# Decision Robustness of Voice Activity Segmentation in unconstrained mobile Speaker Recognition Environments

Andreas Nautsch,[1] Reiner Bamberger[1] and Christoph Busch[1]

**Abstract:** Voice activity detection (VAD) is an essential segmentation process in speaker recognition systems, seperating speech and non-speech segments of voice samples. In speaker recognition, references are modelled purely by concerning speech segments. Different VAD segmentations lead to variations in biometric models, and consequently in system performance. Thus, VAD decisions need to be robust among different conditions.

In this paper, the decision robustness of different VAD algorithms is examined on mobile data by simulating different environmental noise conditions for which we propose a Hamming distance based analysis. By examining speech and speaker recognition based VADs, we further propose to extend a well-performing VAD algorithm, which is based on likelihood ratio comparison of speech to non-speech models, by including most dominant frequency component (MDFC) features for selection of model training segments. Thereby, more robust VAD decisions are conducted by 7%, while sustaining an average EER SNR-sensitivity of 0.76% per dB SNR.

**Keywords:** voice activity detection, robust segmentation, speaker recognition

## 1 Introduction

Biometrics becomes more popular on mobile devices, especially for payment methods. In order to cope with unconstrained mobile environments in speaker recognition, biometric system designers need to ensure front-ends make decisions robustly, i.e. the segmentation of voice samples remains rather constant no matter the environmental condition or signal quality. The segmentation of voice samples into speech and non-speech frames is referred to as voice activity detection (VAD). Based on VAD-selected front-end speech signal features, back-ends extract and compare biometric features. VAD segmentation decision robustness is fundamental to the performance of any speaker recognition back-end e.g., the conventional Gaussian mixture model and universal background model approach (GMM – UBM), Gaussian supervector kernel support vector machines (GSV-SVMs), joint factor analysis (JFA), or intermediate-sized vectors with probabilistic linear discriminate analysis comparison (i-vector/PLDA). Stable segmentation decisions are important to the reliability of biometric systems in unconstrained environments, such as in mobile banking using voice recognition or automated forensic speaker recognition.

In this paper, a novel Hamming distance based VAD metric is proposed, which contrary to conventional VAD metrics stemming from the field of automatic speech recognition (ASR) are more suitable for assessing VAD performance for biometric recognition purposes.

[1] da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany,
andreas.nautsch@h-da.de

Aiming at mobile environments, experiments are carried out on the publicly available MOBIO speaker recognition evaluation task (MOBIO SRE'13) [Mc12, Kh13], which *provides a challenging and realistic test-bed for current state-of-the-art speaker verification* [Kh13]. Different conditions are simulated in terms of additional noise, in particular: white, pink, car and babble noise with white and pink noise representing random channel effects and natural-environment backgrounds, respectively. Based on our VAD analysis of six segmentation algorithms, we propose a more robust variation of the unsupervised GMM-based VAD introduced in [Al14] utilizing MDFC features rather than energy values for the unsupervised training of GMMs modeling speech and non-speech frame segments.

This work is organized as follows: in Section 2 a selection of well-established and recent VAD algorithms is introduced as well as ASR-related VAD metrics, in Section 3 the Hamming distance based analysis is proposed, in Section 4 experimental analysis with respect to VAD, biometric and forensic performance metrics are depicted, which are carried out on noise simulations of the MOBIO database, and in Section 5 conclusions are derived alongside with provided future perspectives.

## 2 Related Work

VAD algorithms are introduced from the field of ASR: in order to recognize the verbalized text under less computational effort, speech is segmented into relevant parts, i.e. words by excluding silence, noise and non-speech sounds. Therefore, VADs are conventionally designed, such that speech is not clipped and non-speech is not falsely segmented as speech.

### 2.1 VAD Algorithms

In this work, emphasis is put on the *ITU recommendation P.56* [IT94][3], the *Voicebox VAD* (VBX) [KS99, Ma01, Br05], the *simple real-time VAD* (SRT) [MH09], the *low-complexity variable framerate VAD* (VFR) [TL10], the *practical, self-adaptive VAD* (PSA) [KR13], and the *unsupervised GMM-based VAD* (USG) [Al14], for which we propose an extension (MUG) incorporating features from the simple real-time VAD approach.

> **P.56** ITU P.56 [IT94] contains a VAD for telephone speech transmission quality in real-time applications. P.56 is a multi-stage VAD, firstly a two-stage exponential averaging on the rectified signal values is performed, secondly initial VAD decisions are conducted by fix-threshold comparison. Thereby, frames are processed in a geometric progression scheme with accumulative activity and hangover counters. The P.56 hangover scheme delays speech to non-speech decisions by 0.2 s, preserving low-energy speech at the end of utterances. Thirdly, activity levels of frames are estimated based on activity counters, which are finally compared to a sample-adaptive threshold based on the long-term energy and a 15.9 dB margin.

[3] P.56 (03/93) is suceeded by P.56 (12/11) with solely changes in annexes. We refer to the P.56 (03/93) Voicebox implementation [Br05].

**VBX** The Voicebox VAD [Br05] is a first-order Markov process modeling of speech with minimum statistic noise estimation. It extends the VAD of [KS99] by conducting a frame-based log-likelihood ratio (LLR) test for speech and non-speech hypotheses examining a-posteriori SNRs estimates based on the power spectrum after discrete Fourier transform (DFT). The noise spectrum is estimated using minimum statistics noise estimation (MSNE) [Ma01] instead of the minimum mean-square error (MMSE) estimator. Thereby, spectral minima are tracked in each frequency band without any speech or non-speech assumptions. The power spectral density (psd) estimation is smoothed by a conditional mean square error estimator, and a-priori and a-posteriori SNRs are computed for each frequency band w.r.t. the variance of the smoothed and bias-compensated psd estimation. A Hidden Markov Model (HMM) based hangover scheme is conducted, such that Bayesian decisions for currents states also depend on previous observations. Speech decisions are conducted on speech posterior probability threshold of 70%. Contrary to conventional hangover schemes, delaying transitions in speech to non-speech decisions, the property of strong correlations in consecutive occurences of speech frames is modeled explicitly by the VBX VAD [KS99].

**SRT** Targeting a simple, efficient and robust algorithm, [MH09] proposed an easy-to-implement and low-complexity VAD for real-time applications based on short term features, i.e. the short-term energy, the spectral flatness measurement (SFM, in dB), and the most dominant frequency component (MDFC), where SFM represents the dB-domain ratio of the geometric mean to the arithmetic mean of the speech spectrum, and MDFC represents the frequency corresponding to the maximum spectral value. For each feature, [MH09] estimates thresholds based on the minimum feature values within first 30 frames assuming them to partially contain non-speech sequences. SRT decides on speech if one of the following votes is positive: short-term energies surpass the minEnergy by an adaptive margin, MDFCs surpass the minMDFC by 185 frequencies, SFMs surpass the minSFM by 5, i.e. the geometric spectral mean is favored over the arithmetic spectral mean by a factor of $\sqrt{10}$. Energy minima $E_{\text{min}}$ estimates increase frame-wise by the number of consecutively observed non-speech frames, such that an adaptive threshold is computed as $E_{\text{thres}} = 40\log(E_{\text{min}})$. SRT examines frames of 10 ms.

**VFR** Contrary to conventional frame window and hop size set-ups in speech processing (25 ms and 10 ms) assuming speech signals to have stationary behavior in short time segments, VFR VAD [TL10] assigns higher frame rates to fast changing and lower frame rates to rather steady events e.g., consonants vs. vowels or silence. Thereby, frames are examined with a frame shift of 100 Hz (1 ms hops), and reliable regions in noisy speech are emphasized on. VAD decisions are carried out on a-posteriori SNR estimate distances of consecutive frames: if accumulated distances of non-speech frames surpass an frame-adaptive threshold, frame segments are denoted as speech. VFR

preserves sigmoidal turning points between 15 dB and 20 dB. In the online available source code to [TL10], the VFR VAD decision is outvoted, if the posterior probability of a frame being voiced is larger than 25% by utilizing the Voicebox pitch tracker [Br05].

**PSA** Targeting NIST speaker recognition evaluations (SREs), Kinnunen and Rajan [KR13] proposed an unsupervised, self-adaptive and practical VAD based on Mel-frequency cepstral coefficients (MFCCs) $\boldsymbol{x}$ and frame energies of denoised and enhanced signals by spectral substractions in magnitude domain, power domain, and Wiener filtering utilizing MSNE for noise tracking. MFCCs of the frames associated to 10% of the lowest and highest clean energy values are utilized in order to train two GMMs representing non-speech and speech $\lambda^{speech}, \lambda^{non\text{-}speech}$, respectively, which take the form of: $p(\boldsymbol{x}\,|\,\lambda^{speech,non\text{-}speech}) = \sum_{c=1}^{C} w_c \, \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu_c}, \boldsymbol{\Sigma_c})$, with mixture weights $w_c$, component means $\boldsymbol{\mu_c}$ and covariances $\boldsymbol{\Sigma_c}$, where $\lambda^{speech}, \lambda^{non\text{-}speech}$ have the same number of components $C$ for simplicity. GMMs are trained using k-means in order to retain low complexity with $C = 16$ codevectors (components) for 12 static MFCCs (including C0) without any normalizations nor deltas. VAD decisions are conducted within the Bayesian decision framework assuming equal priors and costs, such that the LLR test reduces to the nearest-neighbour rule, i.e. to a vector quantization based approach: $\min_c \|\boldsymbol{x_t} - \boldsymbol{\mu_c^{speech}}\|^2 \leq \min_c \|\boldsymbol{x_t} - \boldsymbol{\mu_c^{non\text{-}speech}}\|^2$ , where a simple energy-based VAD decision $\log E(t) \geq -75\,\text{dB}$ needs to hold as well in order to consider a frame as speech.

**USG** In [Al14], an unsupervised GMM based VAD is proposed based on a similar design to the PSA VAD [KR13], where VAD decision making is conducted by LLR and energy decisions, followed by an finite state machine (FSM) based hangover scheme. In our implementation, rastamat [El05] is used for computing energy values. The energy decision are conducted after smoothing $\log E(t)$ values by a 9-frame sliding window moving averaging filter, where the energy threshold $\eta_E$ is the average of the values of the 20% and 80% quantiles of sorted $\log E(t)$ values. Similarly, the sample-adaptive LLR threshold $\eta_{LLR}$ with 23-frame smoothing. As in the PSA VAD [KR13] both speech votes are required for considering a frame as speech. Finally, a hangover scheme is applied that recover speech segments masked by acoustic noise in two distinct ways: transitions from non-speech to speech states are delayed, i.e. in order not to move into the speech state due to false-alarms, all frames in the transition phase need to indicate speech, and transitions from speech to non-speech states are delayed, i.e. if noise is indicated, another transition phase prevents speech misses. Motivated by [DNT06], we refer to 3 and 8 frame states for false-alarm and miss VAD transition phases, respectively.

**MUG** GMM-based VADs are motivated due to the poor performance of energy-based VADs in low-SNR scenarios e.g., on 0 dB SNR, speech and

noise energies are equal, effectively leading to random VAD decisions. Since USG GMMs are self-adaptive to the current speech sample, energy-based selection may result in inadequate-representative training segments, especially in the presence of high-energy noise impulses, such as closing doors or moving nearby objects[4]. Thus, we propose to utilize the lowest and highest MDFCs instead of energy values for initializing speech and non-speech GMMs, respectively. MDFC values are smoothed by a 3-frame sliding window moving averaging filter before sorting in order to exclude impulsive short-time noises from speech GMM training. This paper refers to the proposed extension as MDFC-based unsupervised GMM (MUG) VAD.

### 2.2 VAD Metrics in Speech and Speaker Recognition

In speech recognition, VAD metrics represent how much verbalized context is missed in contrast to how many false-alarms occur in terms of non-speech that is forwarded to ASR systems. Conventional VAD metrics [GS85, Fr89, DNT06, KEWP11] are computed as:

- the front-end clipping (FCE): $\text{FCE} = \frac{N_{front-miss}}{N_{got-speech}}$, (1)
- the middle-speech clipping (MSC): $\text{MSC} = \frac{N_{mid-miss}}{N_{got-speech}}$, (2)
- the non-speech over-hang (OVER): $\text{OVER} = \frac{N_{over-fa}}{N_{got-non}}$, (3)
- the noise detected as speech (NDS): $\text{NDS} = \frac{N_{fa}}{N_{got-non}}$, (4)
- the speech, non-speech, and average hit rates (SHR, NHR, AHR):
  $\text{SHR} = \frac{N_{speech-hits}}{N_{got-speech}}, \text{NHR} = \frac{N_{non-hits}}{N_{got-non}} = 1 - \text{NDS}, \text{AHR} = \frac{1}{2}(\text{SHR} + \text{NHR})$, (5)

with the speech misses in the beginning of an utterance $N_{front-miss}$, during utterances $N_{mid-miss}$, the false alarms of non-speech frames after an utterance and in a sample $N_{over-fa}, N_{fa}$, the amounts of correct speech and non-speech decisions $N_{speech-hits}, N_{non-hits}$, and the ground-of-truth amounts of speech and non-speech frames $N_{got-speech}, N_{got-non}$, requring frame-wise VAD annotated datasets.

In speaker recognition, VAD effects are mostly reported regarding their effect to the biometric and decision performance e.g., in terms of the equal-error rate (EER), NIST SRE decision cost functions (DCFs) or the goodness of LLRs $C_{llr}$ [SS12, KR13, Al14, MG15]. Due to GMM and factor analysis based architectures in state-of-the-art speaker recognition [RQD00, Ke05, De11], contextual information is accumulated, i.e. VAD is relevant to reject speech segments as little as possible in order to estimate higher-certainty Baum-Welch statistics, without regard to a segment's context, in which a segment is omitted (missed) or

[4] Several MOBIO [Mc12] samples comprise short-time noises at the capture start, which may occur due to e.g., doors, chairs or pressing a start recording button.

falsely included (false alarms). Thus, FCE, MSC and OVER are less relevant for the biometric VAD performance assessment, however these metrics remain useful for developing VADs, such that for example FCE and OVER reflect the gains from a two-way hangover scheme, and MSC the benefits of smoothings. Furthermore, AHR equally accounts for SHR and NHR, which is not necessarily optimal from the perspective of retaining speech segments for discriminative biometric recognition, especially if SHR and NHR diverge significantly.

## 3 VAD Decision Performance in unconstrained Environments

Targeting VAD performance assessment for unconstrained mobile environments, VAD decisions shall remain stable under changing conditions impacting sample quality, such as varying background noises stemming from different sources. Since VAD decisions are binary, i.e. speech or non-speech, and environmental effects are conventionally examined in certain levels or steps, such as 0 dB, 5 dB, ..., 20 dB and *clean*, of a quality-adapted undistorted *clean* database, effects on VAD decisions under different environments can be thought of as binary sequences, which have arbitrary but fixed length for each voice sample as depicted in Fig. 1. Given optimal-condition e.g., *clean*, and synthetic-distorted samples, each binary VAD decision sequence stemming from distorted signals can be XOR-compared to the clean, and reported in terms of the average Hamming distance $\overline{d}$ depicting the conditional VAD decision error rate for one sample. In order to report VAD decision robustness, we propose the database-average conditional VAD decision error $\emptyset_{\overline{d}}$ (VDE). Other statistic moments, such as variance, skewness and kurtosis, can aid a $\overline{d}$-distributional summary and VAD development processes, but are not included in further steps for the sake of easy tractability.

| condition | VAD decision examples on 16 segments | |
|---|---|---|
| *clean* | 0 0 0 1 1 1 1 0 0 0 1 1 1 1 0 0 | |
| 20 dB | 0 0 0 **0** 1 1 1 0 0 0 1 1 1 1 0 0 | XOR: 1, $\overline{d} = \frac{1}{16}$ |
| 15 dB | 0 0 0 **0** 1 1 1 **1** 0 0 1 1 **0** 1 0 0 | XOR: 3, $\overline{d} = \frac{3}{16}$ |
| 10 dB | 0 0 0 **0** 1 1 **0** 0 0 **1** 1 1 **0** **0** 0 0 | XOR: 5, $\overline{d} = \frac{5}{16}$ |
| 5 dB | **1** **1** 0 **0** 1 1 **0** 0 0 **1** 1 **0** **0** 1 **1** 0 | XOR: 8, $\overline{d} = \frac{8}{16}$ |
| 0 dB | **1** **1** **1** **0** **0** **0** **0** **1** **1** **1** **0** **0** **0** **0** **1** **1** | XOR: 16, $\overline{d} = \frac{16}{16}$ |

Fig. 1: VAD decision example under changing environmental conditions with segment-wise VAD votes as *speech* (1) and *non-speech* (0), where *clean* denotes the original sample of good quality.

## 4 Evaluation in mobile Speaker Recognition Scenario

Experiments are carried out on the MOBIO SRE'13 task [Kh13]. A standard speaker recognition front-end is used based on *rastamat* [El05] and *jfacookbook* [Gl09]: 60-dimensional speech signal features based on 19 MFCCs with log-Energy and derived $\Delta$

and $\Delta\Delta$ coefficients on a standard hamming window. Feature warping [PS01] is applied using a $3\,s$ sliding window, and for non-speech features removal, the in Sec. 2 depicted VADs were utilized. Raw i-vectors are extracted with 400 dimensions based on a 512-component UBM. Due to the limited data in MOBIO SRE'13, gender-independent systems are trained based on a state-of-the-art PLDA comparator [GREW11]: i-vectors are projected into a 49-dimensional spherical unit subspace by LDA, mean-subtraction, within class covariance normalization (WCCN) [HKS06] rotation, and length-normalization [GREW11]. Reference and probe i-vectors are compared by PLDA [GREW11] in full-subspace to obtain (uncalibrated) LLR scores.

### 4.1 Database Description: MOBIO SRE'13

The speaker recognition subset of the MOBIO database [Mc12, Kh13] was recorded on mobile phones and laptops, however in the MOBIO SRE'13 [Kh13] only data from mobile phones was used. Tab. 1 depicts the amount of speakers and samples for the backround and development (dev-set) containing 50 and 42 subjects in total, respectively. Experiments are solely conducted on the dev-set, in order to prevent data snooping effects for other research on the MOBIO database targeting the MOBIO evaluation set. Due to the small partition of female speakers during PLDA training, results are solely reported w.r.t. data of male speakers.

| Set | Female | | Male | |
|---|---|---|---|---|
| | Subjects | Samples | Subjects | Samples |
| Background | 13 | 2496 | 37 | 7104 |
| dev-set (references) | 18 | 90 | 24 | 120 |
| dev-set (probes) | 18 | 1890 | 24 | 2520 |

Tab. 1: Partitioning of MOBIO database, see [Kh13].

### 4.2 Evaluation Criteria

The biometric performance is reported in accordance to the ISO/IEC IS 19795-1 [IS11] by the Equal-Error-Rate (EER), and the False Non-Match Rate (FNMR) at a 1% False Match Rate (FMR100). As an application-independent performance metric, we emphasize on the minimum cost of log-likelihood ratio (LLR) scores $C_{\text{llr}}^{\text{min}}$, which represents the discrimination power as generalized empirical cross-entropy of genuine and impostor LLRs with respect to Bayesian thresholds assuming well-calibrated systems [BdP06, RGR08].

### 4.3 Experimental Results

Analyses are conducted regarding the VAD decision robustness in noisy conditions of different SNR levels, and the coherence of VAD metrics in terms of sensitivity to the eval-

uation criteria. The performance of baseline speaker recognition systems utilizing VADs are compared in Tab. 2. In terms of EER and $C_{llr}^{min}$, VFR outperforms the other algorithms, while the proposed MUG VAD yields a better FMR100 performance. The performance gain of baseline systems to no VAD segmentation applied is moderate, since the MOBIO task comprises rather prompted speech instead of phone calls, i.e. samples are pre-segmented due to the prompted scenario.

| VAD | P.56 | VBX | SRT | VFR | PSA | USG | MUG | no VAD |
|---|---|---|---|---|---|---|---|---|
| EER (in %) | 11.0 | 10.9 | 12.2 | **10.2** | 10.9 | 11.0 | 10.7 | 11.9 |
| FMR100 (in %) | 42.4 | 41.1 | 45.9 | 40.0 | 43.7 | 41.7 | **39.6** | 46.7 |
| $C_{llr}^{min}$ | 0.377 | 0.376 | 0.376 | **0.355** | 0.373 | 0.377 | 0.361 | 0.407 |

Tab. 2: VAD algorithm performance comparison to no VAD applied by EER, FMR100, and $C_{llr}^{min}$ on male speaker subset of the MOBIO dev-set, i.e. in this work referred to as clean condition.

### 4.3.1 VAD Decision Robustness

In order to analyze the impact of noise conditions (source types and SNR levels) to VAD and biometric recognition performance, pink, white, babbel and street noise are examined in 0 dB, 5 dB, ..., 20 dB SNR levels utilizing the Matlab implementations of [Br05, ZB14, Ly15]. Pink noise is referred to be ubiquitous in many biological and physical systems [BTW87], white (Gaussian) noise represents random signals, babbel noise is conducted utilizing all speakers of the MOBIO background set with random sample selection, street noise is stemming from the QUT-NOISE-TIMIT corpus [De10], which is explictely designed for the purpose of evaluating VAD performance.

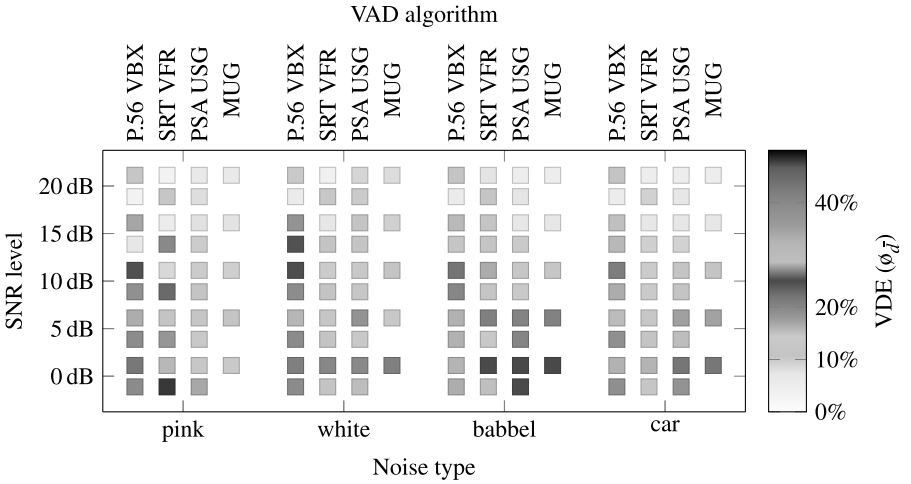


Fig. 2: VAD decision robustness carried out under noisy samples compared to clean samples on dev-set by the proposed VDE metric $\phi_{\bar{d}}$.

Fig. 2 depicts the robustness of VAD algorithms introduced in Sec. 2 regarding the proposed VDE metric, representing the average rate of misconducted VAD segmentation votes. For the majority of VAD algorithms, speech decisions on 15 dB and 20 dB conditions are similarly to the clean condition. On 0 dB and 5 dB, SRT yields the most stable decisions on white, babbel and car noise, and MUG on pink noise. Regarding 10 dB to 20 dB, VFR outperforms other VADs on pink, white, and street noises, while for babble noise, USG and SRT yield more stable VAD decisions on high-SNRs and 10 dB, respectively. Examining SRE-related VADs, the proposed MUG VAD outperforms PSA and USG on pink, white and street noise in 0 dB to 15 dB conditions, and on 20 dB pink and white noise. In other conditions, USG outperformed PSA. On condition-averaged VDE, MUG yields 0.120, USG 0.129, and PSA 0.130, where VFR and SRT yield 0.113 and 0.157, respectively.

### 4.3.2 Sensitivity Coherence: VAD to biometric Recognition Metrics

Sensitivity analyses are conducted in order to provide insights on coherence of the proposed VDE metric to biometric and forensic performance. For tractability purposes, noise conditions are pooled by SNR level and the SNR of clean samples is assumed to be 25 dB. Fig. 3 depicts the SNR sensitivity of the SRT, VFR, USG and MUG algorithms. SRT achieves low sensitivity regarding EER and FMR100, though SRT yields the highest EER and FMR100 results among all examined VADs, cf. Tab. 2. In terms of the proposed VDE metric and FMR100, VFR, USG and MUG VADs perform similarly, however regarding $C_{llr}^{min}$ and EER, USG and MUG result in more stable performance than VFR, especially in the low-SNR region with average EER sensitivitiy of 0.76% and 0.80% in EER per 1 dB SNR.

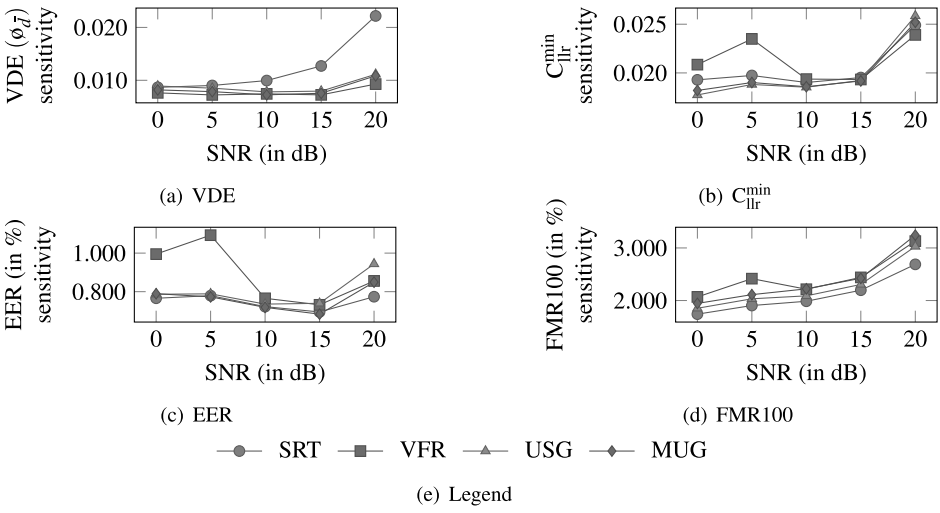


Fig. 3: Sensitivity of VAD performances to different SNR levels of SRT, VFR, USG and MUG approaches by (a) VDE, (b) $C_{llr}^{min}$, (c) EER, (d) FMR100.

### 4.4 Summary and Discussion

VADs are designed for certain applications and target specific environmental constraints, such that none of the examined algorithms is able to outperform other approaches regarding each analyses. NIST SRE-motivated VADs yield more stable segmentation decisions in high-SNR conditions than conventional VADs. However, examining low-SNR conditions, conventional and SRE VADs achieve good performance, in particular: VFR, PSA, USG and MUG. In contrast, SRT achieves better sensitivity results, but has shortcomings in VDE on pink noise conditions and also high biometric error rates on clean condition. Regarding evaluation criteria sensitivity, USG and MUG yield the least SNR-sensitive results, which is coherent to the conducted VAD decision sensitivity analysis. The proposed MUG outperforms other SRE VADs by utilizing benefitial MDFC-features from conventional VADs, where SRT and VFR partially achieve gains by employing SFM, a-posteriori SNR and pitch features in low-SNR white, babbel and car noise conditions.

## 5 Conclusion

The proposed VDE metric reveals the stability of VAD segmentation decisions under different noise conditions. Contrary to well-established metrics in speech recognition, the proposed metric examines the average amount of inconsistent VAD decisions on changing environmental conditions, emphasizing on *where* speech frames are falsely recognized. By conducting the proposed analyses recipe for examining the decision robustness and evaluation criteria sensitvity of VADs, coherent decisions can be made regarding the applicability of VAD segmentation algorithms to speaker recognition tasks. The proposed metric has limitations regarding the location of false segmentation decisions, which can be examined by conventional VAD metrics. However, decision robustness is more valuable to state-of-the-art speaker recognition methods, in which speech frame statistics are accumulated, i.e. the location of VAD errors remain without impact, whereas unstable VAD decisions lead to different frame samplings forwarded to front- and back-end processing. Contrary to conventional VAD metrics, frame-wise annotation voice samples is not required in order to measure VAD performance. Furthermore, the proposed MUG-extension of the USG approach yields promising gains, which are expected to be more extended by incorporating SFM, a-posteriori SNR and pitch features into the VAD decision process.

## Acknowledgement

We like to thank the MOBIO consortium for database sharing. This work has been partially funded by the Center for Advanced Security Research Darmstadt (CASED), and the Hesse government (project no. 467/15-09, BioMobile).

## References

[Al14] Alam, Md.J.; Kenny, P.; Ouellet, P.; Stafylakis, T.; Dumouche, P.: Supervised/Unsupervised Voice Activity Detectors for Text-dependent Speaker Recognition on the

RSR2015 Corpus. In: Proc. Odyssey 2014: The Speaker and Language Recognition Workshop. pp. 123–130, 2014.

[BdP06] Brümmer, N.; du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech & Language, 20(2–3):230–275, 2006. Odyssey 2004: The Speaker and Language Recognition Workshop.

[Br05] Brookes, M.: , VOICEBOX: Speech Processing Toolbox for MATLAB. Department of Electrical & Electronic Engineering, Imperial College, London, 2005.

[BTW87] Bak, P.; Tang, C.; Wiesenfeld, K.: Self-Organized Criticality: An Explanation of 1/f Noise. APS Physical Review Letters, 59(4):381–384, 1987.

[De10] Dean, D.B.; Sridharam, S.; Vogt, R.; Mason, M.W.: The QUT-NOISE-TIMIT corpus for the evaluation of voice activity. In: Proc. Interspeech. pp. 3110–3113, 2010.

[De11] Dehak, N.; Kenny, P.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-End Factor Analysis For Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing (TASLP), 19(4):788–798, 2011.

[DNT06] Davis, A.; Nordholm, S.; Togneri, R.: Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold. IEEE Transactions on Audio, Speech, and Language Processing (TASLP), 14(2):412–424, 2006.

[El05] Ellis, Daniel P. W.: , PLP and RASTA (and MFCC, and inversion) in Matlab. `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`, 2005. [Online; accessed 2013-10-10].

[Fr89] Freeman, D.K.; Cosier, G.; Southcott, C.B.; Boyd, I.: The voice activity detector for the Pan-European digital cellular mobile telephone service. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 369–372, 1989.

[Gl09] Glembek, O.: , Joint Factor Analysis Matlab Demo. `http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo`, 2009. [Online; accessed 2013-10-10].

[GREW11] Garcia-Romero, D.; Espy-Wilson, C.Y.: Analysis of i-vector Length Normalization in Speaker Recognition Systems. In: Proc. Interspeech. pp. 249–252, 2011.

[GS85] Gruber, J.G.; Strawczynski, L.: Subjective Effects of Variable Delay and Speech Clipping in Dynamically Managed Voice Systems. IEEE Transactions on Communications, 33(8):801–808, 1985.

[HKS06] Hatch, A.O.; Kajarekar, S.; Stolcke, A.: In: Proc. Interspeech 2006, International Conference on Spoken Language Processing (ICSLP). pp. 1471–1474, 2006.

[IS11] ISO/IEC: Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. ISO/IEC 19795-1:2006, JTC 1/SC 37, Geneva, Switzerland, 2011.

[IT94] ITU-T: Recommendation ITU-T P. 52 (1993), Volume meters. ITU-T P.56, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 1994.

[Ke05] Kenny, P.: Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Technical report, Centre de recherche informatique de Montréal (CRIM), 2005.

[KEWP11] Kola, J.; Epsy-Wilson, C.; Pruthi, T.: Voice Activity Detection. MERIT BIEN final report, University of Maryland, Department of Electrical & Computer Engineering, College Park, Maryland, USA, 2011.

[Kh13] Khoury, E.; Vesnicer, B.; Franco-Pedroso, J.; Violato, R.; Boulkenafet et al., Z.: The 2013 Speaker Recognition Evaluation in Mobile Environment. In: Proc. IAPR International Conference on Biometrics (ICB). pp. 1 – 8, 2013.

[KR13] Kinnunen, T.; Rajan, P.: A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 7229–7233, 2013.

[KS99] Kim, J. SohnN.S.; Sung, W.: A Statistical Model-Based Voice Activity Detection. IEEE Signal Processing Letters, 6(1):1–3, 1999.

[Ly15] Lyons, J.: , matlab speech feature generation scripts used internally in the Signal Processing Lab at Griffith University. `https://github.com/jameslyons/spl_featgen`, 2015. [Online; accessed 2016-05-26].

[Ma01] Martin, R.: Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. IEEE Transactions on Speech and Audio Processing, 9(5):504–512, 2001.

[Mc12] McCool, C.; Marcel, S.; Hadid, A.; Pietikainen, M.; Matejka, P.; Cernocky, J.; Poh, N.; Kittler, J.; Larcher, A.; Levy, C.; Matrouf, D.; Bonastre, J.-F.; Tresadern, P.; Cootes, T.: Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data. In: Proc. IEEE ICME Workshop on Hot Topics in Mobile Multimedia. pp. 635–640, 2012.

[MG15] McLaren, M.; Graciarena, M.: softSAD: Integrated frame-based speech confidence for speaker recognition. In: Proc. Int. Conf. on Acoustic, Speech and Signal Processing (ICASSP). pp. 4694–4698, 2015.

[MH09] Moattar, M.H.; Homayounpour, M.M.: A simple but efficient real-time Voice Activity Detection algorithm. In: Proc. European Signal Processing Conference (EURASIP). pp. 2549–2553, 2009.

[PS01] Pelecanos, J.; Sridharan, S.: Feature Warping for Robust Speaker Verification. In: Proc. Odyssey 2001: The Speaker and Language Recognition Workshop. 2001.

[RGR08] Ramos, D.; Gonzalez-Rodriguez, J.: Cross-entropy Analysis of the Information in Forensic Speaker Recognition. In: Proc. Odyssey 2008: The Speaker and Language Recognition Workshop. 2008.

[RQD00] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. In: Conversational Speech, Digital Signal Processing. volume 10, pp. 19–41, 2000.

[SS12] Sahidullah, Md.; Saha, G.: , Comparison of Speech Activity Detection Techniques for Speaker Recognition. arXiv:1210.0297, 2012.

[TL10] Tan, Z.H.; Lindberg, B.: Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection. IEEE Journal of Selected Topics in Signal Processing, 4(5):798–807, 2010.

[ZB14] Zhivomirov, H.; Baranski, P.: , Pink, Red, Blue and Violet Noise Generation with Matlab Implementation. `http://www.mathworks.com/matlabcentral/fileexchange/42919-pink--red--blue-and-violet-noise-generation-with-matlab-implementation`, 2014. [Online; accessed 2016-05-26].