Improving Speaker Verification by Periodicity Based Voice Activity Detection

Ville Hautamäki, Marko Tuononen, Tuija Niemi-Laitinen and Pasi Fränti

Speech and Image Processing Unit, Department of Computer Science and Statistics,
University of Joensuu, Finland

Abstract

We propose an improvement to a realtime speaker recognition system, where matching is performed instantly while subject is still speaking to the microphone. We present a voice activity detection (VAD) based on realtime periodicity analysis. Performance of the proposed method is compared against two existing methods: a realtime method based on long-term spectral divergence (LTSD) and a simple energy based method, which needs two passes on the data. The periodicity-based method clearly outperforms the other realtime method (LTSD), and performs comparably with the energy-based method when applied for NIST 2001 and 2006 speaker recognition evaluation corpora. The method is also tested for segmenting surveillance recordings, voice dialog and forensic applications.

1. Introduction

Voice Activity Detection (VAD) [1] aims at classifying a given sound frame as a speech or non-speech as. It is often used as a front-end component in voice-based applications such as automatic speech recognition, speech enhancement [2] and voice biometric [3]. A common property of these applications is that only human sounds (typically only speech) are of interest to the system, and it should therefore be separated from the background. Although VAD is widely needed and reasonably well-defined, existing solutions do not satisfy all the user requirements. The methods either must be trained for the particular application and conditions, or they can be highly unreliable when the conditions change. Demands for working solutions, however, are frequently requested by practitioners.

Traditionally, VAD research has been driven by telecommunications and voice-coding [1,2,4] applications, in which VAD has to operate with as small delay as possible and all the speech frames should be detected. Typically, these voice activity detectors work by modeling the noise signal statistics. Initial noise estimates are usually obtained from the beginning of the signal, which is updated as VAD makes non-speech decisions.

Even though realtime VAD for telecommunications application is maybe the most common, other applications also exist. For those applications, design criteria are usually different than telecom VAD's. In speech segmentation application, realtime operation is not as important as the goal is to process the input speech file (usually hours or even days long) in a background process and then the output will be used for the retrieval or the diarization tasks [5]. Segmentation of the forensic wiretapping is especially difficult task. In those recordings, no close talking microphone is used and the noise does not possess properties assumed in the telecom VAD's (long-term stationarity) [6].

In this study, our primary goal is to use VAD as a preprocessor for the realtime speaker verification. It is clear that including non-speech frames in the modeling process would bias the resulting model, especially if the number of non-speech frame is significant. It is also known that not all speech frames have equal discriminative power as it is well known that voiced phonemes are more discriminative than unvoiced phonemes [7], and therefore, it can be beneficial to drop out unvoiced frames to increase recognition. This is in contrast to telecom or speech recognition application where speech should be accurately detected.

In realtime speaker verification [8], we need to start processing speech frames with as small delay as possible. New recorded speech frame is pushed to signal processing subsystem, and to scoring subsystem while subject keeps speaking to the microphone. Not only is this useful feature in the standard applications, but it is essential in low power mobile phones, where realtime processing is only option [9]. Good results in speaker verification can be achieved by a simple energy based VAD, but the method needs analyze the whole utterance before it can start to make speech/non-speech decisions [10].

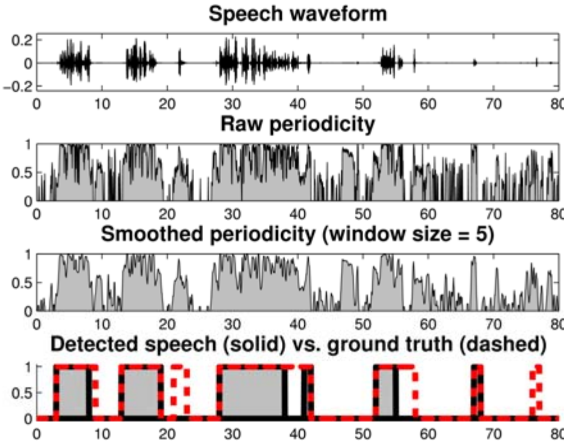

Fig. 1. Example of the proposed system.

Our proposed system is based on the detecting the periodicity of the given frame. In contrast to the telecom voice activity detectors we do not model noise, but we base our decisions on the feature that is known to be short term stationary. Periodicity information has been used before in other speech technology applications, for example speech logging [11]. Our goal is to enhance the speaker verification performance.

2. VAD in Speaker Verification

In the speaker verification, the unknown speech sample is introduced to the system accompanied by a claim. The problem is considered to be a testing hypothesis that claimed person is the author of the utterance. In practice, we calculate *mel frequency cepstral coefficents* MFCC for each frame, which are then fed into either Gaussian mixture modeling algorithm (in training) or to the scoring algorithm (in testing).

We used the adapted Gaussian mixture model [3], in which the target speaker models are trained by adjusting the parameters of a universal background model (UBM) towards the speaker's training data. We used a diagonal covariance matrix GMM, which is trained separately for females and males using expectation maximization (EM) algorithm. EM is initialized by a variant of split algorithm and fine-tuned by K-means [12]. In testing phase, we used the average log-likelihood ratio to calculate a match score. Gender information was used to create separate UBM models for male and female speakers as defined in NIST corpus.

2.1. Periodicity based VAD

Given an accurate periodicity detector, we could voiced phonemes from the speech signal. In practice, we make speech/non-speech decisions on the frame level, so we need to consider how periodic a given frame is. We assume that noise is aperiodic, and therefore (in principle) we can detect speech parts of the utterance by thresholding the periodicity estimate. Although this is not generally true, it is expected that a significant part of speech can be separated from background noise in this way.

We used the periodicity detector defined in the YIN pitch estimation algorithm [13], which is an autocorrelation based algorithm. Autocorrelation function matches lagged version of the same frame on top of the frame itself. When the lagged version matches well to the current frame, we have found a period in the signal, where a pitch estimate can be extracted. Unfortunately, autocorrelation method itself is not very robust in estimation of pitch period because of the influence of the formants, especially the first formant. A complicated post-processing system is implemented in the YIN algorithm as follows.

First, the idea of autocorrelation is realized as the squared difference function, where function value is calculated at different lags:

$$d(\tau)=\sum_{j=1}^{W}(x_j-x_{j+\tau})^2 \,. \tag{1}$$

The minima of $d(\tau)$ will give estimate of aperiodicity and the corresponding τ can be turned into F0 estimate. Next, quadratic function is fitted to each local minima of the $d(\tau)$'s, and the result of the interpolation is then used as the aperiodicity estimate. Each calculated $d(\tau)$ is normalized by cumulative sum of the previous lags. Before normalization we initialize the $d(0)=1$. Cumulative normalization is performed as

$$d'(\tau)=\frac{d(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau}d(j)}\,. \tag{2}$$

$d'(\tau)$ values are evaluated from smallest τ to largest, while applying the user selected threshold to $d'(\tau)$'s. Such $d'(\tau)$ value is selected that is local minimum and smaller than the threshold.

The $d'(\tau)$ now represents the aperiodicity estimate of the frame, where 0 meaning periodic frame and 1 totally aperiodic frame. We obtain the final periodicity estimate p_i for a frame *i* by $p_i=1-d'(\tau)$. As the estimate is very noisy, we need to smooth it by averaging by the sliding window. See Fig. 1 for an example, where upper part contains speech waveform from NIST 2005 evaluation corpus, middle part its raw periodicity and lower part smoothed periodicity by window of length five.

2.2. Energy-based VAD

A simple energy-based approach is often used for voice activity detection. The approach in [10] was used in NIST 2006 evaluation: it measures the intra frame energy by calculating standard deviation of the frame and compares it to the expected SNR value (30 dB by default). A segment is defined as speech segment if it is within this marginal, and if it exceeds a given minimum energy level (-55 dB). This can be easily checked and it has found to work well for its purpose in voice biometric.

In more detail simple energy-based works as follows. First we calculate the logarithm compressed standard deviation of each frame:

$$S_i=20\log_{10}\sqrt{\frac{1}{N-1}\sum_{j=1}^{W}\left(x_j^i-\bar{x}^i\right)^2}\,, \tag{3}$$

where x_j^i is the *j*th sample value of *i*th frame and $\bar{x}^i$ is the sample average of *i*th frame. Then frame *i* is detected as speech if $S_i>(\max_j S_j-T)$ and $S_i>-55$.

2.3. Long Term Spectral Divergence VAD

This approach measures long term spectral divergence (LTSD) between speech and noise. It formulates the speech/non-speech decision rule by comparing the long term spectral envelope to the average noise spectrum. It is also reasonably simple to implement and rather efficient compared to other VAD methods according to the experiments made in [2].

The algorithm operates on a window of $2M+1$ frames, where decision is made on a center frame. $\mathrm{LTSE}_M(k,l)$ is a short term spectral envelope, and it is calculated as follows:

$$\mathrm{LTSE}_M(k,l)=\max_{j=-M}^{M}X(k,l+j)\,, \tag{4}$$

where $X(k,l)$ is the magnitude of the *k*th spectral bin of the *l*th frame. Final decision is performed by thresholding $\mathrm{LTSD}_M(l)$ function:

$$\mathrm{LTSD}_M(l) = 10\log_{10}\left(\frac{1}{\mathrm{NS}}\sum_{k=0}^{\mathrm{NS}-1}\frac{\mathrm{LTSE}^2(k,l)}{\mathrm{N}^2(k)}\right), \tag{5}$$

where $\mathrm{N}(k)$ is the estimated noise magnitude spectrum of the *k*th bin and NS is the number of spectral bins. Noise estimate is updated on the fly when a negative decision has been made. Even though this strategy can account for long term changes in the noise characteristics, it can also bias the estimate and therefore, decrease the detection performance.

3. Speaker Verification Experiments

Feature extraction parameters for speaker verification experiments were set as follows. For the MFCC features, we use the coefficients 1-12, computed from a 27-channel mel-filterbank. The frame length is set to 30 milliseconds, with 33 % overlap. The MFCC vector is appended with its delta and double-delta coefficients at the frame level, yielding 36-dimensional data. Each feature is normalized by subtracting the mean and dividing by the standard deviation estimated from the file.

Voice activity detection parameters are tuned on the NIST 2001 evaluation corpora, in such a way that using one set of VAD parameters all audio files (trial, model and UBM) are segmented during the feature extraction. For periodicity VAD was tried window size parameters 3, 5 and 7. Threshold we varied between 0.60 and 0.90. LTSD and Energy based VAD's were compared by varying their decision thresholds.

In Fig. 2, periodicity VAD parameters are tuned with respect to NIST 2001 evaluation protocol. We can notice that smoothing window size should be small, and the threshold should be relatively low (less than 0.70), which leads to less frames being dropped out. We also notice that 512 size GMM outperforms 64 size model systematically, the difference is only 1% unit.

In Fig. 3, tuning results for energy and LTSD VADs are shown. We notice that LTSD has a sharp minimum around threshold value 40. Energy based VAD has a little bit larger minimum around threshold values of 34-39. Energy based VAD is cleary more stable with respect to its threshold than the LTSD VAD.

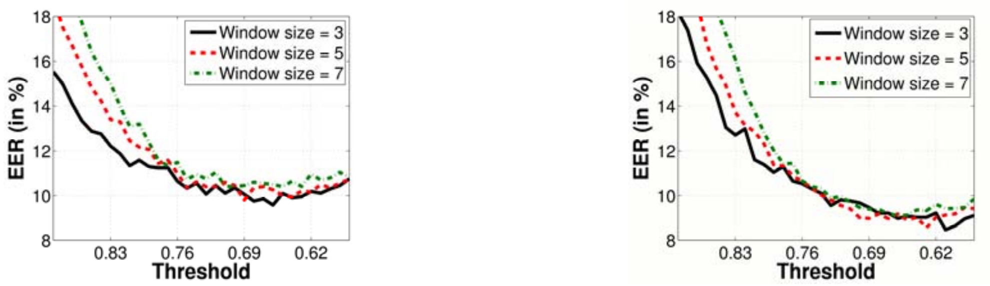

Fig. 2. NIST 2001 results for the proposed method with the model size 64 (left) and the 512 (right).

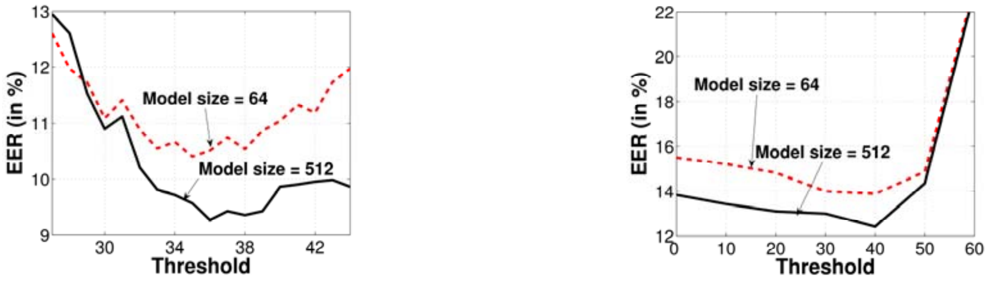

Fig. 3. NIST 2001 tuning results for the Energy (left) and LTSD (right) VAD's.

DET plots of the best tuning results are shown in Fig. 4. It is clear that all methods provide significantly better results than not using VAD at all. Both energy based and periodicity VADs clearly outperforms LTSD for NIST 2001. Periodicity VAD is almost one percent unit better than energy based VAD.

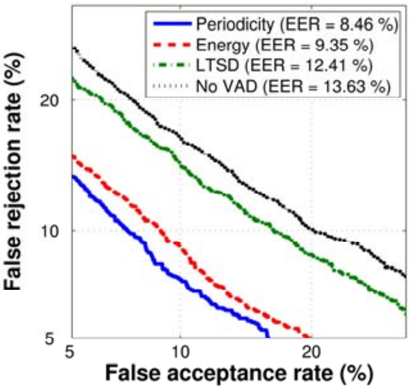

Fig. 4. NIST 2001 best results for the model size 512

Table 1. Summary of speaker verification results (% EER)

	NIST 2001				NIST
	model 512		model 64		model
	EER	Thr	EER	Thr	EER
No VAD	13.63		16.00		44.39
LTSD	12.41	40	13.74	45	35.82
Energy	9.26	36	10.40	35	**16.63**
Periodicity	**8.46**	0.61	**9.58**	0.66	16.76

Summary of the speaker verification results with the corresponding VAD thresholds is shown in Table 1. Best parameters found in tuning with NIST 2001 corpus have been used for the NIST 2006 experiments. We obtained significantly higher error rates for NIST 2006 than NIST 2001, as expected. We also notice that for this corpus the energy based VAD slightly outperforms the periodicity VAD.

4. Speech Segmentation Experiments

Speech segmentation is a task of segmenting the audio signal in time to alternating speech and non-speech blocks. The quality of speech segmentation is evaluated by two measures: *false acceptance* (FA) and *false rejection* (FR) rates.

We used three different materials in our speech segmentation experiments: subset of the 2005 NIST Speaker Recognition Evaluation corpus, Bus-Stop timetable system [14] recordings, and material recorded in our lab [6].

Final comparative results are obtained in the Table 2. We note that the periodicity method is a clear winner in the Lab recording, and provides comparable results to simple energy based method for NIST05 dataset. On the other hand, for the Bus-stop material, periodicity method does not seem to perform adequately. One reason for the higher EER is that periodicity method detects as speech DTMF sounds, which are periodic, but are defined as noise in the ground truth.

Original LTSD VAD assumes that the beginning of the sound file contains noise that is typical to the sessions in question. We noted the poor segmentation results, so we collected non-speech segments from NIST05 evaluation corpus not used in the experiments and trained a separate noise model. This model was then used in all LTSD with trained noise experiments. LTSD VAD seems unstable with respect to the noise modeling scheme used.

Table 2. Summary of segmentation results (% EER).

	Bus-stop	Lab	NIST05
LTSD adaptive	19	14	40
LTSD trained	**6**	15	**1**
Energy	15	17	2

5. Conclusions

Best results are obtained either by energy-based method off-line, or by the proposed periodicity method on-line. In future work, we consider to combine the two measures in a single real-time method either by two-step thresholding, or by a simple classifier fusion.

6. Acknowledgements

The work was supported by the National Technology Agency of Finland (TEKES) as the projects New Methods and Applications of Speech Technology (PUMS).

References

1. *B. A, E. Schlomot, H. Su*, ``ITU-T recommendation g729 annex b: A silence compression scheme for use with g729 optimized for v.70 digital simultaneous voice and data applications," IEEE Communications Magazine, vol. 35, pp. 64-73, 1997.

2. *J. Ramírez, J. Segura, C. Benítez, A. de la Torre, A. Rubio*, ``Efficient voice activity detection algorithms using long-term speech information," Speech Communication, vol. 42, pp. 271-287, 2004.

3. *D. Reynolds, T. Quatieri, R. Dunn*, ``Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol. 10, no. 1, pp. 19-41, 2000.

4. *A. Davis, S. Nordholm, R. Togneri*, ``Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 2, March 2006.

5. *S. Tranter, D. Reynolds*, ``An overview of automatic speaker diarization systems," IEEE trans. on audio, speech and language processing, vol. 14, no. 5, pp. 1557-1565, September 2006.

6. *M. Tuononen, R. González Hautamäki, P. Fränti*, ``Performance of voice activity detection in speech applications," IEEE Trans. on Information Forensics and Security, 2007, submitted.

7. *M. Sambur*, ``Selection of acoustic features for speaker identification," IEEE trans. on Acoustics, Speech, and Singal Processing, vol. 23, no. 2, pp. 176-182, April 1975.

8. *T. Kinnunen, E. Karpov, P. Fränti*, ``Real-time speaker identification and verification," IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 1, pp. 277-288, January 2006.

9. *J. Saastamoinen, E. Karpov, V. Hautamäki, P. Fränti*, ``Accuracy of MFCC based speaker recognition in series 60 device," J. of Applied Signal Processing, no. 17, pp. 2816-2827, September 2005.

10. *R. Tong, B. Ma, K. Lee, C. You, D. Zhou, T. Kinnunen, H. Sun, M. Dong, E. Ching, H. Li*, ``Fusion of acoustic and tokenization features for speaker recognition," in 5th Intl. Sym. on Chinese Spoken Language Processing (ISCSLP 2006), Singapore, December 2006, pp. 494-505.

11. *R. Tucker*, ``Voice activity detection using a periodicity measure," IEE Proc.-I, vol. 139, no. 4, pp. 377-380, 1992.

12. *P. Fränti, T. Kaukoranta, O. Nevalainen*, ``On the splitting method for vector quantization codebook generation," Optical Engineering, vol. 36, no. 11, pp. 3043-3051, November 1997.

13. *A. de Cheveigne, H. Kawahara*, ``YIN, a fundamental frequency estimator for speech and music," The J. of Acoustical Society of America, vol. 111, no. 4, pp. 1917-1930, April 2002.

14. *M. Turunen, J. Hakulinen, A. Kainulainen*, ``Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: similarities and difference," in 9th Intl. Conf. on Spoken Language Processing (ICSLP 2006), Pennsylvania, September 2006, pp. 1057-1060.