined as the number of missed and incorrectly detected speech frames divided by the total number of frames

5.2. Results

We used the test data of 45 minutes split into 30 files, as described in section 2. The ratio of speech frames ranges from 10 to 90 %, with 33 % on average. The sampling rate is 8 khz, with a frame size of 256 samples: with a window of 256 samples, the pitch of human voice being typically in the range 80-250 Hz, we expect at least two pitch periods in each window. We compare the proposed method with three other methods: replacing the enhanced kurtosis with energy, using kurtosis only without enhancement by autocorrelation, and using offline EM instead of on-line EM.

The results are summarized in Table 1. With the proposed method, FAR is kept quite low, which is our main concern in the design of this algorithm. To get a more precise idea of the behavior of our implementation, we give in Figure 4 the frame error rates with respect to the ratio (speech/non speech) for each file from the test database, compared against the implementation using energy instead of enhanced kurtosis. The proposed feature has half as much false alarms as energy, and even though the FRR is a bit higher, it can be seen in Figure 4 that the two methods are not significantly different. On the contrary, it is observed that the energy-based method has significant higher FAR over the data of low speech ratio (below 20 %), which severely degrades GER as well.

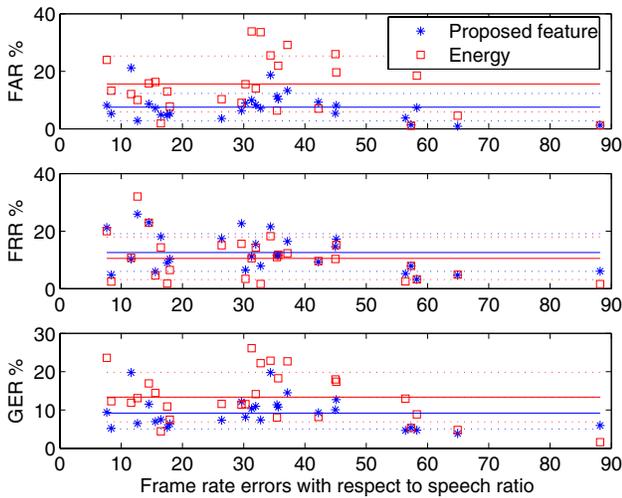


Figure 4: Results of the proposed VAD algorithm in function of the speech/non speech ratio, in comparison with energy-based method. The dashed lines show the standard deviation of each criterion, and solid line the mean.

To see the effect of the enhancement on the cumulant based feature, Figure 5 shows the frame error rates for the proposed feature and the standard normalized kurtosis (we still converted to logarithm to get a Gaussian-like behavior, though). Again, significant degradation due to enormous false alarms is observed for data of low speech ratio. Thus, without the autocorrelation enhancement, using directly the kurtosis is not really effective.

It is also interesting to see how effective the on-line EM is; we compare our results with an offline implementation of EM (i.e. the standard version of EM). The results are at the last row of Table 1. Within our experiment, the on-line EM is almost as effective as the traditional batch version; both results are nearly identical on every test sample, and we lose almost nothing by having a real-time version of the algorithm.

6. Conclusions

We have presented a new real-time algorithm for voice activity detection in natural environments. It works effectively to detect speech, while being robust against various kinds of noises, including other speakers' voices. On-line EM algorithm was also successfully implemented as an on-line adapting algorithm. As the proposed method is based on features which have a low computational cost, and has a small latency, it is suitable for real-time tracking applications.

7. Acknowledgments

This study was partly supported by the National Institute of Information and Communications Technology, Japan.

8. References

[1] Lawrence R. Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

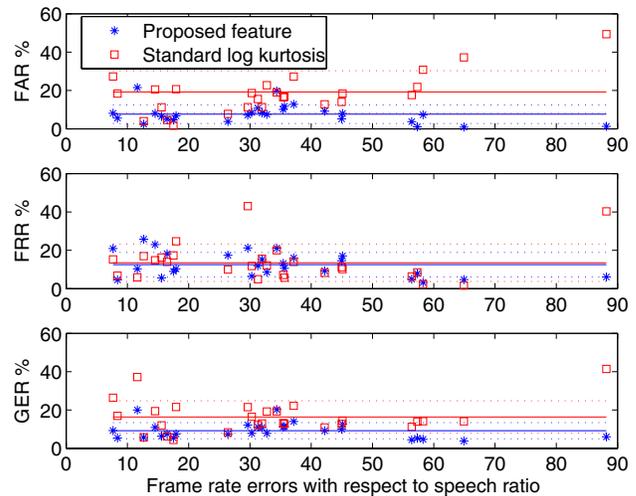


Figure 5: Comparison between the proposed feature and kurtosis-only based feature

[2] Qi Li, Jinsong Zheng, Qiru Zhou, and Chin-Hui Lee, "A Robust, Real-Time Endpoint Detector with Energy Normalization for ASR in Adverse Environments," in *ICASSP01*. IEEE, 2001.

[3] Sumit Basu, *Conversational Scene Analysis*, Ph.D. thesis, MIT, 2002.

[4] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang, "Automatic Analysis of Multimodal Group Actions in Meetings," in *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 2005.

[5] John S. Garofolo, Christophe D. Laprun, and Jonathan G. Fiscus, "The rich transcription 2004 spring meeting recognition evaluation," in *Proceedings of the Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP '04)*, 2004.

[6] Yasuyuki Sumi, Sadanori Ito, Tetsuya Matsuguchi, Sidney Fels, and Kenji Mase, "Collaborative Capturing and Interpretation of Interactions," in *Pervasive 2004 Workshop on Memory and Sharing of Experiences*, 2004.

[7] Elias Nemer, Rafik Goubran, and Samy Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions On Speech And Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[8] Sumit Basu, "A linked-HMM model for robust voicing and speech detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, 2003.

[9] Masa-aki Sato and Shin Ishii, "On-line EM algorithm for the normalized gaussian network," *Neural Computation*, vol. 12, pp. 407–432, 2000.

[10] Masa-aki Sato, "Convergence of on-line EM algorithm," in *7th International Conference on Neural Information Processing*, 2000, vol. 1.