A PRACTICAL, SELF-ADAPTIVE VOICE ACTIVITY DETECTOR FOR SPEAKER VERIFICATION WITH NOISY TELEPHONE AND MICROPHONE DATA

Tomi Kinnunen and Padmanabhan Rajan

School of Computing, University of Eastern Finland (UEF), Joensuu, Finland

ABSTRACT

A voice activity detector (VAD) plays a vital role in robust speaker verification, where energy VAD is most commonly used. Energy VAD works well in noise-free conditions but deteriorates in noisy conditions. One way to tackle this is to introduce speech enhancement preprocessing. We study an alternative, likelihood ratio based VAD that trains speech and nonspeech models on an utterance-by-utterance basis from mel-frequency cepstral coefficients (MFCCs). The training labels are obtained from enhanced energy VAD. As the speech and nonspeech models are re-trained for each utterance, minimum assumptions of the background noise are made. According to both VAD error analysis and speaker verification results utilizing state-of-the-art i-vector system, the proposed method outperforms energy VAD variants by a wide margin. We provide open-source implementation of the method.

Index Terms— Voice activity detection, speaker verification

1. INTRODUCTION

Voice activity detection (VAD) is the task of locating speech segments from an utterance and it plays a crucial role in any speech processing system. The standard VADs such as the g729 [1], ETSI advanced front-end (AFE) [2] and statistical model VADs [3], have been designed with telecommunication and automatic speech recognition (ASR) desiderata in mind, namely, low complexity and real-time operation. But there are also applications that do not require real-time operation, such as speaker diarization and recognition for screening, indexing or forensic use cases. In these applications, it would be beneficial to utilize the full recording for noise modeling and VAD threshold adaptation. Here we focus on text-independent speaker verification [4].

Energy-based VAD [5, 4] is by far the most popular VAD in speaker verification, possibly due to its simplicity. It computes the energy of each short-term frame and assumes that the low and high energy frames, respectively, correspond to nonspeech and speech. The energy threshold is usually adjusted on an utterance-by-utterance basis. For instance, the threshold can be made relative to maximum or average energy of the utterance [5, 4, 6]. It is also common to fit a 2- or 3-component Gaussian mixture model (GMM) to the energy distribution and adjust the threshold according to the GMM parameters (e.g. cross-over point of the two Gaussians) [7, 8]. This may involve iterative re-training of the energy GMM and the thresholds [9].

A well-known shortcoming of the energy-based VADs are their sensitivity to additive (environmental) noise [10, 6]. Some form of speech enhancement pre-processing seems necessary under low signal-to-noise ratios (SNRs). According to [10], energy VAD with spectral subtraction enhancement can outperform more advanced statistical model VAD [3]. Alternative ways to tackle noise include alternative features such as periodicity [11] or phase [12].

Beyond the simple energy VAD, at the other extreme are methods that adopt an off-the-shelf phone recognizer or trainable models for VAD [13, 14, 15, 16, 12]. For instance, phone posterior probabilities can be merged and combined with energy measures [13]. Such VADs are generally complex. Their pre-trained acoustic models may also not generalize well to unseen types of channels or environments. In [6], phone recognizer VAD, similar to energy VAD, was found sensitive to additive noise degradation.

We propose a practical VAD that does not rely on pre-trained acoustic models but is trained only from the recording at hand. This is achieved via an initial energy VAD to label a small number of "reliable" training vectors for that utterance. Two GMMs, one for speech and one for nonspeech, are trained and a likelihood ratio detector is used for labeling all the frames as speech or nonspeech. The proposed VAD is designed with the following criteria in mind:

- **Unsupervised:** It does not require hand-labeled training sets.
- **Self-adaptive:** It does not require pre-trained speech/nonspeech models but adapts itself to a given utterance. It makes no strong assumptions of the type or level of noise.
- **Practical:** It is directly applicable to both telephone and interview data in recent NIST SRE data. It does not use costly Viterbi decoding or iterative re-training procedures.

Relation to previous work: Closest similar works to ours are VADs described in [17, 18, 19], which also use a preliminary VAD to obtain training class labels. Unlike [17, 18] that use maximum a posteriori (MAP) training, we use maximum likelihood training (ML) as [19], but without iterative Viterbi segmentation and re-training to speed up processing. Further speed-up is achieved by use of codebooks, viewed as constrained GMMs [20]. As part of our preparation to the latest NIST SRE 2012 evaluation with I4U coalition [21], another novelty is a systematic study of the effect of the VAD control parameters with a special focus on additive noise degradation with use of multiple enrollment utterances per speaker. The speaker verification experiments are reported using a state-of-the-art i-vector system on the I4U dataset.

2. VOICE ACTIVITY DETECTORS

2.1. Simple Energy VAD

For completeness, we describe here the adaptive energy VAD. Let $x_t[n]$ denote the n^{th} sample of the t^{th} speech frame in an utterance. We first compute the log-energy of each frame as,

$$E_t = 10\log_{10}\left(\frac{1}{N-1}\sum_{n=1}^{N}(x_t[n]-\mu_t)^2+\epsilon\right), \qquad (1)$$

The work was supported by Academy of Finland (proj. 132129, 253120). We thank I4U coalition for devset design and Hanwu Sun (I2R), Pierre Ouellet (CRIM) and Rahim Saeidi (RUN) for useful discussions.

where $\mu_t = (1/N)\sum_{n=1}^{N} x[n]$ is the sample mean of the frame, N is the frame size and $\epsilon = 10^{-16}$ is an arbitrary constant to avoid log of zero. We find the maximum energy $E_{\max} = \max_{t=1,\ldots,T}\{E_t\}$ over all the T frames of the utterance. The VAD decision is simple threshold comparison adjusted according to this maximum level. Additionally, a minimum energy constraint is used to avoid utterances with low energy being falsely tagged as speech. Thus, the energy VAD rule for speech presence is $(E_t > E_{\max} - \theta_{\text{main}}) \wedge (E_t > \theta_{\min})$, where θ_{main} and $\theta_{\min}$, respectively, denote pre-set primary and minimum energy thresholds. These were set to $\theta_{\text{main}} = 30$ dB and $\theta_{\min} = -55$ dB [4] when optimizing the spectral subtraction parameters (Section 4).

2.2. Energy VAD with Spectral Subtraction

For high signal-to-noise ratios (SNRs), the energy VAD works reasonably well but in low SNRs it tends to mark most frames as speech. One strategy to remedy this is to use a plug-in speech enhancement method, intended for increasing SNR, prior to the above-described energy VAD. The spectral subtraction method is based on MATLAB implemention `specsub` in *Voicebox*[1]. Let $|X|^2$ and $|\hat{N}|^2$, respectively, denote the powers of noisy speech and estimated noise in a particular time-frequency FFT bin. Spectral subtraction is achieved by multiplying noisy magnitude $|X|$ by a gain factor g whose general form is [22],

$$g = \max\left\{\left(1 - \left(\alpha\frac{|\hat{N}|^2}{|X|^2}\right)^{\gamma/2}\right)^{e/\gamma}, \min\left(g_h, \left(\beta\frac{|\hat{N}|^2}{|X|^2}\right)^{e/2}\right)\right\},$$

where α is an oversubtraction factor, γ determines the subtraction domain, e is gain exponent, g_h is maximum gain for noise floor and β determines maximum noise attenuation in the power domain. The gain-multiplied magnitude is combined with phase of the noisy signal followed by overlap-and-add signal reconstruction.

We fix $g_h = 1.00$ and $\beta = 0.01$ and focus on (1) subtraction domain, (2) amount of oversubtraction and (3) noise estimator. Regarding the subtraction domain, **magnitude domain** subtraction is obtained by choosing $(\gamma, e) = (1, 1)$, **power domain** spectral subtraction by $(\gamma, e) = (2, 1)$ and **Wiener filter** by $(\gamma, e) = (2, 2)$. Regarding the amount of subtraction, α varies linearly from $\alpha = \alpha_{\max}$ for a frame signal-to-noise ratio (SNR) of -5 dB down to $\alpha = 1$ for SNR = 20 dB; we treat the maximum oversubtraction factor, $\alpha_{\max}$, as a control parameter. Regarding noise estimator, we consider two well known alternatives, *minimum mean square error* (MMSE) [23] and *minimum statistics* (MS) [24] noise trackers.

2.3. Proposed Self-Adaptive VAD

The proposed method is outlined in the pseudocode. First, MFCCs are extracted from the original signal. The signal is then enhanced with spectral subtraction whose purpose is to merely increase the energy contrast between speech and nonspeech without caring of spectral subtraction artefacts. Therefore, relatively aggressive oversubtraction is used [10]. Following this, we sort the energy values and find a fixed percentage of the lowest and highest energy frames (for instance, 10 % of all frames) assumed to correspond, respectively, to reliably-labeled nonspeech and speech frames. Speech and nonspeech models are then trained using the MFCCs corresponding to these frame indices. Finally, all the frames are labeled using the trained models, with an additional minimum-energy constraint.

[1] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html` (URL valid March-2013).

Inputs: Speech signal $s[n]$, frame length (L) and hop (H)
Outputs: Binary VAD labels $\text{VAD}[t]$, $t = 1, 2, \ldots, T$

1. // Extract MFCCs from the **noisy** signal
   $X \leftarrow \text{ExtractMFCC}(s, L, H, \text{MFCCParams})$;
2. // Denoise the speech signal
   $s_{\text{clean}} \leftarrow \text{Specsub}(s, \text{SpecsubParams})$;
3. // Compute frame energies of the **enhanced** signal, Eq. (1)
   $E \leftarrow \text{ComputeEnergy}(s_{\text{clean}}, L, H)$;
4. // Find indices of low/high energy frames (fixed percentage)
   $[i_{\text{low}}, i_{\text{high}}] \leftarrow \text{FindLowestAndHighest}(\text{E}, \text{percentage})$;
5. // Train speech and nonspeech models from the frame **subsets**
   $\lambda^{\text{speech}} \leftarrow \text{Train}(\{\boldsymbol{x}_t \in X | t \in i_{\text{high}}\}, \text{ModelParams})$;
   $\lambda^{\text{nonspeech}} \leftarrow \text{Train}(\{\boldsymbol{x}_t \in X | t \in i_{\text{low}}\}, \text{ModelParams})$;
6. // For **all** frames, pick the more likely hypothesis
   $\text{VAD}[t] \leftarrow \{\log p(\boldsymbol{x}_t|\lambda^{\text{speech}}) \geq \log p(\boldsymbol{x}_t|\lambda^{\text{nonspeech}})\} \wedge E_t \geq \theta_{\min}$; // With min-energy constraint

---

Note that all MFCC processing uses features of the original (noisy) signal rather than the enhanced one that contains spectral subtraction artefacts.

Both the speech and nonspeech models are GMMs of the form $p(\boldsymbol{x}|\lambda) = \sum_{k=1}^{K} P_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with mixing weights P_k, mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k$. Even though different number of Gaussians can be used for speech and nonspeech models [18, 19], we use the same number for simplicity. Two challenges in maximum likelihood training are small amount of data and zero frames found in NIST corpora causing duplicated vectors and numerical problems. To avoid duplicates, we add random Gaussian noise with small amplitude (10^{-9}) to the signal as a preprocessing step, similar to dithering option in HTK. Further, since our goal is to retain low complexity, we use k-means instead of expectation maximization (EM). As discussed in [20, p. 443-444], k-means can be viewed as a limit case of EM with identical covariance matrices $\boldsymbol{\Sigma}_k = \epsilon\boldsymbol{I}$ where $\epsilon \to 0$. Assuming equal speech/nonspeech priors and misclassification costs, the log-likelihood ratio test for speech presence for vector $\boldsymbol{x}_t$, i.e. $p(\boldsymbol{x}_t|\lambda^{\text{speech}}) \geq \log p(\boldsymbol{x}_t|\lambda^{\text{nonspeech}})$, reduces to the nearest-neighbor rule, $\min_k \|\boldsymbol{x}_t - \boldsymbol{\mu}_k^{\text{speech}}\|^2 \leq \min_k \|\boldsymbol{x}_t - \boldsymbol{\mu}_k^{\text{nonspeech}}\|^2$, where $\boldsymbol{\mu}_k^{(\cdot)}$ are the codevectors obtained using k-means.

3. EXPERIMENTAL SETUP

The experiments are divided into two parts. First, VAD parameters are optimized to minimize average frame labeling error. The optimized VAD is then integrated into a speaker verification system.

3.1. VAD development set

To evaluate VAD accuracy, we adopt a simulated additive noise protocol. To this end, we utilize utterances in the development set of the NIST 2010 speaker recognition evaluation (SRE) campaign. This dataset, provided by NIST, contains 2-channel recordings from 18 interview and 36 telephone recordings with supplementary automatic speech recognition (ASR) transcripts. Here we use only the telephone segments due to cross-talk and subsequently incorrect VAD in the interview data. Both sides of the telephone conversations, leading to 36×2 = 72 unique recordings, are used. They

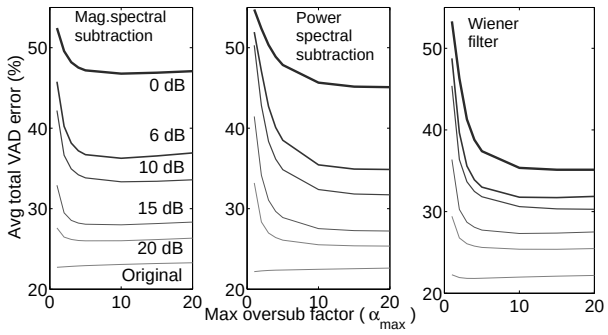

Fig. 1: Energy VAD with spectral subtraction (MMSE noise tracker).

are downgraded with controlled signal-to-noise ratios (SNRs) ranging from [original, 20dB, 15dB, 10dB, 6dB, 0dB] using g729 audio weighting filter for level determination. We use the open-source *Filtering and Noise Adding Tool* (FaNT)[2]. The noise files were selected from *FreeSound*[3] and contain mostly device sounds found at home environments (e.g. airconditioner and washing machine). For a given clean speech file, the noise file was randomly selected and the noise section in the file was also randomized.

Table 1: Energy VAD (average VAD error, %) with and without spectral subtraction in magnitude (**Mag.**), power (**Pow.**) or **Wiener** domain. Max. oversubtraction factor is $\alpha = 10$. **Minstat**.=min. statistics and **MMSE**=min. mean square error noise estimator.

SNR	Baseline	Minstat.			MMSE		
(dB)	(no specsub.)	Mag.	Pow.	Wiener	Mag.	Pow.	Wiener
Orig.	21.90	22.35	22.11	**21.54**	23.08	22.47	21.98
20	44.33	25.79	25.52	**25.32**	26.02	25.52	25.39
15	50.37	27.68	27.36	**27.25**	28.00	27.53	27.32
10	54.30	31.70	31.86	**30.61**	33.34	32.38	**30.61**
6	54.85	36.07	35.05	**31.74**	36.27	35.45	31.76
0	55.63	45.77	45.17	**34.92**	46.78	45.66	35.35

The accuracy of a VAD is evaluated by comparing the predicted VAD labels with a clean reference segmentation obtained from the ASR transcripts provided by NIST. Let $\hat{y}_t(n) \in \{0,1\}$ and $y_t(n) \in \{0,1\}$, respectively, denote the predicted and ground truth VAD label of frame t in file n, and let $T(n)$ denote the total number of frames in utterance n. Our primary metric for VAD tuning is *average total error rate*,

$$\mathcal{E} = \frac{1}{N_{\text{utt}}} \sum_{n=1}^{N_{\text{utt}}} \frac{1}{T(n)} \sum_{t=1}^{T(n)} \mathcal{I}\big\{\hat{y}_t(n) \neq y_t(n)\big\}, \qquad (2)$$

where $\mathcal{I}\{\cdot\}$ is an indicator function and $N_{\text{utt}} = 72$ is the number of utterances. Additionally, average *miss* and *false alarm* rates, corresponding to assertions $\{\hat{y}_t(n) = 0 \wedge y_t(n) = 1\}$ and $\{\hat{y}_t(n) = 1 \wedge y_t(n) = 0\}$, respectively, are reported for selected cases.

3.2. Speaker verification experiments

As part of the pre-evaluation activity for the NIST SRE 2012, the I4U coalition developed a speaker verification devset based on previous years' NIST corpora which is adopted here. The I4U training and test data was drawn from SRE 2006, 2008 and 2010 corpora including both telephone and microphone data. In addition to the original recordings, two noisy versions (15 dB and 6 dB) of each utterance were generated using FaNT where the noises included similar device sounds as the VAD devset but also additional (unintelligible) *crowd* noises. A different set of noise files from VAD optimization is used for the speaker verification part. More details are given in [21].

A state-of-the-art *probabilistic linear discriminant analysis* (PLDA) [25] classifier with an i-vector extractor [26] and length normalization [27] is used. 18 MFCCs are extracted, followed by RASTA filter, delta and double deltas (54 dimensions), frame dropping using one of the compared VADs, and global cepstral mean/variance normalization (CMVN). Gender-dependent universal background models (UBMs) with 1024 diagonal covariance Gaussians are trained from NIST 04, 05, 06 and 08 data. The i-vector extractor (T-matrix) with 600 dimensions is trained (5 iterations) using the same corpora plus Fisher and Switchboard. These are also used for PLDA training with speaker and channel subspace dimensions of 200 and 0, respectively. The original utterances without added noise are used for UBM and T-matrix training, but both the PLDA training and the enrollment data contain original and noisy utterances. A single, averaged i-vector is used as the enrollment representation. To evaluate accuracy, we report both equal error rate (EER) and normalized MinDCF following NIST 2010 speaker recognition evaluation plan ($P_{\text{tar}} = 0.001$, $C_{\text{miss}} = C_{\text{fa}} = 1$).

4. RESULTS: VAD OPTIMIZATION

We first optimize the denoising parameters in the energy VAD. The first parameters of interest are the maximum oversubtraction factor and spectral subtraction domain (magnitude, power or Wiener). The results, using minimum mean square error (MMSE) noise tracker of [23], are given in Fig. 1. As expected, accuracy drops dramatically with decreasing SNR. Regarding oversubtraction, aggressive oversubtraction is helpful as noted earlier [10]. Stabilization occurs roughly for $\alpha \geq 5$ for magnitude subtraction and $\alpha \geq 10$ for power subtraction and Wiener filter across all SNRs. Regarding the subtraction domain, the results are similar for high SNRs but for SNRs less than 15 dB, Wiener filter clearly wins.

We next study the influence of the noise tracker. We fix maximum oversubtraction factor to $\alpha = 10$ and compare the MMSE [23] and minimum statistics [24] methods in Table 1. The results for baseline energy VAD without spectral subtraction are also displayed. Clearly, any enhancement yields remarkable improvement. Regarding noise estimator, minimum statistic method wins in most cases, but the difference is small. This was further confirmed by visually comparing the estimated noise spectra which appeared very similar. The MMSE tracker implementation executes faster and is fixed the rest of the experiments.

We now turn attention to the proposed VAD, where the optimized spectral subtractor (Wiener domain, MMSE tracker, $\alpha = 10$) is used for training vector labeling. 12 MFCCs (including C0) without any normalizations or deltas are extracted. 10 % of the frames are used for training speech and nonspeech models with $K = 16$ codevectors each. Table 2 shows the results for energy VAD without (column 1) and with (column 2) spectral subtraction. The proposed self-adaptive VAD is reported using energy VAD without (column 3) and with (column 4) enhancement of energy. The last column gives the result when both the energy *and* the MFCCs are extracted from Wiener-filtered signal. We observe the following from Table 2:

[2] `http://dnt.kr.hsnr.de/download.html`
[3] `http://www.freesound.org`

Table 2: Energy VAD and proposed VAD (codebook $K = 16$) with and without spectral subtraction (**SS**), average VAD errors (%). In the last column the MFCCs are also extracted from denoised speech.

SNR (dB)	Energy VAD		Self-adaptive VAD		
	No SS.	SS.	No SS.	SS. (energy)	SS. (energy & MFCCs)
Orig.	21.90	21.98	**10.90**	12.46	12.54
20	44.33	25.39	26.63	23.15	**22.29**
15	50.37	27.32	30.21	**25.24**	25.61
10	54.30	30.61	36.01	**28.21**	28.59
6	54.85	31.76	40.05	**30.00**	30.32
0	55.63	35.35	45.75	**34.04**	34.49

- Column 1 vs 3: proposed VAD systematically outperforms energy VAD when spectral subtraction is turned off.
- Column 3 vs 4: spectral subtraction for energy cleaning further improves accuracy of the proposed VAD.
- Column 2 vs. 4: when spectral subtraction is included, there is a slight yet systematic advantage of using the proposed VAD. The same energy VAD is used in both, so this additional gain is due to the trained MFCC-based VAD.
- Column 4 vs 5: there is not much difference whether the MFCCs are extracted from the original or enhanced signal.

To sum up, the proposed VAD outperforms energy VAD systematically across all SNRs. For the rest of the experiments, we use spectral subtraction to enhance energy only. As a final analysis, Fig. 2 displays separately the average miss and false alarm rates for varying number of codevectors and percentage of training frames in an utterance (20 dB and 0 dB). Comparing the scales of the two graphs, majority of the VAD errors come from missed speech. Increasing the codebook size decreases miss rate and slightly increases false alarm rate. Regarding the amount of training vectors, larger training set reduces miss rate and increases false alarm rate.

The miss rate should be interpreted with caution because of erroneous ASR transcripts used as reference; a large proportion of "missed speech" likely originates from speech-internal pauses considered as speech according to ASR transcript but which, for speaker verification, should be considered nonspeech. Considering noisy utterances, we would like to mainly retain low false alarm rate without removing too many speech frames; we fix the amount of training data to 10 % with codebook size $K = 16$ for the speaker verification part. The minimum energy threshold (same for all the three VADs) was re-adjusted to $\theta_{\min} = -75$ dB and main threshold of energy VAD to $\theta_{\text{main}} = 45$ dB following [5]. These selections were confirmed by listening and viewing spectrograms of the VAD devset.

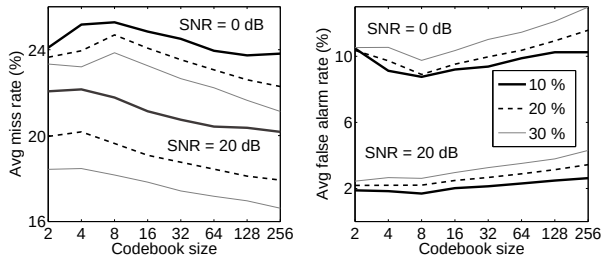

Fig. 2: Effect of the amount of data used for VAD initialization (10 %, 20 %, 30 %) and codebook size (2,4,...,256).

Table 3: Speaker verification accuracy (EER %) on the female trials of the I4U devset. **En.** = energy VAD, **SS-En.** = energy VAD with spectral subtraction, **Prop.** = Proposed self-adaptive VAD.

Test	tel phn			mic phn			mic int		
SNR	En.	SS-En.	Prop.	En.	SS-En.	Prop.	En.	SS-En.	Prop.
Orig	1.59	0.94	**0.85**	7.64	3.10	**1.82**	0.87	**0.35**	0.48
15dB	4.89	2.13	**0.94**	8.07	4.57	**2.25**	1.16	**0.69**	0.94
6dB	7.51	4.47	**1.45**	9.18	6.12	**3.69**	3.20	2.26	**1.77**
All	4.66	2.51	**1.08**	8.29	4.50	**2.58**	1.74	1.10	**1.06**

Table 4: Same as Table 3 but for MinDCF.

Test	tel phn			mic phn			mic int		
SNR	En.	SS-En.	Prop.	En.	SS-En.	Prop.	En.	SS-En.	Prop.
Orig	0.37	0.20	**0.19**	0.64	0.33	**0.27**	0.24	0.10	**0.09**
15dB	0.63	0.36	**0.20**	0.70	0.42	**0.29**	0.28	0.14	**0.10**
6dB	0.81	0.62	**0.26**	0.91	0.69	**0.38**	0.59	0.36	**0.20**
All	0.60	0.39	**0.21**	0.75	0.48	**0.31**	0.37	0.20	**0.13**

5. RESULTS: SPEAKER VERIFICATION

The speaker verification results (female trials only), in terms of equal error rate (EER) and MinDCF are shown in Tables 3 and 4. Since we use multi-condition training including multiple SNRs and channel types and a variable number of training segments per speaker, we report the results considering the test file SNR and data type. The latter includes phone conversations with telephone channel (**tel-phn**) and microphone channel (**mic-phn**), as well as interview scenario with microphone channel (**mic-int**). The results for pooled trials across all test SNRs are are given in the last rows of each table.

As expected, accuracy drops with decreasing SNR. Energy VAD without spectral subtraction yields highest error rates as expected. Including this simple enhancement yields a considerable boost. The proposed self-adaptive VAD further improves on this by a large margin. In the case of interview data EER for the original and 15 dB test, energy VAD with spectral subtraction slightly outperforms the proposed VAD. Regarding the noisiest 6 dB case and MinDCF, the proposed VAD wins again. Our cross-talk cancellation strategy, which was implemented as logical operation (interviewee **AND NOT** interviewer) [5] may be suboptimal. In summary, the overall results indicate that the proposed VAD can indeed handle data with different channel type and SNR without breaking down; it experiences much smaller relative degradation with decreasing SNR in comparison to the energy VAD variants.

6. CONCLUSIONS AND POINTER TO PROGRAM CODE

We studied a simple VAD for speaker verification with promising results over spectral subtraction VAD which is considered one of the high-performance VADs in speaker verification [10, 9]. Consistent behavior on telephone, microphone data, clean and noisy data was observed. Results should be further compared to similar utterance-by-utterance adaptive VADs (e.g. [19]). The problems of threshold selection to maximize speaker verification accuracy and detect nonspeech-only utterances deserve attention as well. A MATLAB implementation of our VAD is available at `http://cs.uef.fi/pages/tkinnu/VQVAD/VQVAD.zip`.

7. REFERENCES

[1] A. Benyassine, E. Schlomot, and H.Y. 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] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI ES 201 108 Recommendation, 2002.

[3] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[4] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[5] H. Sun, B. Ma, and H. Li, "Frame selection of interview channel for NIST speaker recognition evaluation," in *Proc. 7th Int. Symposium on Chinese Spoken Language Processing (ISCSLP 2010)*, Nantou, Taiwan, December 2010, pp. 305–308.

[6] M. Sahidullah and G. Saha, "Comparison of Speech Activity Detection Techniques for Speaker Recognition," *ArXiv e-prints*, Oct. 2012.

[7] I. Magrin-Chagnolleau, G. Gravier, B. Guillaume, and R. Blouet, "Overview of the 2000-2001 ELISA consortium research activities," in *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)*, Crete, Greece, June 2001, pp. 61–66.

[8] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D.A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[9] M. McLaren and D. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *Proc. NIST SRE 2011 workshop*, Atlanta, US, December 2011.

[10] H.B. Yu and M.W. Mak, "Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation," in *Proc. Interspeech 2011*, Florence, Italy, August 2011.

[11] V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, and P. Fränti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. 12th International Conference on Speech and Computer (SPECOM 2007)*, Moscow, Russia, October 2007, pp. 645–650.

[12] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2026–2038, September 2012.

[13] E. Marcharet, G. Potamianos, K. Visweswariah, and J. Huang, "The IBM rt06 evaluation system for speech activity detection in CHIL seminars," in *Proc. of the Third international conference on Machine Learning for Multimodal Interaction (MLMI'06)*, 2006, pp. 323–335.

[14] N. Brümmer, L. Burget, P. Kenny, P. Matejka, E. Villiers de, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*, Brno Univ. Tech., 2010, pp. 1–20.

[15] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J.H. Černocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1979–1986, September 2007.

[16] S. Ganapathy, P. Rajan, and H. Hermansky, "Multi-layer perceptron based speech activity detection for speaker verification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2011)*, October 2011, pp. 321–324.

[17] M. Huijbregts, C. Wooters, and R. Ordelman, "Filtering the unknown: Speech activity detection in heterogeneous video collections," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007, pp. 2925–2928.

[18] H. Sun, T. L. Nwe, B. Ma, and H. Li, "Speaker diarization for meeting room audio," in *Proc. Interspeech 2009*, Brighton, UK, 2009, pp. 900–903.

[19] P. Kenny, P. Ouellet, and M. Senoussaoui, "The CRIM system for the 2010 nist speaker recognition evaluation," in *Proc. NIST 2010 Speaker Recognition Evaluation*, Brno Univ. Tech., 2010.

[20] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, 2006.

[21] R. Saeidi with $>$ 30 coauthors, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Submitted to Interspeech 2013*, 2013.

[22] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP 1979*, 1979, vol. 4, pp. 208–211.

[23] T. Gerkmann and R.C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, pp. 1383–1393, 2012.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[25] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S.J.D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, January 2012.

[26] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[27] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.