Combining Voice Activity Detection Algorithms by Decision Fusion

Evgeny Karpov, Zaur Nasibov, Tomi Kinnunen, Pasi Fränti

Speech and Image Processing Unit, University of Eastern Finland, Joensuu, Finland

ekarpov@cs.joensuu.fi, znasibov@cs.joensuu.fi, tkinnu@cs.joensuu.fi, franti@cs.joensuu.fi

Abstract

This paper presents a novel method for voice activity detection (VAD) by combining decisions of different VAD. To evaluate the proposed technique we include several well known industrial methods to compute VAD decisions on three data sets of varying complexity. We use the outputs of these methods as an input for our decision-level fusion algorithm to produce new VAD labeling and compare them to the original results. Our experiments indicate that the fusion is useful especially when low speech miss rate is desired. The best results were obtained on the most challenging Lab dataset, with low false alarm rate and comparable miss rate.

1. Introduction

Voice activity detection (VAD) is a classification task that aims at partitioning a given speech sample into speech and non-speech segments. It has an important role in various modern speech processing methods and telecom standards [1]. While being a relatively well studied problem, acceptable solution that works in different acoustic conditions is yet to be found.

A large number of VADs have already been proposed. The simplest methods use features such as zero crossing rate, frame energy or spectral entropy to distinguish non-speech frames from speech frames. Other more sophisticated methods use statistical methods to model background noise characteristics and utilize them in decision making [2-4]. However, different methods tend to work inconsistently in varying acoustic conditions or noise levels. For example, the G729 standard [5] method works usually well in moderate noise conditions but provides unacceptable speech detection accuracy with increased noise level. Another example is AMR [6] that works best in very low SNR noise conditions but its conservative behavior degrades its non-speech detection accuracy [9]. Thus, it seems natural to ask whether such complementary information in different methods can be utilized for high-accuracy voice activity detection by fusion. Even though a few studies have been done to combine different features to improve VAD accuracy [13], we are unaware of comprehensive study of decision-level combination of different VAD algorithms. In this paper, we propose to use *majority voting* over short-term temporal contexts to combine different VAD methods. Our base method pool consists of the following methods found in various industrial standards: ITU G729B [5], ETSI AMR option 1 and 2 [6], ETSI AFE [7], emerging Silk codec used in Skype [8] and a simple energy method [14]. In the experiments, we compare these different VAD methods and their fusion on three independent data sets. The first data set (*NIST05*), a subset of the NIST 2005 speaker recognition evaluation (SRE) corpus, is representative data in telephone-based speaker recognition. The second data set (*Bus stop*) consists of speech data found in a speech user interface application. Finally, the third data set (*Lab*) consists of data recorded using low-quality microphone in far-field recording setting, and it emulates wiretapping material found in forensics.

2. Base classifiers: the individual VADs

2.1. Energy VAD

The *energy VAD* is representative method of a simple non-realtime speech detector used often in speech technology research [14]. We first compute the energies of all frames in a given speech utterance. The detection threshold is then set to 30 dB below the maximum frame energy and, additionally, minimum absolute energy threshold of -55 dB is used for rejecting frames with very low energy. These thresholds were originally determined to maximize speaker recognition accuracy on the telephony NIST 2005 and 2006 speaker recognition evaluation corpora [15].

2.2. G729

As an extension to G729, ITU has also published Annex B in order to support discontinuous transmission (DTX) by means of VAD. The G729 VAD operates on 10ms frames and uses background noise model and the following four parameters for decision making [1, 5]:

- a full-band energy difference between input signal and noise model
- a low-band energy difference between input signal and noise model
- a spectral distortion
- a zero crossing rate difference between input signal and noise model

The algorithm has shown to be robust in moderate noise conditions but yields low speech detection rate with increasing noise level [9]

2.3. AMR

AMR option 1 decomposes signal into nine subbands using filterbanks with emphasis on higher frequency bands. For each subband, it calculates energy and signal-to-noise ratio (SNR) estimates. The sum of SNRs is then compared with adaptive threshold to make a VAD decision, followed by a *hangover* scheme [1, 6]. AMR option 2 is similar to option 1 but it uses FFT instead of filterbanks, has 16 subbands, and adapts background noise energy for every band during nonspeech frames [1, 6]. In general, AMR works well in varying noise conditions. However, its conservative behavior degrades its non-speech detection accuracy [9].

2.4. AFE

ETSI advanced feature extraction (AFE) algorithm uses simple energy-based voice activity detection with forgetting factor for updating noise estimate [7]. AFE first computes logarithmic energy of 80 samples of the input signal. It is used to compute mean energy and later these two energy values are used to estimate frame as silence or speech [7].

2.5. Silk

Silk is a speech codec developed by *Skype* [8] for voice over IP communications. It uses VAD algorithm to support discontinuous transmission (DTX) mode where silent frames are dropped from transmission channel. Silk uses a sequence of half-band filterbanks to split the signal in four subbands. For every frame, the signal energy and signal-to-noise ratio (SNR) per subband are computed. VAD decision is then made based on the average SNR and a weighted average of the subband energies [8].

3. Decision-level combination of the base VADs

Most of the standard VADs - as reviewed in the previous section - produce *hard* decisions (speech / non-speech labels) and therefore, *decision-level combination* of VADs is the most natural choice. Selecting an appropriate decision fusion is a research topic in itself [12]. However, to our knowledge, fusion

techniques have not been yet widely applied to voice activity detection problem. There are only a few attempts to utilize decision fusion from different classifiers. In [13] the authors propose two complementary systems whose outputs are merged using fusion. The first system uses non-Gaussianity score feature based on normal probability testing and the second system a histogram distance score feature to detect changes in the signal through template-based similarity measure between adjacent frames [13].

The reason why decision-level combination of VADs has received little attention is because the industrial VADs are mainly used in real-time applications. Having several classifiers running at the same time can be a computationall burden. However, fusion technique has potential uses in non real-time applications like forensic data analysis, voice search and other speech processing tasks that do not require real-time operation.

For our experiments we select two basic strategies: *majority voting* and *temporal context voting*. We describe these algorithms in more details in the following subsections.

3.1. Majority Voting

The idea of majority voting is simple: for each frame we collect decisions from *N* base VADs and then classify each frame as majority of methods report. Basically the more methods vote for certain classification more likely it will be the correct one.

3.2. Including Temporal Context to Majority Voting

As speech-to-non-speech changes occur slowly compared to usual frame duration of about 15 ms, it is useful to smooth results by utilizing contextual information [11]. This is often implemented using a *hangover* scheme [11], which is a state transition machine that helps in correcting mislabeled data. For example, in the VAD output 00100100000, the two isolated ones are most likely mislabeled than short speech segments.

A hangover scheme is usually experimentally determined using method-dependent *ad hoc* rules. The goal in the proposed *temporal context voting* is the same as in hangover – to correct erroneous frame decisions – except that we now combine temporal information from *several* VADs. This is done by extending majority voting over a context of *C* frames. Thus, with *N* base VADs, majority voting is carried out on the concatenated decision vector of *N* x *C* binary decisions. With the context size *C*=1, it reduces back to simple frame-level majority voting rule as a special case.

As an example consider *N*=3 with giving the following frame-level decisions:

VAD1 0 1 1 0 0 0 ...

VAD2 0 1 0 1 0 1 ...

VAD3 0 0 1 1 1 0 ...

The decision function (for context size *C*=3) for the second and third frames on these vectors is the following:

Fusion(2) = round((0+0+0 + 1+1+0 + 1+0+1) / 9) = 0

Fusion(3) = round((1+1+0 + 1+0+1 + 0+1+1) / 9) = 1.

4. Experimental Setup

4.1. Data Sets

In the experiments, we use the datasets listed in Table 1.

The first dataset is a subset of the NIST 2005 speaker recognition evaluation (SRE) corpus, consisting of conversational telephone-quality speech with 8 kHz sampling rate [10]. We have selected this corpus to evaluate algorithms on telephone quality speech material. NIST SRE corpora are commonly used for evaluating speaker verification algorithms where VAD plays an important role.

The second data set, *Bus stop*, consists of timetable system dialogues recorded in 8 kHz sampling rate. The data mainly contains human speech commands that are mainly very short, as well as synthesized speech that provides rather long explanations about bus schedules. This data is a good example of a typical speech dialogue application [16].

The third dataset, *Lab*, consists of a one long continuous recording from the lounge of our laboratory in 44.1 kHz, using a low-quality Labtec PC microphone not specifically designed for far-field recordings. People are often passing our laboratory lounge, which causes false alarms due to, for instance, opening and closing the doors. In addition, our pantry is located in the same facility, so other background sounds include, for instance, sounds from a water tap and microwave oven. The distance of the microphone to the speakers is several meters and the signal-to-noise ratio of these recordings is very low. The goal of this material is to simulate wiretapping material found in forensics or audio surveillance applications, where it is not always practical to install a high-quality microphone to facility being monitored. Due to the massive amount of data in such application – imagine continuous recording for several days in a row – a VAD plays an important role in helping the forensic investigator to rapidly locate speech segments.

	NIST 2005	**Bus stop**	**Lab**
Recording equipment	Telephone	Telephone	Labtec PC Microphone
Total amount of data	12 h 23 min	2h 48min	4 h 12 min
Amount of speech	49%	80%	7%

Table 1. *Data sets used in the experiments and their properties*

4.2. Measuring VAD Accuracy

We measure VAD accuracy in terms of *miss rate* (MR) and *false alarm rate* (FAR) defined as percentage of all actual speech or silence frames that were misclassified as silence or speech respectively.

$$\mathrm{MR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} * 100\% \qquad (1)$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} * 100\% \qquad (2)$$

Here, TP (true positive) and TN (true negative) are the number of real speech and non-speech frames in the evaluation dataset and FN (false negative) and FP (false positive) are the number of misclassified speech and non-speech frames, respectively. Low miss rate for algorithm corresponds to its ability to correctly identify speech frames, whereas low false acceptance rate corresponds to better non-speech detection properties of the algorithm.

5. Results and Discussion

We first utilize the NIST05 data set for selecting the best combination of VADs. The miss and false alarm rates are shown in Table 2 for different selection of base VADs and the context size *C*.

Combined VADs	C=1	C=3	C=5	C=7	C=9	C=11
G729, AMR1, AMR2	23.5	14.4	12.9	12.1	11.4	**10.9**
G729, AMR1, SILK	23.5	13.4	11.1	9.62	8.60	**7.81**
G729, AMR2, SILK	21.3	11.6	9.95	8.78	7.91	**7.24**
SILK, AMR1, AMR2	22.1	13.8	11.6	10.2	9.18	**8.38**

Table 2. *Miss rates (%) for NIST05 with varying context size (C, frames) and base VAD pool.*

Combined VADs	C=1	C=3	C=5	C=7	C=9	C=11
G729, AMR1, AMR2	**38.2**	54.1	57.3	59.8	61.8	63.6
G729, AMR1, SILK	**39.1**	61.4	66.3	70.2	73.2	75.7
G729, AMR2, SILK	**44.5**	65.4	69.5	72.7	75.2	77.4
SILK, AMR1, AMR2	**42.4**	65.3	71.6	75.9	79.2	81.7

Table 3. *FAR (%) for NIST05 with varying context size (C, frames) and base VAD pool.*

Combining G729, AMR2 and SILK produces the best miss rate using context of *C*=11 frames, whereas combining G729, AMR1 and AMR2 produces the smallest false alarm rate with a simple majority voting (context size *C*=1).

In the following, we evaluate how these two combination strategies generalize to our other datasets. Table 4 summarizes the miss rates for the combination of G729, AMR2 and Silk with context of *C*=11 frames (later referred as Fusion 1). Table 5, in turn, shows the result for combination of G729, AMR1 and AMR2 with simple majority voting, e.g. *C*=1 (later referred as Fusion 2). We also show corresponding MR and FAR for both fusion methods to evaluate how these methods affect both metrics.

Corpus	Energy	G.729	AMR1	AMR2	Silk	AFE	Fusion 1	Fusion 2
NIST05	63.9	22.1	25.0	19.1	20.0	17.0	**7.24**	23.5
Bus stop	33.3	12.5	9.26	11.5	14.7	9.97	**1.01**	16.0
Lab	70.9	67.8	63.8	46.6	37.2	33.0	**9.7**	59.3

Table 4. *Miss rates (%) comparison for all methods*

Corpus	Energy	G.729	AMR1	AMR2	Silk	AFE	Fusion 1	Fusion 2
NIST05	**14.9**	40.0	34.4	46.8	50.3	55.5	77.4	38.2
Bus stop	**26.6**	59.3	48.0	46.8	62.8	43.3	94.7	36.7
Lab	30.8	10.8	**8.5**	12.2	37.2	27.3	80.0	9.47

Table 5. *False alarm rates (%) comparison for all methods*

5.1. Discussion

The first fusion strategy (Fusion 1) achieves very low miss rates but increases false alarm rates unusably high. The second fusion strategy with a simple frame-level majority voting (Fusion 2), on the other hand, yields comparable accuracy to the base VADs; it gives the second smallest false alarm rates on the Bus stop and Lab data sets, and third smallest false alarm rate on the NIST '05 data. The miss rates, in turn, are the 5th on NIST '05 and Bus stop and 4th on Lab. Overall, the most promising results are obtained on the extremely noisy Lab data set.

6. Conclusion

In this paper we studied decision-level combination of several well-known voice activity detectors. According to our experiments, simple majority voting gives comparable or better accuracy compared to standard VADs. Using temporal information was not found successful in our experiments. The best results were obtained on the most challenging Lab dataset, with low false alarm rate and comparable miss rate. Accuracy might be further improved by trainable fusion such as weighted voting, so that accuracies of the individual VADs are taken into account. This is left as a future work.

7. References

1 *A.M. Kondoz,* "Digital Speech: Coding for Low Bit Rate Communication Systems", John Wiley & Sons, Ltd. ISBN 0-470-870007-9

2 *J.-H. Chang, N.S. Kim and S.K. Mitra,* "Voice Activity Detection Based on Multiple Statistical Models", IEEE Trans. Signal Processing, 54(6), June 2006, pp. 1965-1976.

3 *J. Ramírez, J.C Segura, C. Benítez, A. de la Torre, A. Rubio* (2004) "Efficient voice activity detection algorithms using long-term speech information". Speech Comm. 42, pp. 271–287.

4 *J. Ramírez, P. Yelamos, J.M. Gorriz, J.C. Segura* (2006) "SVM-based speech endpoint detection using contextual speech features". Elec.Letters 42(7), 2006.

5 ITU-T Recommendation G.729-Annex B. (1996). A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.

6 ETSI EN 301 708 Recommendation: Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, ETSI, Sophia Antipolis, Dec. 1999

7 ETSI ES 202 050 Recommendation: Speech processing, Transmission and Quality aspects (STQ);Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, 2000

8 Silk codec: http://tools.ietf.org/html/draft-vos-silk-00, accessed on 19 May 2011.

9 *A. de la Torre, J. Ramirez, C. Benitez, J. C. Segura, L. Garcia, and A. J. Rubio,* "Noise robust model-based Voice Activity Detection," in Proc. INTERSPEECH2006, USA, 17-21 Sep. 2006, pp. 1954-1957.

10 National Institute of Standards and Technology, NIST speaker recognition evaluations. http://www.nist.gov/speech/tests/spk/, accessed on 19 May 2011.

11 *J. Ramírez, J.C Segura, C. Benítez, A. de la Torre, A. Rubio* (2004) "Efficient voice activity detection algorithms using long-term speech information". Speech Comm. 42, pp. 271–287.

12 *Dymitr Ruta and Bogdan Gabrys,* "An Overview of Classifier Fusion Methods", Computing and Information Systems, 7 (2000) p.1-10

13 *H. Ghaemmaghami, D. Dean, S. Sridharan, I. McCowan.* "Noise robust voice activity detection using normal probability testing and time-domain histogram analysis", in proc. ICASSP 2010, USA, 14-19 March, 2010

14 *T. Kinnunen and H. Li,* "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", Speech Communication 52(1): 12--40, January 2010

15 *R. Tong, B. Ma, K.A. Lee, C.H. You, D.L. Zou, T. Kinnunen, H.W. Sun, M.H. Dong, E.S. Ching and H.Z. Li*, "Fusion of acoustic and tokenization features for speaker recognition", in Proc. ISCSLP, pp. 566-577, Singapore, 2006.

16 *M. Turunen, J. Hakulinen, K.-J. Räihä, E.-P. Salonen, A. Kainulainen, and P. Prusi*, "An architecture and applications for speech-based accessibility systems," IBM Systems Journal, vol. 44, pp. 485-504, 2005.