IEICE TRANS. INF. & SYST., VOL.E91–D, NO.12 DECEMBER 2008

PAPER

Voice Activity Detection Based on High Order Statistics and Online EM Algorithm

David COURNAPEAU[†a)], *Nonmember and* **Tatsuya KAWAHARA**[†], *Member*

SUMMARY A new online, unsupervised voice activity detection (VAD) method is proposed. The method is based on a feature derived from high-order statistics (HOS), enhanced by a second metric based on normalized autocorrelation peaks to improve its robustness to non-Gaussian noises. This feature is also oriented for discriminating between close-talk and far-field speech, thus providing a VAD method in the context of human-to-human interaction independent of the energy level. The classification is done by an online variation of the Expectation-Maximization (EM) algorithm, to track and adapt to noise variations in the speech signal. Performance of the proposed method is evaluated on an in-house data and on CENSREC-1-C, a publicly available database used for VAD in the context of automatic speech recognition (ASR). On both test sets, the proposed method outperforms a simple energy-based algorithm and is shown to be more robust against the change in speech sparsity, SNR variability and the noise type.

key words: *speech recognition, voice activity detection, high order statistics, online EM*

1. Introduction

Voice activity detection (VAD), which automatically detects speech from audio signals, is a classical problem in speech processing. For example, it is often used as a front-end for automatic speech recognition (ASR) [1]. The problem has recently received attention because the effectiveness of the VAD front-end is crucial for the performance of the speech recognizer in noisy environments; when the background noise is high, the number of insertion errors becomes large [2], and having a VAD robust against noisy environments can significantly reduce the word error rate (WER). Other tasks where VAD is useful include speech coding and speaker recognition.

The work described in this paper aims at detecting speech in the context of human-to-human interaction. This situation poses several challenges, mainly because some of the assumptions usually made for ASR or speech coding, such as the signal containing speech most of the time, are not met in human-to-human interaction. Thus, a VAD algorithm needs to cope with this sparsity. Also, as several people are involved, it should be able to discriminate between the different speakers involved. One solution to this problem is using an array of microphones [3]. If use of close microphone recording is allowed, discriminating between the wearer of the microphone and other persons is possible if we can find a feature whose behavior is different in close-talk and far-field speech. The work presented here follows this approach. Even when close-talk recording is available, the simple strategy based on energy gives unsatisfactory results [4], mainly because of cross-channel talk and varying noise conditions.

The method we propose in this article aims at solving those problems. It incorporates a novel feature based on high-order statistics (HOS) for discriminating between close-talk speech and far-field speech, and an online, unsupervised classification scheme based on Online Expectation Maximization (OEM) algorithm to cope with varying noise condition and change of speech proportion. HOS can be defined from the moments of a random variable, and give information that is absent from the most commonly used moments: mean and variance. Using HOS for VAD has been suggested, for example in [5], whose strategy was refined in [6]. It was shown in [5] that the HOS of the LPC residual is an increasing function of the number of harmonics in the signal, using a sinusoidal model of speech [7]. As HOS are immune to Gaussian noise, it can be used for VAD in some noisy, Gaussian-like environments. However, HOS are sensitive to other kind of noises such as transient noises (noises which have a high energy and are well localized in time, which can occur for example when there is a physical contact with the microphone). In this paper, we combine HOS with another metric, derived from the normalized autocorrelation, to enhance its robustness to non-Gaussian-like noises. We investigate the effectiveness of the enhanced HOS to discriminate far-field speech and close-talk speech. We also propose a new scheme for online classification, using online EM. This method has an advantage of estimating online the noise and speech level concurrently with classification, without relying on a separate SNR estimation scheme such as the one used in G.729B VAD [8]. Figure 1. gives an overview of the whole scheme: first, the speech signal is divided in frames which are pre-processed and go through LPC analysis, whose residual is used as an input to the rest of the method. Kurtosis and the normalized autocorrelation, whose main peak is extracted, are computed, combined with each other, and used as the input of the online EM algorithm, which does the classification and model updating simultaneously.

The organization of this paper is as follows. Section 2 describes the use of high-order statistics for speech detection, its rationale, limitations and the method to alleviate those limitations. Section 3 describes the classification

Manuscript received June 24, 2008.
Manuscript revised August 5, 2008.

[†]The authors are with the School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: david@ar.media.kyoto-u.ac.jp

DOI: 10.1093/ietisy/e91–d.12.2854

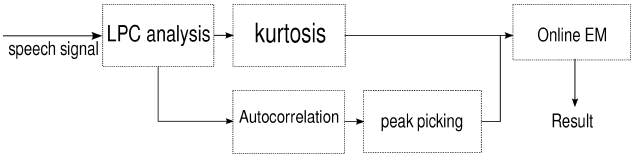

Fig. 1 Overview of proposed method.

scheme based on online EM. Section 4 evaluates the proposed method on both in-house data and a public database for VAD in the context of ASR in noisy environments, the CENSREC-1 database [9], and compares its performances to standard algorithms.

2. High Order Statistics for Speech Detection

2.1 Property of Close-Talk and Far-Field Speech

A variety of features have been proposed for VAD, such as energy, auto-correlation, cepstrum peaks [10], and MFCC [11]. The goal is to find a feature whose underlying distribution is different for speech signal and for non-speech signal. As our goal is to detect close-talk speech only, the feature also has to be able to discriminate between close-talk and far-field speech. The most obvious feature to discriminate between close-talk and far-field speech would be energy, but this does not work as expected; [4] identifies cross-channel talk and varying noise conditions as the main causes for degradation of performances in the simple energy-based algorithms. Also, as noted in [12], normalization of the feature with respect to the energy is crucial for online VAD. For those reasons, we focus on features independent of the energy.

We plot in Fig. 2 the LPC residual of close-talk speech and far-field speech, since it is known to relate with glottal excitation. The signal was recorded with two microphones, one close-talk and the other far-field, and we display the time-synchronized extract from both microphones. In both cases, the spectral envelope (middle column) is similar, and the pulses corresponding to the air flow variations as well as their periodicity are visible on the LPC residual (right column). But in the close-talk case, the LPC residual is relatively uncorrupted by the noise and the pulses have a much stronger amplitude on average. Thus, in close-talk speech, the signal amplitude $x(t)$ is either outside $[-\sigma, \sigma]$ range or approximately 0 (i.e. $|x| \ll \sigma$). On the other hand, for far-field speech, the amplitude is more likely to be around σ (i.e. $|x| \approx \sigma$).

There are several explanations for this difference: first, because the SNR is lower for distant speech and as such speech is embedded in the noise, and also because of reverberation, its LPC residual distribution is more Gaussian-like. Another possible explanation for this difference could be the proximity effect of the microphone. Most close-talk microphones are directional, and because directional microphones use two diaphragms, this results in the proximity effect of directional microphones. This proximity effect increases the low spectrum of the received signal for close signals (a few centimeters away from the microphone).

In summary, the distribution of the LPC residual is more likely to get extreme values (far from the mean, or close to the mean) in the close-talk case than in the far-field case. Following this property, discrimination between close-talk and far-field speech is reduced to discrimination between fat-tailed, peaky distribution against fat mid-range distribution. Kurtosis, which is one of HOS, is a standard statistics used for this purpose.

2.2 Definition of HOS

The HOS, also called cumulants, of random variables X are defined from the cumulant generating function ψ:

$$\psi(t) \triangleq \log \Phi(t) = \log \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} \kappa_n^X \frac{t^n}{n!} \quad (1)$$

that is, the cumulant generating function is defined as the logarithm of the moment generating function Φ, and the cumulant of order n κ_n is the nth coefficient of Taylor expansion divided by $n!$. There is a direct relationship between the cumulants of a random variable X and its central moments. For the first four cumulants, those are:

$$\kappa_1^X = \mathbb{E}[X - \mathbb{E}[X]] = 0 \quad (2)$$
$$\kappa_2^X = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sigma^2 \quad (3)$$
$$\kappa_3^X = \mathbb{E}[(X - \mathbb{E}[X])^3] \quad (4)$$
$$\kappa_4^X = \mathbb{E}[(X - \mathbb{E}[X])^4] - 3\sigma^{2^2} \quad (5)$$

The cumulant of order 2 is simply the variance. The most common high order statistics, skewness s_X and kurtosis k_X, are defined as normalized version of cumulants of order 3 and 4, respectively, with the normalization factor being σ^n, where σ is the standard deviation and n the order of the statistics:

$$s_X \triangleq \kappa_3^X/\sigma^3 = \mathbb{E}[(X-\mu)^3]/\sigma^3 \quad (6)$$
$$k_X \triangleq \kappa_4^X/\sigma^4 = \mathbb{E}[(X-\mu)^4]/\sigma^4 - 3 \quad (7)$$

One motivation of this definition is the additivity property for independent random variables, which is a direct consequence of the property of the moment generating function for independent random variables. Another direct consequence is that all cumulants of order $n \geq 3$ are 0 for Gaussian random variables, since the cumulant generating function of a Gaussian variable is a polynomial of order 2.

2.3 Kurtosis and Shape of Distribution

Kurtosis has been used for long in the statistics literature as a measure of non-Gaussianity, peakedness or tailedness of a random variable [13], [14]. Whereas the first and second moments of a random variable X may be seen as simply a translation and scale parameter, respectively, HOS contains

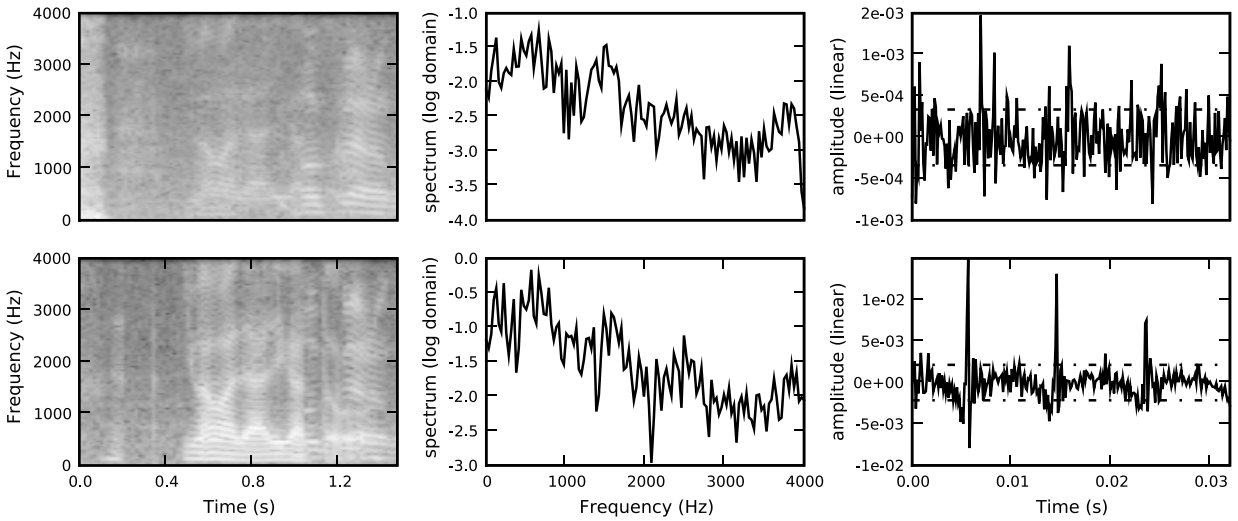

Fig. 2 Comparison between far-field (top) and close-talk speech (bottom). They correspond to the same signal. The left column displays the spectrogram of approximately one second of signal, the middle (spectrum) and right column (LPC residual) for one particular frame from the extract. The dashed line represents the standard deviation of each signal.

information on the shape of the distribution. Kurtosis measures both the peakedness and the tailedness of random variables, and both those characteristics have to be taken into account when comparing two random variables [13]. More formally, for two symmetrical random variables X and Y of equal mean and variance, if there are a and b such that

$$\forall x, |x| \leq a \text{ or } |x| \geq b, f_X(x) \geq f_Y(x) \tag{8}$$

while

$$\forall x, a \leq |x| \leq b, f_X(x) \leq f_Y(x) \tag{9}$$

then the fourth moment of X is higher than that of Y (see [14] for a proof). Taking a Gaussian random variable as a reference, an example of distribution having heavier tails and being more peaky than a Gaussian is the Laplace distribution, as displayed in Fig. 3. Whereas a Gaussian has a kurtosis of 0, the Laplace distribution has a kurtosis of 3 (both Laplace and Gaussian have a kurtosis which is independent of their parameters).

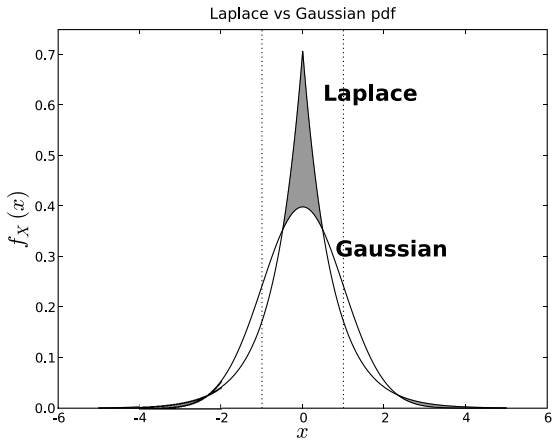

Fig. 3 Comparison of heavy-tailed and peaky distribution (Laplace) against Gaussian. Both have same mean (0) and variance (1), are symmetric, but Laplace has a kurtosis of 3, compared to 0 for Gaussian. Filled area emphasizes the ranges where values are more likely for Laplace.

2.4 Enhancing Kurtosis with Autocorrelation

Following the above discussion, kurtosis is expected to be a candidate to discriminate between far-field and close-talk speech. For the example of Fig. 2, the kurtosis is 15.4 for the close-talk speech, and 0.4 for the far-field speech. However, as already noted in [5], using HOS directly for VAD is not effective, for several reasons; the standard estimators for kurtosis, based on a sample estimator of moments, slowly converge to the true value, and are sensitive to outliers; also, non-Gaussian noises may not have a low value for kurtosis. In fact, typical noises on close-talk conditions such as contact noises, which are of highly transient nature, have a large kurtosis. This is observed in Fig. 4. The figure represents one extract of 37 seconds recorded by a close-talk microphone: the signal contains mostly speech starting around 17 second, but the whole beginning of the signal contains background speech, whose low frequency spectral lines can be seen on the spectrogram. It also contains transient noises around second 9-10, which are visible both on the spectrogram and the energy plot. The figure shows that kurtosis behaves differently for far and close-talk speech: it is mostly low value and stable for remote speech, whereas it has high value for close-talk speech. However, the kurtosis has some spikes, particularly for the transient noises around 9 second.

To enhance the above property, we propose a method to improve the kurtosis for transient noises, while keeping its desired behavior for discriminating far-field speech against close-talk speech; we combine it with the normalized auto-

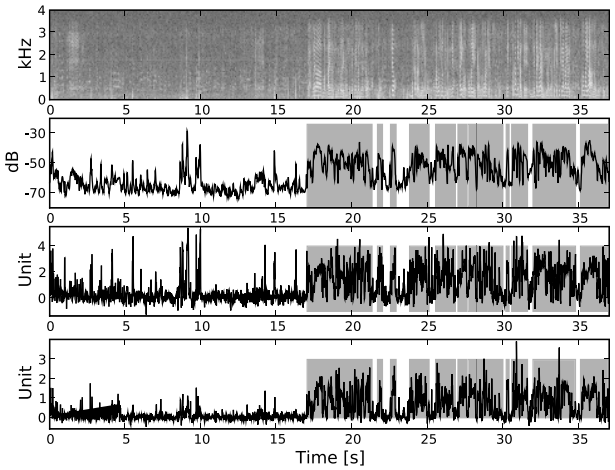

Fig. 4 Sample of in-house speech recording with natural noise. Spectrogram (top), energy (2nd), log-kurtosis (3rd) and proposed feature (4th). Filled areas are speech to be detected.

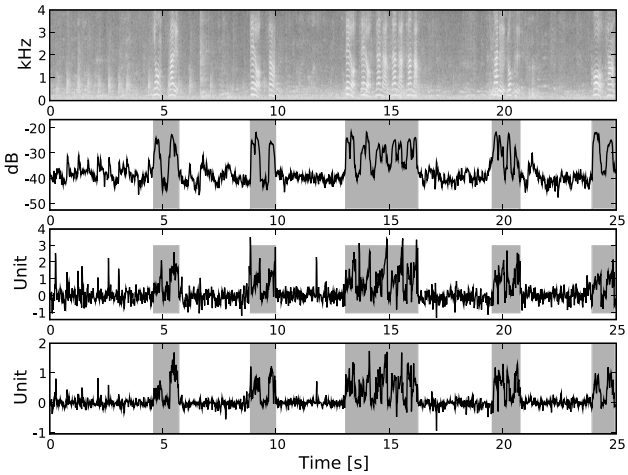

Fig. 5 Sample of speech from CENSREC-1 (high SNR). Spectrogram (top), energy (2nd), log-kurtosis (3rd) and proposed feature (4th).

correlation peak. Auto-correlation is a good cue to indicate pitch, and is fairly robust against transient noises; for those reasons, it has often been used for VAD (for example in [15]). To enhance robustness against energy variation of the signal, we use the normalized auto-correlation $a[k]$ for a frame $X = (x_t) = \{x_0, \ldots, x_{N-1}\}$, given by the following formula:

$$a[k] = \frac{\sum_{n=k}^{N-1} x[n] \cdot x[n-k]}{\sum_{n=0}^{N-1} x[n]^2} \tag{10}$$

For periodic signals of T samples, the auto-correlation has maxima at multiple of T lags. We detect a peak if its value is strictly bigger than its nearest neighbors on both sides (discarding the first one at $k = 0$, which is always equal to 1 by definition of normalized autocorrelation). Because of the normalization process, peaks can appear for low-energy noises which have a sharp spectrum (an example of such noise is motor noise). Also, it cannot be used by itself to discriminate between the main speaker's speech and background voices. However, in this study, the motive to use the auto-correlation is that its peaks have low amplitude for transient noises, which are the most problematic noises when using HOS.

We then combine the auto-correlation peak m_X and the kurtosis of the LPC residual k_X to obtain the new feature f_X as follows:

$$f_X \triangleq m_X \cdot \log(1 + k_X) \tag{11}$$

We use the log-kurtosis to give a more Gaussian-like behavior of the feature, which will be useful for the classification, and also compensate for high values which may occur for strong voiced, close-talk frames. The enhanced kurtosis is shown in Fig. 4, where the improvement over the original kurtosis is apparent. The enhanced kurtosis still has low values for far-field speech, and is more stable in noisy frames. Another example, taken from the CENSREC-1 dataset, is shown on Fig. 5, where we can observe the same characteristics. In particular, we can observe that the enhanced kurtosis is more robust to 'onset' noises in the first 5 seconds (they correspond to step's noise from someone walking); the enhancement compared to simple kurtosis is also apparent on the 2nd and 5th speech section.

3. Classification with Online EM

3.1 EM and Unsupervised Learning

Some VAD algorithms rely on thresholding the feature, with a threshold whose value is generally computed from the estimated background noise level; the frame-level speech/non-speech classification is then converted to speech boundaries through a state-machine (also called hangover); for unsupervised VAD algorithm, it is the most straightforward way for classification (for example [16]). Here, we propose a scheme of unsupervised classification, but without relying directly on a threshold, which would require noise level estimation.

If we suppose that each class (speech and non-speech) has a probabilistic distribution, an optimal decision can be made by choosing the class which maximizes $p(\text{class}|X)$; this is maximum a posteriori classification (MAP). Here $p(X|\text{class})$ is modeled as a parametric density $p(.;\theta)$, and we try to estimate the parameter set θ. We use a parametric clustering method, as represented in Fig. 6. If we choose a Gaussian distribution for $p(.;\theta)$, the model is a simple binary mixture of Gaussians, where each component of the mixture represents a class (one for speech, one for non-speech). Expectation Maximization (EM) algorithm [17] estimates parametric models with latent variables based on the maximum likelihood principle. In this case, the latent variable is the class membership C. The EM algorithm is an iterative algorithm, and each iteration i requires two steps: the E step, where the conditional expectation of the log-likelihood for the complete data (X, C) given the observed

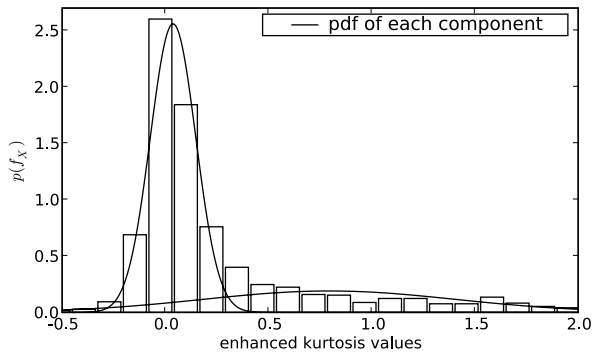

Fig. 6 Histogram of the enhanced kurtosis for the same extract as Fig. 4, and a simple 2 component mixture model estimated by standard EM algorithm (the latent variable has two states, either speech or non-speech).

data X is computed:

$$J_{\theta_{i-1}}(\theta) \triangleq \mathbb{E}[\log p(X, C; \theta)|X; \theta_{i-1}] \tag{12}$$

and the M step, where J is maximized with respect to θ to give a new estimated parameter set for step i

$$\theta_i \triangleq \arg\max_{\theta} J_{\theta_{i-1}}(\theta) \tag{13}$$

The key of EM algorithm is that the above scheme guarantees that the log-likelihood for the observed data X at point θ_i is higher than at the point θ_{i-1}, and that J has a closed form for a large class of models, including finite mixtures of exponential models, such as Gaussian mixture models. In the case of finite mixtures, and given T Independent and Identically Distributed (IID) observations $X = \{X_t\} = \{X_1, \ldots, X_T\}$, computing $J_{\theta_{i-1}}$ is reduced to computing $\zeta_t^i(c) \triangleq P(C_t = c|X_t; \theta_{i-1})$ for all t and c, where C_t is the latent variable corresponding to X_t and c is a choice of the membership. The new parameter θ_i can then be computed from ζ and statistics which depend directly on the data; in the case of Gaussian mixture models, they are X_t and X_t^2. In other words, (ζ_t^i, X_t, X_t^2) are Sufficient Statistics (SS) for θ_{i-1}.

For online classification, this cannot be used directly, because the E step requires the whole data set X. For example, in standard EM, the mean of the component c at step i $\mu^i(c)$ is given by:

$$\mu^i(c) = \sum_t \zeta_t^i(c) \cdot X_t / \sum_t \zeta_t^i(c) \tag{14}$$

Instead, we have to find a method to update μ_t (and other model parameters) at frame t from the observed data X_t and the previously estimated parameter θ_{t-1}. As noted in [18], there have been several approaches to this problem. We adopt the same scheme as proposed in [18] and [19]; the quantities of interest for the E step are replaced by a stochastic approximation, and the M step is kept the same. In online EM, we replace every statistic averaged by ζ by its stochastic approximation, which is updated every time a new frame is fed to the algorithm. For the mean, the statistic $\sum_t \zeta_t(c) \cdot X_t$ is replaced by $\hat{X}_t(c)$, which is updated at every new frame:

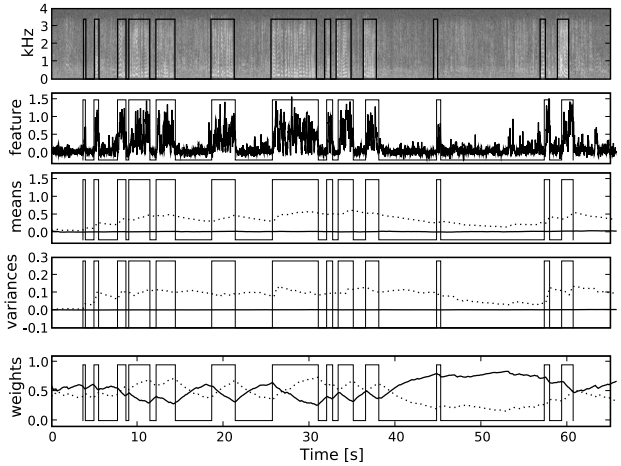

Fig. 7 Spectrogram of audio segment (1st), the enhanced kurtosis (2nd), means (3rd), variances (4th), and weights (5th) of the components as estimated by online EM (dashed red for speech, plain green for noise).

$$\hat{X}_t(c) = \hat{X}_{t-1}(c) + \gamma_t(\zeta_t(c) \cdot X_t - \hat{X}_{t-1}(c)) \tag{15}$$

Note that the i suffix for ζ is dropped, as the step and frame index are the same for online EM; also, the approximated statistics now depends on c through $\zeta_t(c)$. This approximation is then plugged into the mean estimator given by standard EM (equation 14):

$$\hat{\mu}_t(c) = \hat{X}_t(c)/\hat{\zeta}_t(c) \tag{16}$$

So, instead of averaging the statistics $\zeta_t \cdot X_t$ for all t at once, online EM successively averages between the current frame and the previous frame and the term $\zeta_t(c) \cdot X_t - \hat{X}_{t-1}(c)$ can be seen as the approximation error of the procedure [20]. The conditions on the sequence γ_t such that the above procedure converges are given in ([18], [19]); a more complete review of the theory behind this kind of procedures is given in [20].

As with the standard EM, we need to initialize the algorithm. Several strategies are possible. The simplest strategy is to initialize using random values; another solution, which we adopted, is based on a k-mean algorithm to initialize the means, with the weights for equi-probable cluster distribution. We then compute the variance for each cluster. The k-mean algorithm is run on a small subset of the data for each signal: the first second of the signal in our implementation. To give an idea about the online adaption of the EM, we plot in Fig. 7 the means of the two components. Although the two means have much the same value in the initial three seconds, where speech is not present, we can observe that the model effectively adapts itself to the signal when some close speech is present in the signal. Speech is always assumed to be the component with the higher mean.

Concerning the computing cost, standard EM can be split into three parts (for each iteration i): computing the responsibilities $\zeta_t^i(c)$, computing Sufficient Statistics (ζ_t^i, X_t, X_t^2), and updating the mixture model (Eq. (14) and its equivalent for the weights and covariance matrices). Only

the computation of SS is different in online EM (Eq. (15)), but the difference is negligible in terms of computation amount. The cost of running online EM on a given dataset is thus roughly the same as one iteration of EM. Since online EM is a recursive algorithm, and the features are computed frame by frame, it has a latency of one frame once it is initialized.

4. Experimental Evaluation

4.1 Evaluation Framework

We evaluate the proposed method, and compared it with several algorithms. We use two test sets, an in-house data which consists in close recording done during an open lab. For comparison purpose, we also use the CENSREC-1-C database [9]. For both test sets, we use a speech frame of 32 ms with an overlap of 16 ms (e.g. 256 samples at a sampling rate of 8 kHz, 50% overlap), and exactly the same algorithm (e.g. no tuning for the hangover system).

As an evaluation measure, we use frame-level classification errors, that is:

- False Rejection Rate (FRR), defined as:

$$\frac{\text{number of missed speech frames}}{\text{number of speech frames}}$$

- False Alarm Rate (FAR), defined as

$$\frac{\text{number of incorrectly detected speech frames}}{\text{number of non-speech frames}}$$

- Global Error Rate (GER), defined as

$$\frac{\text{number of missed + incorrectly detected frames}}{\text{number of frames}}$$

4.2 Evaluation on In-House Data

The data during the open lab were recorded in the following conditions: people were wearing a head-mounted device equipped with a head-set microphone. They were talking with other people in poster presentations. Audio data were recorded on another PC-like embedded device, carried by each person (the sound stream is converted to digital and recorded on the hard-disk of the device). The data contain several kinds of noises (air conditioning, other people, cars running on the street, etc.). The test set contains around 45 minutes of audio data, split into around 30 files of the same length. They are different in speakers, gender, language (mainly Japanese, but also English), sparsity and SNR (between 10 dB and 25 dB). For each file of the test set, the proposed online EM algorithm was initialized with the first second data. The goal of this evaluation is to assess whether the proposed method can adapt to various SNR and sparsity conditions. The ratio of speech frames in this data ranges from 10 to 90%, with 33% of speech on average. We compare the proposed method with methods by replacing the enhanced kurtosis with different features: energy and kurtosis. The results are summarized in Table 1. The enhanced kurtosis achieves a significantly better FAR and GER, while being slightly worse FRR than energy. It also shows that kurtosis alone is not effective.

Table 1 Frame error rates for the proposed algorithm (1st row), using online EM on energy (2nd), and online EM on kurtosis only (3rd).

	FAR	FRR	GER
Proposed method	7.8 %	13.0 %	9.5 %
Using energy	15.8 %	10.6 %	13.3 %
Using kurtosis only	19.0 %	13.8 %	16.3 %

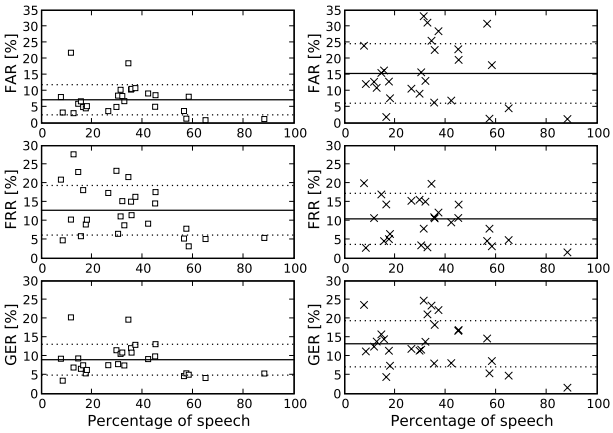

Fig. 8 Results of the proposed VAD algorithm (left) in function of the speech/non speech ratio, in comparison with energy-based method (right). The dashed lines show the standard deviation of each criterion, and solid line the mean along all files from the database.

Table 2 Comparison between online EM and standard EM.

	FAR	FRR	GER
Online EM (proposed method)	7.8 %	13.0 %	9.5 %
Standard EM	8.0 %	12.0 %	9.5 %

To get a more precise idea of the robustness of the proposed classification scheme against sparsity, we give in Fig. 8 the frame error rates with respect to the sparsity (i.e. the ratio speech/non-speech) for each file. Whereas the proposed method and energy-based method perform similarly for signals where speech is dominant, the FAR significantly increases for the energy-based algorithm for sparser signals. Also, the dashed line, representing the standard deviation of the error rates on the whole data, shows that the proposed method is less sensitive to sparsity variation, thus giving more stable results.

We also compared the effectiveness of the online EM to the standard EM algorithm. Both used the enhanced kurtosis as a feature. The results are summarized in Table 2. Online EM is found to give similar performance to the offline EM.

4.3 Evaluation on CENSREC-1 Database

For comparison purpose, we also tested the proposed

Table 3 Frame error rates for the proposed method on close recordings of CENSREC-1-C.

Noise condition	FAR	FRR	GER
Restaurant, high SNR	10.3 %	6.9 %	9.1 %
Restaurant, low SNR	9.9 %	8.5 %	9.3 %
Street, high SNR	7.2 %	13.8 %	9.7 %
Street, low SNR	8.7 %	13.4 %	10.7 %
Average	9.0 %	10.6 %	10.7 %

Table 4 Frame error rates for the energy-based method on close recordings of CENSREC-1-C.

Noise condition	FAR	FRR	GER
Restaurant, high SNR	10.1 %	20.6 %	13.7 %
Restaurant, low SNR	10.4 %	26.8 %	16.7 %
Street, high SNR	12.7 %	29.8 %	19.7 %
Street, low SNR	12.2 %	28.4 %	18.4 %
Average	11.2 %	26.0 %	16.7 %

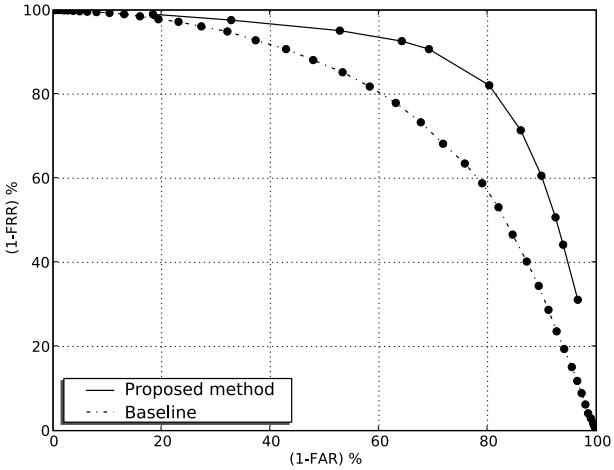

Fig. 9 ROC of baseline vs. proposed method, in remote recordings condition (low and high SNR averaged).

method on a public database, CENSREC-1 [9]. This database consists of noisy contiguous digit utterances in Japanese. The recordings were done in two kinds of noisy environments (street and restaurant), and high (SNR > 10 dB) and low SNR ($-5 \leq$ SNR ≤ 10 dB). For each of these conditions, close and remote recordings were available [9]. The algorithm used is exactly the same as previous section, and the online EM was initialized with the first second for every file of the database.

First, the results for close recordings of several noise conditions are given in Table 3. Each case has a total length of 30 minutes approximately. From Table 3, it is observed that the figures are much the same for low and high SNR, both for restaurant and street environments in the close recording case. The noise type seems more significant than the SNR condition. We also compared the proposed method with a method that uses energy instead of the enhanced kurtosis as a feature. The results are given in Table 4, which confirms that the enhanced kurtosis gives better performances than energy.

Finally, as suggested by the designers of CENSREC, we show a comparison with the baseline along with its ROC for remote recordings, although our method is intended for the close-talking condition. The ROC is computed for the average between low and high SNR, and is plotted in Fig. 9. The baseline uses a simple energy-based algorithm [9]. It should be noted that this baseline method is an offline algorithm, and the classification is done a posteriori knowing the whole signal. This gives the baseline an advantage, however, our algorithm outperforms the baseline.

5. Conclusion

A new online method for VAD has been proposed. The method uses HOS enhanced by autocorrelation to improve the robustness against non-Gaussian noises. The use of HOS for discriminating against far-field speech, which is a significant problem in human-to-human situations, has also been investigated. The classification is done online by an online clustering method based on EM algorithm. The Expectation step is replaced by a recursive stochastic approximation to enable the algorithm to change its state for each new frame, thus providing a noise estimation without requiring a separate scheme for SNR estimation.

The method has been evaluated on two test sets, and compared with an energy-based method and standard HOS. On the in-house data, the robustness of the algorithm against sparsity compared to the energy-based algorithm has been confirmed. On CENSREC-1-C, the method has been confirmed to be robust with respect to SNR, while significantly outperforming the energy-based algorithm for close recordings. For remote recordings, the method outperformed the baseline, too.

This scheme can be further improved. Online EM estimation is not so reliable until some speech data are available. By explicitly considering prior knowledge and comparing models, Bayesian treatment of online EM may solve this problem and provide better accuracy. This issue will be addressed in the future work.

References

[1] L.R. Rabiner and B.H. Juang, Fundamentals of speech recognition, Prentice Hall, 1993.
[2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM Spine evaluation system," ICASSP, 2002.
[3] G. Lathoud and I. McCowan, "Location based speaker segmentation," Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), 2003.
[4] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 2001.
[5] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," IEEE Trans. Speech Audio Process., vol.9, no.3, pp.217–231, 2001.
[6] K. Li, M.S.S. Swamy, and M.O. Ahmad, "An improved voice activity detection using high order statistics," IEEE Trans. Speech Audio Process., vol.13, no.5, pp.965–974, 2005.
[7] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based

on a sinusoidal representation," IEEE Trans. Acoust. Speech Signal Process., vol.34, no.4, pp.744–754, 1986.
[8] "ITU-T: A silence compression scheme for G.729 optimized for terminals conforming to recommendation v.70, annex B.," Nov. 1996.
[9] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, K. Yamamoto, T. Takiguchi, and S. Nakamura, "CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment," IPSJ SIG Technical Report, 2006.
[10] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," IEEE Trans. Speech Audio Process., vol.7, no.3, pp.333–338, May 1999.
[11] J.K. Shah, A.N. Iyer, B.Y. Smolenski, and R.E. Yantorno, "Robust voiced - unvoiced classification using novel features and Gaussian mixture model," IEEE ICASSP'04, 2004.
[12] Q. Li, J. Zheng, Q. Zhou, and C.H. Lee, "A robust, real-time endpoint detector with energy normalization for ASR in adverse environments," ICASSP01, IEEE, 2001.
[13] L.T. DeCarlo, "On the meaning and use of kurtosis," Psychological Method, vol.2, pp.292–303, 1997.
[14] H.M. Finucan, "A note on kurtosis," J. Royal Statistical Society. Series B (Methodological), vol.26, pp.111–112, 1964.
[15] S. Basu, "A linked-HMM model for robust voicing and speech detection," IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03), 2003.
[16] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03), pp.432–435, 2003.
[17] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statistical Society. Series B (Methodological), vol.39, pp.1–38, 1977.
[18] O. Cappe, M. Charbit, and E. Moulines, "Recursive EM algorithm with applications to DOA estimation," Proc. 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006, 2006.
[19] M. Sato and S. Ishii, "On-line EM algorithm for the normalized Gaussian network," Neural Comput., vol.12, pp.407–432, 2000.
[20] H.J. Kushner and G.G. Yin, Stochastic approximation algorithms and applications, Springer-Verlag, 1997.

David Cournapeau received a MSc. degree from UPMC Paris (Universite Pierre Marie Curie), specializing in acoustics and music signal processing in 2003, and a M.E from Telecom Paristech in 2004. In 2006, he enrolled in the PhD program from Kyoto University, under the supervision of Prof. Kawahara. His research interests include digital signal processing, speech and music signal processing, signal representation, statistical signal processing and pattern recognition.

Tatsuya Kawahara received the B.E. degree in 1987, the M.E. degree in 1989, and the Ph.D. degree in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Adjunct Professor in the School of Informatics, Kyoto University. He is also an Invited Researcher at ATR Spoken Language Communication Research Laboratories. He has published more than 150 technical papers covering speech recognition, spoken language processing, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/). Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of the IEEE SPS Speech Technical Committee. He was a general co-chair of IEEE Automatic Speech Recognition & Understanding workshop (ASRU-2007).