Evaluation of Real-time Voice Activity Detection based on High Order Statistics

David Cournapeau and Tatsuya Kawahara

Graduate School of Informatics, Kyoto University
ACCMS South. Bldg., Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

Abstract

We have proposed a method for real-time, unsupervised voice activity detection (VAD). In this paper, problems of feature selection and classification scheme are addressed. The feature is based on High Order Statistics (HOS) to discriminate close and far-field talk, enhanced by a feature derived from the normalized autocorrelation. Comparative effectiveness on several HOS is shown. The classification is done in real-time with a recursive, online EM algorithm. The algorithm is evaluated on the CENSREC-1-C database, which is used for VAD evaluation for automatic speech recognition (ASR) [1], and the proposed method is confirmed to significantly outperform the baseline energy-based method.

Index Terms: Voice activity detection, online EM, high order statistics

1. Introduction

VAD can be simply described as detecting speech boundaries from audio signal. It is used in most speech processing tasks as a pre-processing step. For example, the GSM 729 standard defines two VAD modules used for variable bit speech coding; VAD robust to noise is a critical step for ASR in noisy environments. Recently, it has also an important role in the task of multi-modal human-to-human interactions, such as meetings [2]. In the latter situations, the problem is complicated by the fact that in the case of natural speech, it is difficult to make assumptions, generally made for VAD in ASR contexts, that most of the signal contains speech. This means the classification scheme has to somewhat adapt to sparsity of the speech. Also, as several people are involved in those situations, it is necessary to be able to discriminate between speakers. One solution for this problem is to use microphone arrays, for example in [3]. If using wearable microphones is possible, the problem is reduced to find a feature good enough to discriminate between close-talk and far-field speech: this is the approach we have taken. Note that even with close-talking microphones, it is not easy to eliminate background speech with a simple energy-based method in sparsely uttered situations as in meetings.

We have already presented the basic concept of the approach and a preliminary test with an in-house data in [4]; the global scheme is depicted in Figure 1. In this paper, we give thorough description of the method and investigate how the high order statistics of order 4 (a.k.a. kurtosis) performs better than statistics of order 3 (a.k.a. skewness) : this is developed in section 2. The on-line classification scheme is explained in section 3. Finally, the algorithm is evaluated on a publicly available database designed for VAD evaluation in section 4.

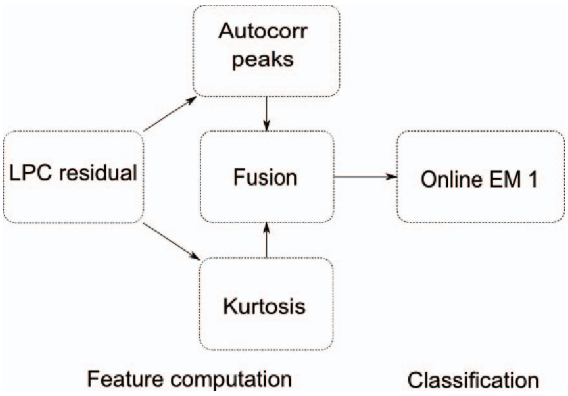

Figure 1: Overview of proposed method

2. Proposed feature

2.1. Kurtosis as discriminative feature against far-field speech

Many features have been suggested for VAD: energy, autocorrelation, cepstrum peaks ([5]) and MFCC ([6]). The idea is that the underlying distribution of the feature is different for speech and non-speech parts, and that those differences can be easily detected. We are also interested in a feature which is robust against far-field speech. Also, for real-time speech detection, as noted for example in [7], normalization of the feature is critical to avoid classification errors. We focus on normalized features, that is features which are independent on the signal energy.

To discriminate against far-field speech, we use several properties of our setting (close microphone); first, obviously, because energy received by the microphone is dependent on the distance between the source and the microphone, far-field speech and close-talk speech have different energy; but more interestingly, close-talk microphones being directional (that is their sensitivity is not uniform across all directions), they have a so called boost-effect. This boost effect increases the amplitude in the low spectrum for audio sources which are really close to the microphone; all directive microphones have this effect, which is often used by sound engineers for music recording. This effect is also sometimes called proximity effect.

Figure 2 shows an example of this boost effect on the LPC residual for close-talk and far-field speech. In a simple source-filter model of speech production, the LPC residual can be seen as the glottal excitation; as such, the impulse train produced by the vibration of the vocal cords can be observed for periodic speech frames. In both cases, the pulses corresponding to the opening of the vocal cords can be seen as well as their period-

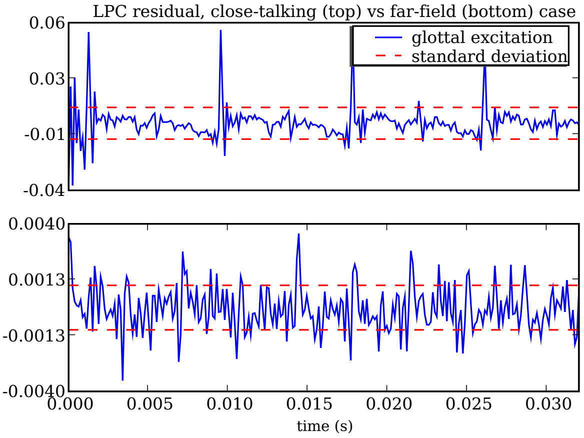

Figure 2: LPC residual of close-talk speech (top) and far-field speech (above)

icity, but because of the boost effect, the pulses have a clearly higher energy amplitude in close-talking speech. On the contrary, for far-field speech, those pulses are weaker; also, because of environmental noises, the rest of the signal is more likely to take higher values compared to close-talk.

If we consider the distribution of the amplitude of the LPC residual for close-talk speech, most samples will be around 0 inside the range $[-\sigma, \sigma]$, where σ is the standard deviation. On the contrary, for far field speech, more samples will have high values. In other words, because of the proximity effect, pulse amplitudes are emphasized, and for far-field speech, they are comparatively de-emphasized. Statistically, the distribution of the LPC residual in close-talk has a high peak around the mean, and a fat tail (relatively high number of extreme values), whereas the distribution for far-field speech has a fatter midrange, that is many values around σ. From this point of view, discriminating between close-talk and far-field speech can be done by discriminating fat-tailed distributed LPC residual against fat-midrange distributed LPC residuals.

To measure such a difference, we use High Order Statistics (HOS) derived as cumulants of the LPC residual. Cumulants of a random signal X are defined by the cumulant generating function, defined as the following:

$$\begin{aligned} \log \Phi(t) &= \log \mathbb{E}[e^{tX}] \\ &= \sum_{n=0}^{\infty} \kappa_n \frac{t^n}{n!} \end{aligned} \qquad (1)$$

The cumulant generating function is the log of the moment generating function, and cumulants of order n, κ_n, are to the cumulant generating function what the moments of order n are to the moment generating function. Kurtosis is defined as the cumulant of order 4. Another commonly special case is the cumulant of order 3, called skewness. An explicit relationship between skewness, kurtosis and the number of harmonics in the LPC residual was given in [8]. Kurtosis has high values for random signals which are heavy tailed, and has low values for random signals which have values in the midrange of their distribution (generally located around σ and $-\sigma$ for centered signals); see for example [9] for a proof.

Table 1: Comparison between kurtosis and skewness

	FAR	FRR	GER
Proposed algorithm (kurtosis)	7.8 %	13.0 %	9.5 %
Proposed algorithm (skewness)	8.2 %	14.6 %	10.6 %

2.2. Comparison between kurtosis and skewness

We use the normalized version of the excess kurtosis; excess kurtosis is defined such that the kurtosis of a normally distributed signal is 0, and the normalization factor is equal to $1/\sigma^4$. As an example, for the signals represented on Figure 2, the kurtosis is 15.4 for the close-talk case (top), and 0.4 for the fair-field case (bottom). If we remove a few samples corresponding to the pulses in the top LPC residual, the kurtosis quickly drops to small values.

We experimentally compared the difference between kurtosis and skewness. We ran the VAD method with the exact same algorithm except that kurtosis was replaced by skewness (again, normalized). The test set is the in-house data used in [4], which consists of 45-minute conversation by a number of people wearing head-set microphones. The results are given in Table 1; FAR is the False Alarm Rate (ratio of non-speech detected as speech), FRR the False Rejection Rate (ratio of speech frames not classified as speech) and GER, the Global Error Rate, (ratio of all misclassified frames against the total number of frames). In this result, kurtosis is confirmed to be more effective than skewness.

2.3. Enhancing kurtosis

As we already noted in [4], and as noted in [8], kurtosis alone cannot be used, because it is really sensitive to some kind of noises, particularly transient noises, that is noises which are well localized in time and have really high energy (e.g. noises corresponding to physical contact to the microphone). To compensate those problems, we combine the kurtosis with a feature robust against transient noises: the second main peak of normalized autocorrelation (the first peak, at lag 0, is always 1 by definition of normalized autocorrelation). We compute the autocorrelation of the LPC residual. The LPC residual is supposed to contain most of the pitch information, and this makes the peaks sharper. Thus, we combine it with log kurtosis to make its behaviour more Gaussian:

$$f = m \cdot \log(1 + \kappa) \qquad (2)$$

where m is the amplitude of the main peak of the normalized autocorrelation and κ the log-kurtosis. An example on Figure 3 shows that the enhanced kurtosis has lower values than kurtosis alone for transient noises (such as the parts around second 10).

3. Classification scheme

3.1. Online EM

Some VAD algorithms rely on state-machine like scheme to classify speech and non-speech states; for unsupervised VAD algorithm, it is the most straightforward way for classification (for example [10]). 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

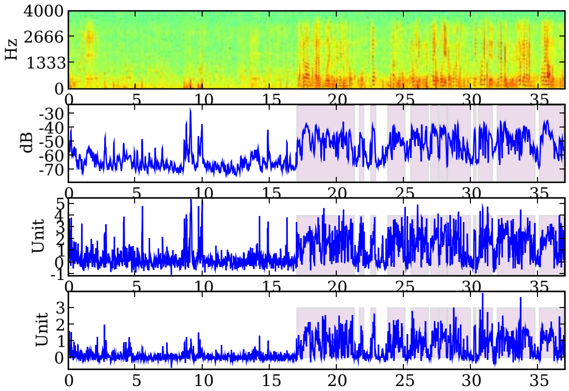

Figure 3: Audio example with spectrogram (top), comparing energy (middle top) to log-kurtosis (middle bottom) and enhanced kurtosis (bottom). The red boxes correspond to speech sections.

each class (speech and non-speech) has a probabilistic distribution, an optimal threshold can be determined in a Bayesian context as the class which maximizes $p(\text{class}|x)$ where x is the observed data (e.g. enhanced cumulant values); this is a maximum a posteriori classification (MAP). The problem of course is to be able to compute $p(\text{class}|x)$.

In a parametric context, $p(x|\text{class})$ is modeled as a parametric density $p(.;\theta)$, and we try to estimate the parameters θ. Expectation Maximization algorithm [11] is a well known algorithm to estimate parametric models with so called hidden feature, also called latent variables; in our case, the latent variable is the class membership. If we choose a Gaussian distribution for $p(.;\theta)$, the model estimated by the EM algorithm is a simple binary mixture of Gaussian, where each component of the mixture represents one class (one for speech, one for non speech). The EM algorithm is an iterative algorithm, and each iteration i requires two steps, an Expectation step, where the latent variable distribution is estimated using the parameters of the former iteration θ_{i-1}, and a Maximization step, where the sufficient statistics of the model are estimated from the latent variable distribution and used to update the parameters of the model (C is the random variable representing the class membership):

1. E step: estimate $\zeta_c^i \triangleq p(C = c|x, \theta_{i-1})$
2. M step:
   (a) M 1: estimate sufficient statistics SS_i from ζ_c^i
   (b) M 2: estimate θ_i from sufficient statistics SS_i

For real-time classification, this cannot be used directly, because both E step and M step (part1) requires the whole data set. As noted in [12], there have been several approaches to solve this problem. One method, described in [13], consists in recursive approximation of the model's parameters, that is for a new observation x_n,

$$\theta_n = \theta_{n-1} + \gamma_n {I_f}^{-1}(\theta_{n-1})U(x_n; \theta_{n-1}) \tag{3}$$

where γ is a non-decreasing scalar sequence, ${I_f}^{-1}(\theta_{n-1})$ the Fisher Information Matrix of one complete observation (x_n, c_n), and U a score function defined by $U \triangleq \nabla_\theta \log p(x;\theta)$. A more EM-like approach, where both E and M steps are still used, has been proposed more recently ([14], [12]); the E step is replaced by a stochastic approximation, and the M step is kept the same:

$$SS_n = SS_{n-1} + \gamma_n(SS(x_n;\theta_{n-1}) - SS_{n-1}) \tag{4}$$

The conditions on the sequence γ such that the above procedure converges are given in ([14], [12]); a more complete review of the theory behind this kind of procedures is given in [15]. We used the latter approach in our implementation. To give an idea about the online adaption of the EM, we plot in Figure 4 the means of each component; we can observe that the state of the model effectively adapts itself to the signal after a few frames.

We compared the effectiveness of the online EM to the standard EM algorithm. Both used the enhanced kurtosis as a feature. They were tested on the test-set as in section 2.2. Although online EM is slighthly worse than offline for FRR (13.0 % vs 12.0 %), they got comparable FAR (7.8% vs 8.0 %) and the same GER (9.5 %). Online EM is found to give similar results to the offline EM.

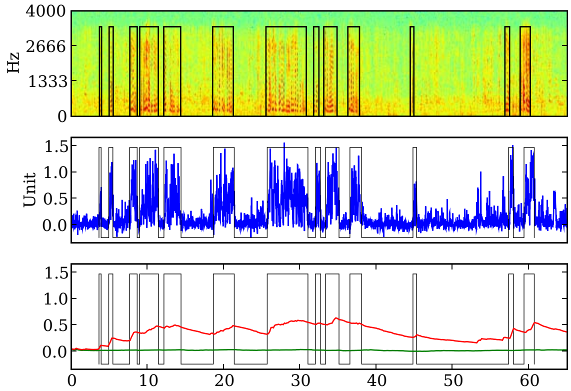

Figure 4: Spectrogram of audio segment (top), the enhanced kurtosis (middle) and means of components estimated by online EM (red for speech, green for noise)

4. Evaluation

We applied our method to the CENSREC-1 database for more comprehensive evaluation. This database consists of noisy continuous digit utterances in Japanese; the recordings were done in two kinds of noisy environments of street and restaurant, and high (SNR $>$ 10 dB) and low SNR (-5 $\leq$ SNR $\leq$ 10 dB). For each of these situations, close and remote recordings were available [1]; we used a speech frame of 32 ms with an overlap of 16 ms (eg 256 samples at a sampling rate of 8 khz, 50 % overlap).

First, the results for close recordings are given in Table 2; each case has a total length of 30 minutes approximately. From Table 2, it is observed that the results are much the same for low/high SNR, both for restaurant and street environments in the close recording case. The noise type seems more significant than the SNR condition. For comparison purposes, we also compare the proposed algorithm with an algorithm which still uses online EM for classification, but uses energy instead of the enhanced kurtosis as a feature. The results averaged over different SNR and two types of environments are given in Table 3. This confirms the effectiveness of the proposed algorithm.

Table 2: Frame error rates for the proposed algorithm on close recordings of CENSREC-1-C

Close Case	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 %

Table 3: Frame error rates for the proposed algorithm compared to energy-based method

Close Case	FAR	FRR	GER
Proposed algorithm	9.0 %	10.6 %	9.8 %
Energy-based	11.2 %	26.0 %	16.7 %

Finally, we show a comparison with the baseline along with its ROC for remote recordings, although our method is designed for close-talking. The ROC is computed for the average between low and high SNR, and is plotted in Figure 5. The baseline uses a simple energy based algorithm [1]. It should be noted that this baseline algorithm 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.

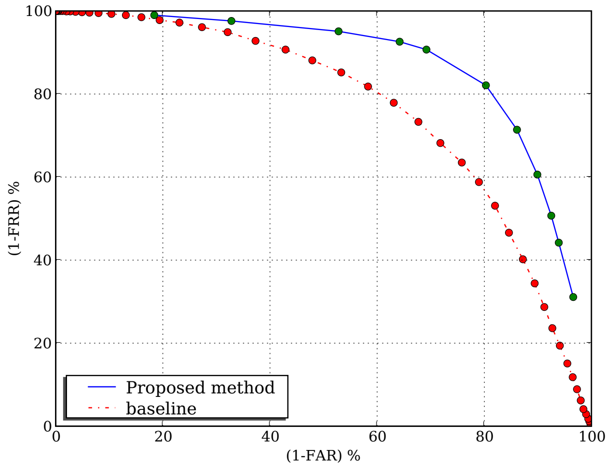

Figure 5: ROC of baseline vs proposed algorithm, in remote recordings conditions (low and high SNR averaged)

5. Conclusion

An unsupervised VAD algorithm has been presented and evaluated on a publicly available database. The proposed method outperforms the baseline by a significant margin. As the feature is not computationaly intensive and the classification does not have a high latency, the method is suitable for real-time VAD.

6. References

[1] 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, Tech. Rep., 2006.

[2] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, “Automatic Analysis of Multimodal Group Actions in Meetings,” in *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 2005.

[3] G. Lathoud and I. McCowan, “Location Based Speaker Segmentation,” in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, 2003. [Online]. Available: ftp://ftp.idiap.ch/pub/lathoud/segmentation.pdf

[4] D. Cournapeau, T. Kawahara, K. Mase, and T. Toriyama, “Voice activity detection based on enhanced cumulant of lpc residual and on-line em algorithm,” in *Proceedings of Interspeech06*, 2006.

[5] S. Ahmadi and A. S. Spanias, “Cepstrum-Based Pitch Detection Using a New Statistical V/UV classification algorithm,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, May 1999.

[6] J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, “Robust voiced - unvoiced classification usgin novel features and gaussian mixture model,” in *IEEE ICASSP’04*, 2004.

[7] Q. Li, J. Zheng, Q. Zhou, and C.-H. Lee, “A Robust, Real-Time Endpoint Detector with Energy Normalization for ASR in Adverse Environments,” in *ICASSP01*. IEEE, 2001.

[8] E. Nemer, R. Goubran, and S. 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.

[9] H. M. Finucan, “A note on kurtosis,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, pp. 111–112, 1964.

[10] I. Shafran and R. Rose, “Robust speech detection and segmentation for real-time ASR applications,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’03)*, vol. 1, 2003, pp. 432–435.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[12] O. Cappe, M. Charbit, and E. Moulines, “Recursive em algorithm with applications to doa estimation,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006.

[13] D. M. Terrington, “Recursive parameter estimation using incomplete data,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, pp. 257–267, 1984.

[14] M. Sato and S. Ishii, “On-line EM algorithm for the normalized gaussian network,” *Neural Computation*, vol. 12, pp. 407–432, 2000.

[15] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*. Springer-Verlag, 1997.