SPEECH-CODEBOOK BASED SOFT VOICE ACTIVITY DETECTION

Florian Heese, Markus Niermann, and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{heese,niermann,vary}@ind.rwth-aachen.de

ABSTRACT

A novel noise-robust soft Voice Activity Detector (VAD) operating in the short-time Fourier domain is presented. A speech energy gain is obtained by frame-wise processing of a noisy speech signal with a speech codebook algorithm. This gain can be used for robust voice detection. A speaker-independent speech codebook, consisting of spectral envelopes, is created in the training process. While applying the algorithm, the codebook is adapted in every frame to the current speaker by combining the harmonic pitch structure of the actual noisy speech frame with the codebook entries. Soft VAD values ranging from zero to one are calculated by post-processing of the speech gain which is obtained using gain shape vector quantization. A binary VAD is carried out by applying a threshold. The proposed method does not rely on noise *a-priori* knowledge and is robust w.r.t. highly non-stationary noise and adverse SNR conditions. In addition, it is possible to compromise between the detection-rate and the false-alarm-rate by varying a threshold without increasing the total number of mis-detections. Compared to state-of-the-art VAD systems, the proposed method is characterized by better detection-rates at significant lower false-alarm-rates.

Index Terms— Voice activity detection, Codebook, Noise robust

1. INTRODUCTION

The rapid progress of digital and mobile speech communication enables audio-visual communication from almost anywhere in the world. As a consequence, voice is often transmitted from acoustically disturbed environments. The objective of a Voice Activity Detector (VAD) is to detect the presence or absence of human speech in a microphone signal which might be degraded by background noise. Algorithms such as noise and echo control as well as speech coding and speech recognition are often supported by a robust VAD. In a video conference a joint speaker dependent VAD and a video face tracking enables new applications such as an artificial scene composition where the audio signals and the active speakers are emphasized.

1.1. Relation to prior work

Early VAD systems extract simple energy features such as SNR estimations, that respond while speech is present, and compare the quantified values to a fixed or adaptive threshold for a VAD decision, e.g., [1, 2, 3]. In the GSM cellular radio system the VAD [4] is basically an energy detector whose accurary is improved by adaptive filtering to increase the speech-noise ratio. Since the encountered noise in mobile environments may be constantly changing with time and frequency the adaptive filter is only updated when speech is absent, the signal seems stationary, and does not include a pitch component which is inherent in voiced speech.

However, energy based techniques do not work reliably under adverse acoustic conditions, e.g., at signal-to-noise ratios of 0 dB or below. Recent systems mainly employ statistical models, also including additional features like the zero crossing rate, pitch, tone, complex-signal correlation, and the energy levels of frequency bands [5, 6, 7, 8]. Adding more microphones, the voice activity detection accuracy can be improved (see, e.g., [9, 10]). All these approaches cope with moderate, mainly stationary noise. However, for many applications, they are not sufficiently robust with respect to highly non-stationary noise.

Sohn [6] proposed a likelihood ratio test, combined with a markov process, that models speech occurrences in order to obtain a VAD. Cho [7] analyzes this method and improves some fundamental problems at speech offset regions using a smoothed likelihood ratio for the adaptation of the noise variance, resulting in an improved decision of voice activity. Tan [11] employs a likelihood ratio test and modifies the handling of voiced frames by selecting exclusively the harmonic components for computing. Ghosh [8] introduces a "long-term signal variability measure", which represents the degree of non-stationarity. Combined with the assumption that speech is significantly less stationary than noise, this measure discriminates between noise and noisy speech, resulting in a robust VAD performance.

In this contribution, a new approach is proposed that uses a speech codebook as *a-priori* knowledge similar to our speech enhancement approach [12]. Acoustically degraded signals are frame-wise compared with the speech codebook in order to determine a similarity measure between the input signal and typical spectral speech compositions. This new technique is robust to highly non-stationary noises and reliably detects speech also in adverse SNR conditions of -5 dB. Since the speech codebook is designed speaker-independently and as we do not rely on a noise codebook, the algorithm is not restricted to known speakers or known noise types.

2. SIGNAL MODEL

It is assumed that the noisy input signal $x(k)$ consists of clean speech $s(k)$ degraded by an additive noise $n(k)$ according to:

$$x(k) = s(k) + n(k), \tag{1}$$

where k is the discrete time index. The samples $x(k)$ are obtained by analog-digital conversion with a sampling frequency of $f_s = 16$ kHz. Since the proposed VAD algorithm is performed in the frequency domain, $x(k)$ is segmented into 50 % overlapping frames of length L_F, followed by windowing (square root Hann-window) and zero-padding. Subsequently, each frame is transformed by applying the Fast Fourier Transform (FFT) of length $M \geq L_F$. The spectral coefficients of the input signal $x(k)$ at frequency bin μ and frame λ are given by:

$$X(\lambda, \mu) = S(\lambda, \mu) + N(\lambda, \mu), \tag{2}$$

where $S(\lambda,\mu)$ and $N(\lambda,\mu)$ correspond to the spectral coefficients of the clean speech signal and the noise signal. The proposed VAD algorithm operates on the short-term energy spectrum (STES) $|X(\lambda,\mu)|^2$.

3. PROPOSED VAD ALGORITHM

This section describes the training of the speaker-independent codebook (Sec. 3.1) and the codebook matching algorithm (Sec. 3.3), which provides a robust voice indicator. The soft VAD is obtained by further post-processing. The codebook is adapted in every frame to the current speaker by incorporating the pitch from the actual degraded speech signal using a cepstral approach (Sec. 3.2).

3.1. Codebook Creation

The codebook is trained with speech files from several speakers. The training sequence is transformed into frames of STES according to Sec. 2. All frames with an energy below a threshold are discarded. This removes on the one hand silent parts of the training data which may be over-represented in the codebook training. On the other hand, it eliminates frames with mere recording noise. To make the codebook speaker-independent, the speaker-dependent excitation is removed by using the cepstral approach described in Sec. 3.2. Afterwards, the frames are normalized to an energy of one.

The LBG algorithm [13] with the Itakuro Saito distance

$$d_{\mathrm{IS}}(A(\mu),B(\mu)) = \sum_{\mu=0}^{M-1}\left[\frac{A(\mu)}{B(\mu)} - \log\frac{A(\mu)}{B(\mu)} - 1\right], \quad (3)$$

as distance measure is used for the codebook training, with A and B as placeholders for two STESs. Finally, a codebook with N_{cb} entries, consisting of STESs $|E_l(\mu)|^2$ with the entry indices $l = 1, ..., N_{\mathrm{cb}}$, is created. Each STES is normalized to an energy of one.

3.2. Cepstral Processing

Due to the source-filter model [1], human speech is composed of a spectral envelope and its excitation. The speaker-dependent pitch frequency f_{p} of the excitation is assumed to be in the range between 50 Hz and 500 Hz [1]. A cepstral approach, like in [14], is applied to separate the spectral envelope and the excitation. Therefore, the STESs $|X(\lambda,\mu)|^2$ are frame-wise transformed to the cepstral domain:

$$c_{|X(\lambda)|^2}(q) = \frac{1}{2}\sum_{\mu=0}^{M-1}\log\left(|X(\lambda,\mu)|^2 \mathrm{e}^{j2\pi\frac{\mu q}{M}}\right), q=0,...,M-1. \quad (4)$$

A pitch frequency of f_{p} is represented in the cepstrum as a peak in the cepstral bin $\left\lfloor \frac{f_{\mathrm{s}}}{f_{\mathrm{p}}} \right\rfloor$ (e.g., [14, 15]). Assuming that pitch frequencies are bounded to be lower than 500 Hz and considering the symmetry of the cepstral coefficients, the range $q_p < q < M - q_p$ with $q_p = \left\lfloor \frac{f_{\mathrm{s}}}{f_{\mathrm{p}}} \right\rfloor$ is called the excitation part in the following.

Before generating the codebook, the speaker-dependent excitation is removed from the training sequence by setting the corresponding cepstral coefficients to zero:

$$c_{\left|\tilde{X}(\lambda)\right|^2}(q) = \begin{cases} 0 & q_c < q < M - q_c \\ c_{|X(\lambda)|^2}(q) & \text{else.} \end{cases} \quad (5)$$

Afterwards, the modified cepstrum $c_{\left|\tilde{X}(\lambda)\right|^2}$ is transformed back to the spectral domain:

$$\left|\tilde{X}(\lambda,\mu)\right|^2 = \exp\left(2\cdot\sum_{q=0}^{M-1} c_{\left|\tilde{X}(\lambda)\right|^2}(q) e^{-j\frac{2\pi}{M}\mu q}\right). \quad (6)$$

3.3. Codebook Matching

The concept of codebook matching is to compare the noisy speech signal frame-wise with the speech codebook entries in order to find the entry $\left|E_{l_{\mathrm{opt}}}(\mu)\right|^2$ which fits best the current noisy frame. In a second step, the speech gain is determined to scale $\left|E_{l_{\mathrm{opt}}}(\mu)\right|^2$ to the correct energy. Since the speaker-independent codebook, in contrast to the noisy frames, contains only spectral envelopes, its harmonic structure has to be re-established. The goal is a comb-like structure whose pitch frequency equals the one of the current input speech frame. This is realized by means of a cepstral approach, i.e. the excitation part from the noisy STES $|X(\lambda,\mu)|^2$ is extracted and incorporated into each codebook entry $|E_l(\mu)|^2$. This procedure is repeated for each input frame. The cepstral representation $c_{|E_l|^2}(q)$ of the codebook entries is calculated analogously to Eq. (4) and $c_{|X(\lambda)|^2}(q)$ is the cepstrum of the noisy speech signal. The envelope from $c_{|E_l|^2}(q)$ and the pitch from $c_{|X(\lambda)|^2}(q)$ are combined according to:

$$c_{\left|\tilde{E}_l(\lambda)\right|^2}(q) = \begin{cases} c_{|X(\lambda)|^2}(q) & q_c < q < M - q_c \\ c_{|E_l|^2}(q) & \text{else,} \end{cases} \quad (7)$$

transformed to the spectral representation analogously to Eq. 6 and normalized to an energy of one. The result $\left|\tilde{E}_l(\lambda,\mu)\right|^2$ is a codebook entry which is adapted to the current speaker, i.e. with a corresponding harmonic frequency structure.

In the next step, the index l_{opt} of the best fitting codebook entry is determined by

$$l_{\mathrm{opt}}(\lambda) = \arg\min_{l} \operatorname*{dist}_{\mu}\left(|X_E(\lambda,\mu)|^2, \left|\tilde{E}_l(\lambda,\mu)\right|^2\right), \quad (8)$$

where

$$|X_E(\lambda,\mu)|^2 = \frac{1}{\sum_{\mu'=0}^{M-1}|X(\lambda,\mu')|^2}|X(\lambda,\mu)|^2 \quad (9)$$

is the energy-normalized noisy STES, and dist refers to the relative energy distance,

$$\operatorname*{dist}_{\mu}\left(|X_E(\mu)|^2, |\tilde{E}_l(\mu)|^2\right) = \frac{\sum_{\mu=0}^{M-1}\left||X_E(\mu)|^2 - |\tilde{E}_l(\mu)|^2\right|}{\sum_{\mu=0}^{M-1}|X_E(\mu)|^2}, \quad (10)$$

which turned out to be the best metric. Using the index of the best codebook entry, $l_{\mathrm{opt}}(\lambda)$, a speech gain σ^2_{vad} is calculated,

$$\sigma^2_{\mathrm{vad}}(\lambda) = \arg\min_{\sigma^2} \operatorname*{dist}_{\mu}\left(|X(\lambda,\mu)|^2, \sigma^2\left|\tilde{E}_{l_{\mathrm{opt}}(\lambda)}(\lambda,\mu)\right|^2\right), \quad (11)$$

where the possible gains are equally distributed in N_σ steps:

$$\sigma^2 \in \left\{\frac{i}{N_\sigma - 1}\cdot\sum_{\mu'=0}^{M-1}\left|X(\lambda,\mu')\right|^2 \,\middle|\, i = 0,...,N_\sigma - 1\right\}. \quad (12)$$

While speech is present, a suitable codebook entry and a gain σ^2_{vad} close to the speech frame energy can be found. In turn, during speech pauses, no suitable codebook-entry is available in general. Thus, the spectral envelopes of the codebook entry and the normalized noise frame differ significantly. The resulting relative energy distance is very high and in general greater than the distance of $|X|^2$ to zero. Therefore, a small gain minimizes the distance measure. Thus, the speech gain σ^2_{vad} is used as speech presence indicator. Since spectral overlaps while speech absence between X and $\tilde{E}_{l_{\mathrm{opt}}}$ cannot be excluded, a noise floor in the gain is observed. Further post-processing is necessary to obtain a reliable VAD.

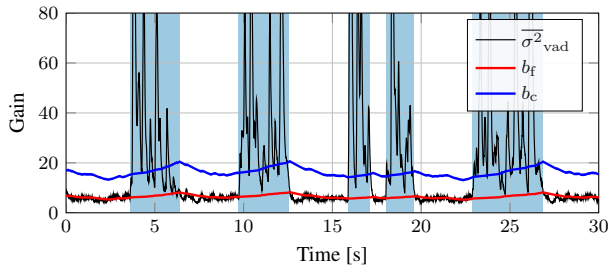

Fig. 1. Example of smoothed gain $\overline{\sigma^2}_{\text{vad}}$, floor estimation b_{f} and ceiling estimation b_{c} for male speech and jackhammer noise (SNR = 5 dB). A blue background indicates true speech activity.

3.4. Speech gain σ^2 post-processing

In a first step of the post-processing, the speech gain σ^2_{vad} is smoothed recursively by:

$$\overline{\sigma^2}_{\text{vad}}(\lambda) = \left[\alpha\sqrt{\overline{\sigma^2}_{\text{vad}}(\lambda-1)} + (1-\alpha)\sqrt{\sigma^2_{\text{vad}}(\lambda)}\right]^2. \quad (13)$$

The smoothing parameter $0 < \alpha < 1$ determines the smoothing intensity and is chosen differently for rising or falling values in order to control on- and offset of voice activity differently:

$$\alpha = \begin{cases} \alpha_+ & \sigma^2_{\text{vad}}(\lambda) \geq \bar{\sigma}^2_{\text{vad}}(\lambda-1) \\ \alpha_- & \sigma^2_{\text{vad}}(\lambda) < \bar{\sigma}^2_{\text{vad}}(\lambda-1). \end{cases} \quad (14)$$

The smoothed sequence $\overline{\sigma^2}_{\text{vad}}(\lambda)$ is a reliable speech presence indicator with a range of values in $[0, \infty)$. An example of this indicator is given in Fig. 1. It shows a noise floor during speech pauses and considerably higher levels during speech presence. However, soft VAD-values between zero and one are desired, which requires a mapping of $\overline{\sigma^2}_{\text{vad}}$ to a range between zero and one (see Eq. 17). Therefore, a baseline tracing of the noise floor $b_{\text{f}}(\lambda)$ and ceiling $b_{\text{c}}(\lambda)$ is employed. This method is similar to noise estimation known from speech enhancement [16, 17]. The noise floor tracing is implemented according to:

$$b_{\text{f}}(\lambda) = b_{\text{f}}(\lambda-1) + \text{sign}\left[\overline{\sigma^2}_{\text{vad}}(\lambda) - b_{\text{f}}(\lambda-1)\right]\Delta'(\lambda) \quad (15)$$

$$b_{\text{c}}(\lambda) = \max\left[\eta \cdot b_{\text{f}}(\lambda), b_{c,min}\right], \quad (16)$$

where $b_{c,min}$ defines a minimum value for the ceiling $b_{\text{c}}(\lambda)$ and the factor η controls the upper clipping of $\overline{\sigma^2}_{\text{vad}}$. In each frame, the noise floor $b_{\text{f}}(\lambda)$ is updated by shifting $\pm\Delta'(\lambda)$ in order to follow $\overline{\sigma^2}_{\text{vad}}(\lambda)$ slowly. Finally, soft VAD values for $\overline{\sigma^2}_{\text{vad}}$ between b_{f} and b_{c} are interpolated linearly according to

$$v_{\text{soft}}(\lambda) = \max\left(\min\left(\frac{\overline{\sigma^2}_{\text{vad}}(\lambda) - b_{\text{f}}(\lambda)}{b_{\text{c}}(\lambda) - b_{\text{f}}(\lambda)}, 1\right), 0\right). \quad (17)$$

Gains lower or equal to the noise floor are mapped to zero, whereas gains higher or equal to the ceiling $b_{\text{c}}(\lambda)$ are mapped to one. The resulting soft values are robust to different noise floor levels in the speech gain which may result from low SNR and varying noise types.

In order to be independent of the sampling frequency f_{s} and the frame advance L_{A}, a relative shift Δ is introduced with dimension $\frac{\%}{time}$ such that $\frac{L_{\text{A}}}{f_{\text{s}}}\Delta$ is the relative change per frame. Moreover, it is desirable to update the noise floor mainly in cases of speech absence, yielding the absolute shift to

$$\Delta'(\lambda) = \begin{cases} \frac{L_A}{f_{\text{s}}} \cdot \Delta \cdot b_{\text{f}}(\lambda-1) & \overline{\sigma^2}_{\text{vad}}(\lambda) \leq b_{\text{c}}(\lambda-1) \\ \frac{L_A}{f_{\text{s}}} \cdot \Delta \cdot b_{\text{f}}(\lambda-1) \cdot \beta_{\text{sp}} & \overline{\sigma^2}_{\text{vad}}(\lambda) > b_{\text{c}}(\lambda-1). \end{cases} \quad (18)$$

Parameter	**Settings**
Sampling frequency f_{s}	16 kHz
Frame length L_{F}	320 ($\hat{=}$ 20 ms)
Frame advance L_{A}	160 ($\hat{=}$ 10 ms)
FFT length M	512 (including zero-padding)
Frame overlap	50% ($\sqrt{\text{Hann-window}}$)
Speech codebook entries N_{cb}	128
Number gains N_σ	10
Smoothing parameter $\alpha_+ \vert \alpha_-$	0.8 | 0.91
Gain ceiling factor η	2.5
Ceiling minimum $b_{c,min}$	3
Relative shift Δ	$0.2\,\text{s}^{-1}$
Speech presence factor β_{sp}	$\frac{1}{4}$

Table 1. Simulation system settings

If the speech gain exceeds the ceiling b_{c}, speech presence is assumed and the tracing speed is reduced by the factor $0 < \beta_{\text{sp}} < 1$. It is not set to zero in order to avoid a status in which the system gets stuck in case that the floor and ceiling estimation are completely wrong. Experiments confirmed that the relative shift over time Δ should be in the range between $\frac{0.2\,\%}{20\,\text{ms}}$ and $\frac{0.8\,\%}{20\,\text{ms}}$, i.e., the noise floor changes by the given percentage during 20 ms, a time in which speech is considered to be stationary [1].

If a binary VAD is desired, it can be calculated by a simple comparison with a threshold $0 < thr < 1$ according to

$$v_{\text{bin}}(\lambda) = \begin{cases} 0 & \text{if } v_{\text{soft}}(\lambda) < thr \\ 1 & \text{if } v_{\text{soft}}(\lambda) \geq thr\ . \end{cases} \quad (19)$$

4. EVALUATION

The proposed VAD system is compared in a benchmark with three reference methods proposed by Sohn [6], Tan [11], Ghosh [8] and GSM VAD [4]. All algorithms except GSM VAD provide soft VAD values. Since the objective scores require a binary VAD, Eq. (19) is applied for different thresholds varying between zero and one.

To evaluate the binary VAD $v_{\text{bin}}(\lambda)$, a reference VAD $v_{\text{true}}(\lambda)$ is necessary. In this simulation, the clean speech and the scaled noise, from which the noisy signal is additively generated, are separately available. The squared magnitude of the clean speech signal is compared sample-wise with a fixed threshold of 10^{-5}, which fits well to the TIMIT database [18]. If the threshold is exceeded at least once during the last 2 ms, speech is assumed ($v_{\text{true}}(\lambda) = 1$).

4.1. Simulation

The parameters for the simulation are given in Tab. 1. The speech codebook is trained according to Sec. 3.1 with randomly chosen speech files from the training set of the TIMIT database [18], resulting in a total training sequence length of 938 s. The configuration of the remaining algorithms are chosen as suggested in [6, 11, 8].

For the benchmark, 24 randomly chosen sentences belonging to 12 male and 12 female, randomly chosen speakers from the test set of the TIMIT database are selected and concatenated. The test set is not included in the training set. Three seconds of silence are inserted at the beginning and the end of the sequence as well as between the sentences. The resulting 160 s speech sequence is mixed with 10 types of noise (white, pink, jackhammer, wind, outside traffic, inside car, train station, nature, inside train, pub noise) from the ETSI database [19] at different SNR values from -5 dB to 20 dB in 5 dB steps, resulting in 60 different noisy signals, respectively 160 minutes. The threshold varies for all tested algorithms in 24 steps from zero to one.

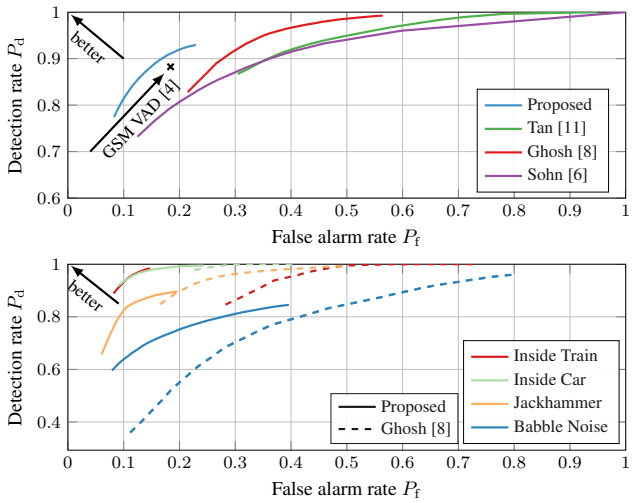

Fig. 2. The upper part depicts the ROC curves for varying thresholds. The ROC curve for 4 exemplary noises is shown in the lower plot for the proposed and 2nd best algorithm Gosh [8] at varing thresholds.

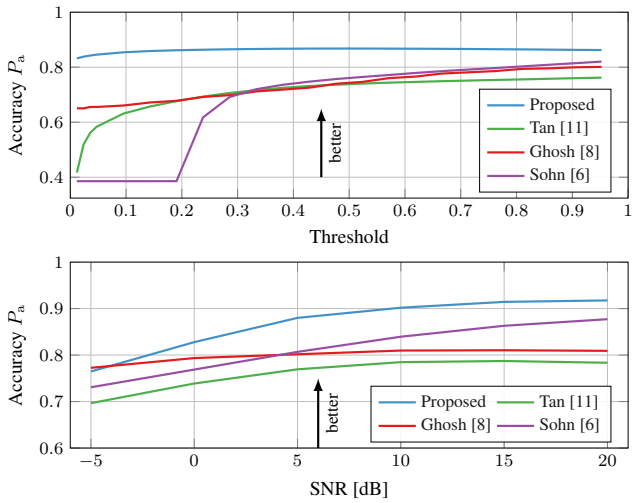

Fig. 3. In the upper plot, the average accuracies for varying thresholds are depicted, while the lower plot shows the average accuracies over the SNR. For each algorithm, the most favorable threshold is chosen.

4.2. Objective Scores

Let Q be the set of frame indices for which holds $v_{\text{true}}(\lambda)=1$ and $\overline{Q}$ the complementary set for $v_{\text{true}}(\lambda)=0$. $K=Q\cup\overline{Q}$ is the set of all frames. The first 160 frames, i.e., 1.6 s, are not included in the evaluation to ignore transient effects. The measures are defined as:

$$P_{\text{a}} = 1 - \frac{1}{|K|}\cdot\sum_{\lambda\in K}\left|v_{\text{bin}}(\lambda) - v_{\text{true}}(\lambda)\right|,$$

$$P_{\text{d}} = \frac{1}{|Q|}\cdot\sum_{\lambda\in Q} v_{\text{bin}}(\lambda),\ P_{\text{f}} = \frac{1}{|\overline{Q}|}\cdot\sum_{\lambda\in\overline{Q}} v_{\text{bin}}(\lambda).$$

The accuracy rate P_{a} is the percentage of frames in which the VAD-estimation is correct. The detection rate (or true positive rate) P_{d} is the fraction of active speech frames that are detected correctly. The false alarm rate (or false positive rate) P_{f} is the fraction of frames without speech that are classified erroneously as speech.

4.3. Results

When applying a VAD, a compromise between detection-rate and false-alarm-rate has to be made by choosing an appropriate threshold (resp. a working point on the receiver operating characteristic (ROC) curve). The upper plot of Fig. 2 shows the ROC curve which is generated by averaging the objective scores (detection rate, false alarm rate) for all permutations of the SNR and noises, separately for each threshold thr. It shows the achievable combinations of detection-rate and false-alarm-rate that result from varying the threshold. In addition the binary GSM VAD [4] is depicted as reference.

For the proposed VAD system, it is obvious that it holds the best relationship between the false-alarm-rate and the detection-rate. The false-alarm-rate never exceeds 23 % with a maximum detection-rate of 93 %. In order to achieve the same detection-rate, significantly higher false-alarm-rates of 32 % (Ghosh), 42 % (Tan) or 44 % (Sohn) must be tolerated. However, the reference VAD systems achieve a higher maximum-detection-rate, but at the expense of a significantly higher false-alarm-rate.

The lower plot of Fig. 2 depicts the same ROC curve as above, but for different noise types. For the sake of clarity, only the proposed VAD and the best reference method, i.e., Ghosh [8], are visualized. For all noise types, the proposed method holds best performance. Moreover, the proposed algorithm performs well for stationary noise types, e.g. inside train and car, and for instationary noises like the jackhammer. However, a reliable voice detection during babble noise is not possible because this sort of noise is very similar to speech-codebook entries. Hence, babble noise is frequently classified as speech, leading to a high false alarm rate, yet better than Ghosh [8].

In Fig. 3 (upper plot), the average accuracies for varying thresholds are visualized. An advantage of the proposed technique is the flatness of this measure. Because of that, it is possible to set any desired working point on the ROC curve by adjusting the threshold without losing accuracy. In addition, it holds the best accuracy (especially for thresholds up to 0.4) over the complete threshold range. The accuracy of the remaining VAD algorithms increases with the threshold, with similar performance among them for $thr > 0.3$.

In order to analyze the accuracy over the SNR, the best thresholds for each algorithm are selected from Fig. 3 (upper plot). Using those thresholds, the accuracy is depicted over the SNR in the lower plot of Fig. 3, where the proposed VAD also provides the best scores.

5. CONCLUSION

A novel robust VAD system is presented, which utilizes a speech codebook to provide a speech energy gain in each frame. This gain provides a stable speech indicator and may contain a noise floor, especially at low SNR. A baseline tracing algorithm, known from noise reduction, is employed during the post-processing and subsequently the gain is mapped to soft VAD values between zero and one. The speaker-independent codebook is created by training a vector quantizer, using only the spectral envelopes of speech. While processing, the codebook is adapted in every frame to the current speaker by incorporating the pitch from the noisy input signal. The new VAD does not rely on noise *a-priori* information, which makes it robust also to highly non-stationary noise and adverse SNR conditions (e.g., -5 dB). If desired, a binary VAD can be calculated by applying a threshold. Instrumental measurements confirmed a consistent improvement in comparison to state-of-the-art systems [6, 11, 8], resulting in better detection rates at a significant lower false alarm rates. In addition, it is possible to adjust the compromise between a higher detection-rate versus a higher false-alarm-rate by changing the threshold without increasing the total number of miss-detections.

6. REFERENCES

[1] Peter Vary and Rainer Martin, *Digital Speech Transmission - Enhancement, Coding & Error Concealment*, p. 10, John Wiley & Sons, Ltd., Chichester, UK, Jan. 2006.

[2] R. McAulay and Mi Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, 1980.

[3] Dirk Van Compernolle, "Noise adaptation in a hidden markov model speech recognition system," *Computer Speech & Language*, vol. 3, no. 2, pp. 151–167, 1989.

[4] ETSI Recommendation, "GSM recommendations for VAD: GSM 06.32, GSM 06.42, GSM 06.82," *Voice activity detection for full rate speech traffic channels*.

[5] Antti Vähätalo and Ingemar Johansson, "Voice activity detection for GSM adaptive multi-rate codec," in *IEEE Workshop on Speech Coding Proceedings*, 1999, pp. 55–57.

[6] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[7] Yong Duk Cho and A Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *Signal Processing Letters, IEEE*, vol. 8, no. 10, pp. 276–278, Oct. 2001.

[8] P.K. Ghosh, A Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 600–613, Mar. 2011.

[9] Justinian Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proceedings of EUSIPCO*, 2002.

[10] Mohammad J. Taghizadeh, Philip N. Garner, Hervé Bourlard, Hamid R. Abutalebi, and Afsaneh Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 92–97.

[11] Lee Ngee Tan, B.J. Borgstrom, and Abeer Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, Mar. 2010, pp. 4466–4469.

[12] Florian Heese, Christoph Matthias Nelke, Markus Niermann, and Peter Vary, "Selflearning codebook speech enhancement," in *ITG Fachtagung Sprachkommunikation*. Sept. 2014, VDE Verlag GmbH.

[13] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *Communications, IEEE Transactions on*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[14] Tobias Rosenkranz, "Noise codebook adaptation for codebook-based noise reduction," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv*, 2010.

[15] R. Martin, P.U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*, p. 123, Wiley, 2008.

[16] Christin Baasch, Vasudev Kandade Rajan, Mohamed Krini, and Gerhard Schmidt, "Low-complexity noise power spectral density estimation for harsh automobile environments," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 219–223.

[17] Florian Heese and Peter Vary, "Noise PSD estimation by logarithmic baseline tracing," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Apr. 2015.

[18] John S. Garofolo and Linguistic Data Consortium, *TIMIT: acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, 1993.

[19] *ETSI EG 202 396-1 background noise database*, 2014, Stand 14. März 2014.