Offline Voice Activity Detector Using Speech Supergaussianity

Ivan J. Tashev
Microsoft Research Labs
One Microsoft Way, Redmond, WA 98051, USA
Email: ivantash@microsoft.com

***Abstract*—Voice Activity Detectors (VAD) play important role in audio processing algorithms. Most of the algorithms are designed to be causal, i.e. to work in real time using only current and past audio samples. Off-line processing, when we have access to the entire voice utterance, allows using different type of approaches for increased precision. In this paper we propose an algorithm for off-line VAD based on the different probability density functions (PDFs) of the speech and noise. While a Gaussian distribution is a very good model for noise, the speech PDF is peakier. The proposed VAD algorithm works in frequency domain and estimates the speech signal presence probability for each frequency bin in each audio frame, the speech presence probability for each frame and also provides a binary decision per bin and frame. Provides improved precision compared to the streaming real-time VAD algorithms.**

I. Introduction

Voice Activity Detectors (VAD) are algorithms for detecting the presence of a speech signal in the mixture of speech and noise. They are part of noise suppressors, double talk detectors, codecs, and automatic gain control blocks, to mention a few. The VAD output can vary from simple binary decision (yes/no), trough soft decision (probability of speech presence in the current audio frame), to probability of speech presence in each frequency bin of each audio frame. The commonly used VAD algorithms are based on the assumption of quasi-stationary noise, i.e. noise changes much slower than the speech signal. A classic VAD algorithm works in real time and makes the decisions based on the current and previous frames, i.e. it is causal. Most of these algorithms work in frequency domain for better integration in the audio processing chain and provide estimation for each frequency bin separately. One of the frequently used as a baseline VAD algorithm is standardized as ITU-T Recommendation G.729-Annex B [1]. An improved and generalized VAD is described in [2], where authors create a soft decision VAD assuming Gaussian distribution of the noise and speech signals. A simple HMM is added to create a hangover scheme in [3] and to finalize the decision utilizing the timing of switching the states. This algorithm can be optimized for better performance as described in [4].

Most of the VAD algorithms assume Gaussian distribution of the noise and speech signals. It is well known that while the distribution of noise amplitudes in time domain is well modelled with the Gaussian distribution, the distribution of the amplitudes of the speech signal has higher kurtosis than the Gaussian distribution. Gazor and Zhang [5] published a study for the speech signal distribution in time domain, later in [6] this study was extended with models of the Probability Density Functions (PDFs) of the speech signal magnitudes in frequency domain. In the literature are published several attempts to utilize the non-Gaussianity of the speech signal for better noise suppression rules [7] and [8], or for better VAD [9]. In most of the cases it is very difficult to find analytical form of the suppression rules, or speech presence probability, and the proposed solutions are either approximate or computationally expensive.

In many cases we do the processing off-line and have access to the entire speech utterance. These scenarios include speech enhancement of the recorded utterances sent for speech recognition in the cloud, post-processing of the audio track in video shots, off-line encoding of audio signals. This type or processing allows building more precise statistical models and is more tolerant to computationally expensive algorithms. The question is how much the prior knowledge of the PDFs of the speech and noise signals can help for better parameters estimation, which leads to better VAD algorithms.

In this paper we propose an algorithm for off-line VAD based on the statistical processing of the entire utterance. In section II we formulate the problem and present the statistical models in a VAD. Sections III, IV, and V describe the application of the statistical modeling of the noise and speech PDFs for estimation of the speech and noise signal parameters for normalization and VAD. In section VI we describe the experimental results and conclude in VII.

II. Problem Definition

A. Modelling

We have a limited discrete signal in time domain $x(lT)$ where $l \in [0, L-1]$, $x_{\min} \leq x(lT) \leq x_{\max}, \forall l$, and T is the sampling period. An example of such signal is shown in figure 1. This signal is a mixture of two limited, discrete, and uncorrelated signals $x(lT) = n(lT) + s(lT)$, noise $n(lT)$ and speech $s(lT)$ respectively. After framing, windowing, and converting to frequency domain we have $X_b^{(n)} = N_b^{(n)} + S_b^{(n)}$, where $b \in [0, B-1]$ is the frequency bin, B is the number of frequency bins, $n \in [0, N-1]$ is the frame number, and N is the number of audio frames. As we do the processing off-line the same can be written in matrix form $\mathbf{X} = \mathbf{N} + \mathbf{S}$, where all are $B \times N$ complex matrices representing the spectra

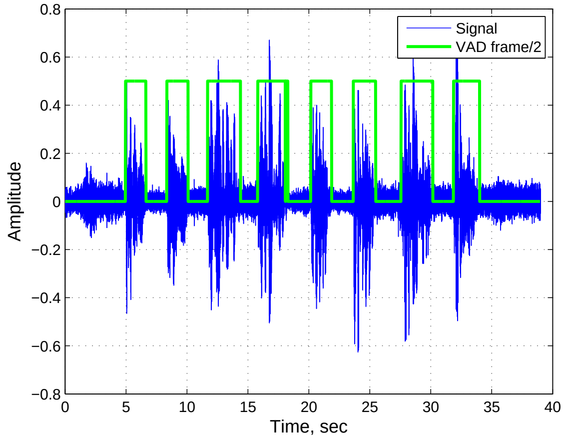

Fig. 1. Noisy signal in time domain with SNR=10 dB.

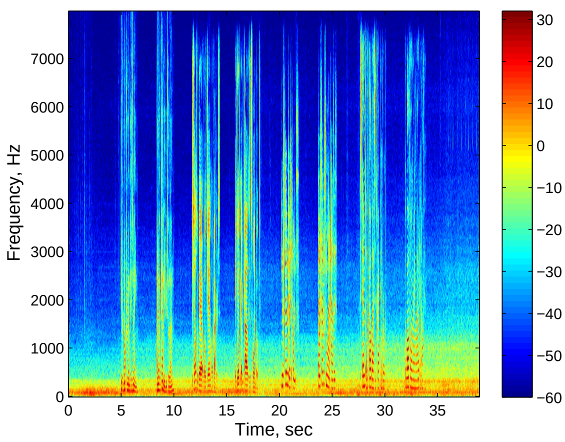

Fig. 2. Noisy signal in frequency domain with SNR=10 dB.

of the signal, the noise, and the speech components. This representation is visualized in figure 2, where the magnitudes are in a decibel scale.

In each frame and/or frequency bin we consider two hypotheses:

H_0: speech is absent, $\mathbf{X} = \mathbf{N}$
H_1: speech is present, $\mathbf{X} = \mathbf{N} + \mathbf{S}$

The goal of the VAD algorithm is to produce for each frequency bin and for each frame (column in the matrices above) the probability of speech signal presence, $P_b^{(n)}(H_1)$ and $P^{(n)}(H_1)$ respectively. An example of the expected VAD decision per frame is shown in figure 1.

B. Voice Activity Detectors

Let assume that the noise and speech signals are fully characterized by their respective variances σ_n^2 and σ_s^2 and we have a prior knowledge of the PDFs of these two signals $p_n\left(a\left|\sigma_n^2\right.\right)$ and $p_s\left(a\left|\sigma_s^2\right.\right)$. The PDF of a mix of two uncorrelated signals is the convolution of the PDFs of the two signals:

$$p_x\left(a\left|\sigma_n^2, \sigma_s^2\right.\right) = p_n\left(a\left|\sigma_n^2\right.\right) * p_s\left(a\left|\sigma_s^2\right.\right). \qquad (1)$$

Note that this equation has analytical solution for a small number of distribution pairs, it has to be solved numerically for most of the cases.

The probability $P(H_1|a)$ of signal with amplitude a to contain speech is derived after direct applying of the Bayesian rule:

$$P(H_1|a) = \frac{p(a|H_1)P(H_1)}{p(a|H_1)P(H_1) + p(a|H_0)P(H_0)}. \qquad (2)$$

Here $P(H_1)$ and $P(H_0) = 1 - P(H_1)$ are the prior probabilities for speech and noise presence. After dividing by $p(a|H_0)P(H_0)$ we have:

$$P(H_1|a) = \frac{\varepsilon\Lambda}{1 + \varepsilon\Lambda}, \qquad (3)$$

where $\varepsilon = P(H_1)/P(H_0)$, and Λ is the likelihood of speech signal presence:

$$\Lambda = \frac{p_x\left(a\left|\sigma_n^2, \sigma_s^2\right.\right)}{p_n\left(a\left|\sigma_n^2\right.\right)}. \qquad (4)$$

The proportion of the prior probabilities for speech and noise ε can be assumed constant and known. Then if we can estimate the noise and speech variances - we can estimate the speech presence presence probability in each frame and/or frequency bin.

The binary flag $V^{(n)}$ for speech presence (1) or absence (0) can be obtained by comparing the likelihood Λ or the speech presence probability $P^{(n)}$ with fixed threshold η:

$$V^{(n)} = \left| \begin{array}{lll} 1 & \text{if} & P^{(n)}(H_1) > \eta \\ 0 & \text{if} & P^{(n)}(H_1) \leq \eta \end{array} \right. \qquad (5)$$

For practical purposes a small hysteresis is added to prevent ”ringing” of the flag when the probability is close to the threshold.

C. Parameters Estimation

For given set of M equally spaced amplitudes $a_m = x_{\min} + m\Delta$, where $m \in [0, M-1]$, $x_{\min} < a < x_{\max}$, $\Delta = (x_{\max} - x_{\min})/(M-1)$, we always can build the histogram $h_M(a_m)$ of the input signal and compute the probability density function:

$$p(a_m) = \frac{h_M(a_m)}{\Delta \sum\limits_m h_M(a_m)}. \qquad (6)$$

To make the PDF $p(a_m)$ independent of the pauses between the clean speech phrases we remove the bin of the histogram, which contains the zero magnitude.

The distance between the two probability distributions is given by the Jensen-Shannon divergence [10], which is a symmetrized and smoothed version of Kullback-Leibler divergence [11] $D_{KL}(p\|q)$:

$$D_{JS}(p\|q) = \frac{1}{2}\left(D_{KL}(p\|m) + D_{KL}(q\|m)\right), \qquad (7)$$

where $m = (p+q)/2$. The Kullback-Leibler divergence is defined as:

$$D_{KL}(p\,\|q) = \sum_i p(x_i)\log\frac{p(x_i)}{q(x_i)} \tag{8}$$

and measures the expected number of extra bits, required to code samples from q when using a code based on p, rather than using a code based on q. Lower Jensen-Shannon divergence D_{JS} indicates a better fit of the model to the measured histogram.

We can estimate the noise and speech variances by minimizing the divergence between measured PDF $h_M(a)$, derived from the histogram of amplitudes, and the models of the noise and speech signals PDF $p_x\left(a\left|\sigma_n^2,\sigma_s^2\right.\right)$:

$$\left[\hat{\sigma}_n^2,\hat{\sigma}_s^2\right] = \arg\min_{\sigma_n^2,\sigma_s^2}\left(D_{JS}\left(h_M(a)\left\|p_x\left(a\left|\sigma_n^2,\sigma_s^2\right.\right)\right.\right)\right) \tag{9}$$

Considering the fact that it is unlikely to have analytic solution for the PDF of the mixed signal (see II-B) we will have to solve the minimization problem in equation 9 numerically.

Once we have estimation of the noise and speech variances we can compute the likelihood using equation 4 and then compute the speech presence probability for each audio frame and/or frequency bin using equation 3. We are going to apply this technique several times further in this paper.

III. Speech Signal Normalization

Before to convert to frequency domain and proceed with the VAD algorithm we found useful to normalize the speech signal. In many cases the recorded signal has low level, contains clicks, bursts of wind noise. The last two are the reason why a simple normalization of the peak amplitude will not work - typically these are short peaks with high amplitude. It is OK to clip them and to normalize the amplitude of the actual speech signal.

We can build the histogram of the signal in time domain using the procedure described in II-C and compute the PDF of the mixed signal $p_{xt}(a_m)$ according to equation 6. The PDF of the noise signal is modeled as zero-mean Gaussian distribution:

$$p_{nt}\left(a\left|\sigma_{nt}^2\right.\right) = \frac{1}{\sigma_{nt}\sqrt{2\pi}}\exp\left(-\frac{a^2}{2\sigma_{nt}^2}\right) \tag{10}$$

where σ_{nt}^2 is the noise variance.

The PDF of the speech signal is modelled with zero-mean Generalized Gaussian Distribution [12] as:

$$p_{st}\left(a\left|\sigma_{st}^2\right.\right) = \frac{\beta_t}{2\alpha_t\Gamma(1/\beta_t)}\exp\left(-\left(\frac{|a|}{\alpha_t}\right)^{\beta_t}\right) \tag{11}$$

where $\Gamma(\cdot)$ is the gamma function, α_t is the scale, and β_t is the shape parameter. When the shape parameter $\beta_t = 2$ we have Gaussian distribution, when $\beta_t = 1$ we have Laplace distribution, and the clean speech distribution has even higher kurtosis. We assume that the shape parameter is constant for the clean speech signal and known in advance. The only parameter of the clean speech signal PDF which needs to be estimated is the scale, related to the speech signal variance σ_{st}^2 as:

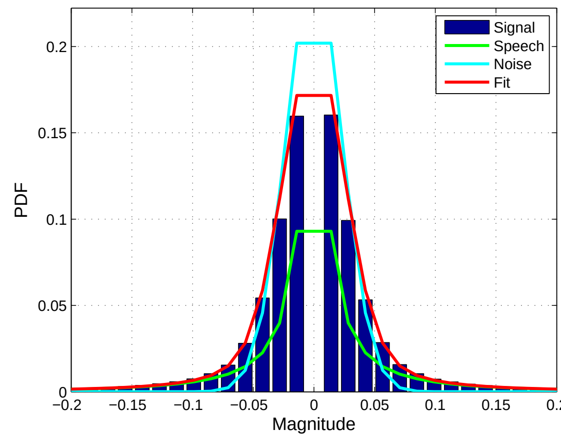

Fig. 3. Distribution fit for signal in time domain with SNR=10 dB.

$$\alpha_t = \sqrt{\frac{\sigma_{st}^2\Gamma(1/\beta_t)}{\Gamma(3/\beta_t)}}. \tag{12}$$

We can use equation 9 to estimate the noise and speech variances σ_{nt}^2 and σ_{st}^2 respectively. Figure 3 shows an example of the distributions of signal in time domain with SNR=10 dB. This is enlarged view of the central part of the distribution. Note the missing histogram bin with zero values and the heavier tail of the distribution of the speech signal.

The input signal in time domain can be normalized by applying gain $G_s = \sigma_{st0}/\sigma_{st}$ aiming to have desired deviation of the speech component σ_{st0}. To handle corner cases, such as only noise or too small amount of speech signal, this gain is limited as

$$G = \min\left[G_s, G_p\right], \tag{13}$$

where G_p is the gain which will cause 90% of the samples to be clipped. It can be obtained from the cumulative distribution function, easily computed from the estimated PDF, or directly from the histogram.

IV. Per-frame VAD algorithm

After normalization and conversion to frequency domain we can compute the RMS of the signal frames as:

$$A^{(n)} = \sqrt{\frac{1}{N}\sum_{b=B_{beg}}^{B_{end}}\left(X_b^{(n)}\right)^2}. \tag{14}$$

Here we implicitly applied a band-pass filter by using only the frequency bins from B_{beg} to B_{end}.

We model the PDF of magnitudes of noise only frames with Weibull distribution [13]:

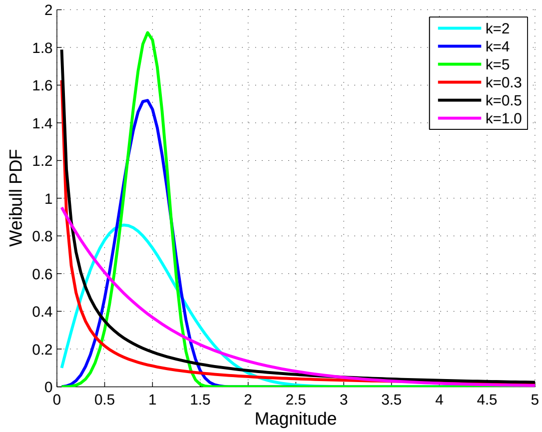

Fig. 4. Weibull distribution.

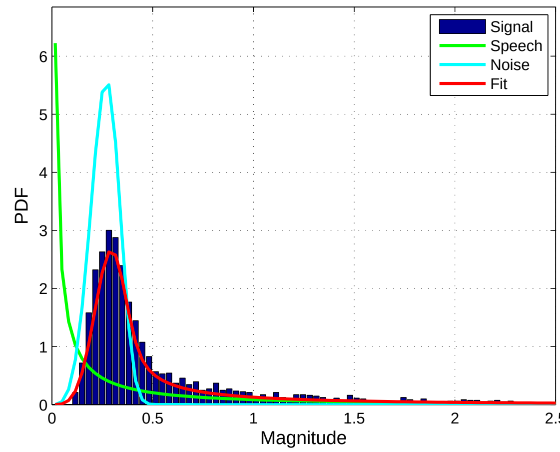

Fig. 5. Distribution fit of the frame magnitudes for signal in frequency domain with SNR=10 dB.

$$p_{nf}\left(A\left|\sigma_{nf}^2\right.\right)=\frac{k_{nf}}{\lambda_{nf}}\left(\frac{A}{\lambda_{nf}}\right)^{k_{nf}-1}\exp\left(-\left(\frac{A}{\lambda_{nf}}\right)^{k_{nf}}\right). \tag{15}$$

Here λ_{nf} is the scale parameter and k_{nf} is the shape parameter. This distribution is defined for $A \geq 0$ and is shown in figure 4 for various shape parameters. When $k_{nf} = 2$ Weibull distribution becomes Rayleigh distribution, which is the PDF of the magnitudes of a complex numbers with a zero-mean Gaussian distribution of the real and imaginary parts, i.e. noise. We can state that this is the distribution of the magnitudes of frames with one sample. With increasing of the shape parameter above 2 the Weibull distribution becomes more and more narrow bell-shaped keeping the same mean. This models well the reduction of the variation of the frame RMS with increasing the number of the samples in the frame. Here we assume known shape parameter $k_{nf} > 2$, which is frame size dependent.

The only parameter we have to estimate is the scale parameter λ_{nf}, which is related with the noise variance as:

$$\lambda_{nf}=\frac{\sigma_{nf}}{\Gamma\left(1+\frac{1}{k_{nf}}\right)}. \tag{16}$$

In this case σ_{nf} is the deviation of the noise, i.e. the mean of the PDF.

We model the PDF of magnitudes of the audio frames with only speech $p_{sf}\left(A\left|\sigma_{sf}^2\right.\right)$ with the same Weibull distribution (Figure 4), but with different shape parameter k_{sf}. When the shape parameter $k_{sf} = 1$ the Weibull distribution becomes exponential distribution and the actual clean speech PDF has even higher kurtosis. We assume known shape parameter $k_{sf} < 1$ of the PDF of the speech audio frames and then the only parameter we have to estimate is the scale parameter λ_{sf}, which is a function of the speech signal variance σ_{sf}^2 as shown in equation 16.

Using the approach described in section II-C we can build the histogram and estimate the noise only and speech only variances σ_{nf} and σ_{ns} by minimizing the Jensen-Shannon divergence. Figure 5 illustrates the result from this distribution fitting for a signal with SNR=10 dB converted to frequency domain.

With the obtained noise only and speech only variances σ_{nf} and σ_{sf} we can estimate the frame presence probability using the process described in section II-B.

V. Per-Bin VAD algorithm

The approach for per-bin estimation of the speech presence probability is the same. Here we will omit the bin index wherever is possible. The noise PDF if given by Rayleigh distribution [14]:

$$p_{nb}\left(A_b\left|\sigma_{nb}^2\right.\right)=\frac{A_b}{\sigma_{nb}^2}\exp\left(-\frac{A_b^2}{2\sigma_{nb}^2}\right), A_b \geq 0 \tag{17}$$

Here σ_{nb}^2 is the noise variance for frequency bin b.

The clean speech signal distribution $p_{sb}\left(A\left|\sigma_{sb}^2\right.\right)$ is modelled also with Weibull distribution according to equation 15. λ_{sb} is the scale parameter and k_{sb} is the shape parameter, which is assumed known and constant across all frequency bins [6] with light dependency of the frame size. The scale parameter is related with the speech variance per equation 16. Here we also will have to omit the first magnitude bin because of the same reasons as in section IV. An example for fitting of the PDF for the frequency bin corresponding to 1000 Hz is shown in figure 6. The estimated speech presence probability per frequency bin is show in figure 7 for the same signal with SNR=10 dB.

For relatively high SNRs the complex and computationally expensive minimizations can be avoided by direct estimation of the noise and clean speech variances for each bin separately. Given per-frame speech signal presence probability $P^{(n)}\left(H_1\right)$ we can estimate the noise and speech plus noise variances as an weighted average:

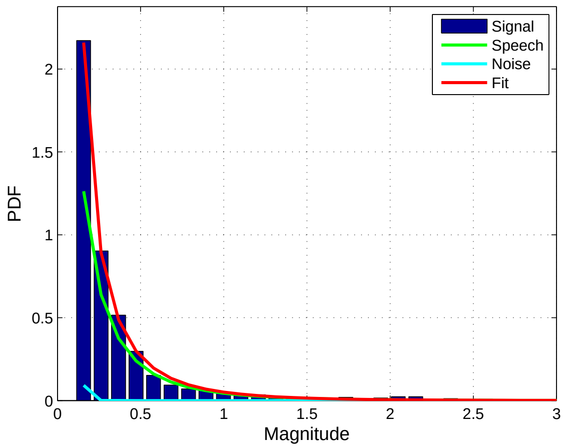

Fig. 6. Distribution fit of the magnitudes in 1000 Hz frequency bin for signal in frequency domain with SNR=10 dB.

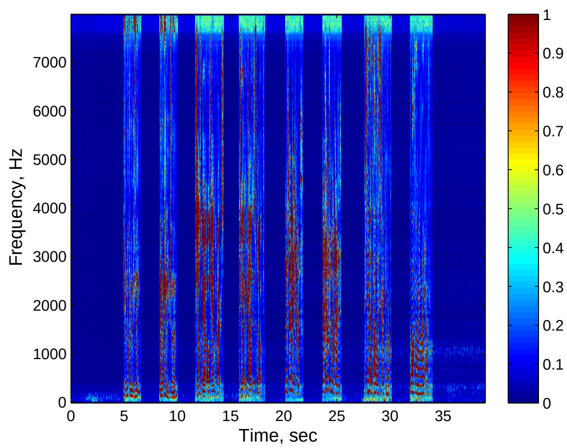

Fig. 7. Estimated speech presence probability per frequency bin, SNR=10 dB.

$$\hat{\sigma}_{nb}^2 = \frac{\sum_n \left(X_b^{(n)}\right)^2 \left(1 - P^{(n)}\left(H_1\right)\right)}{\sum_n \left(1 - P^{(n)}\left(H_1\right)\right)}, \tag{18}$$

$$\hat{\sigma}_{s+n,b}^2 = \frac{\sum_n \left(X_b^{(n)}\right)^2 P^{(n)}\left(H_1\right)}{\sum_n P^{(n)}\left(H_1\right)}. \tag{19}$$

Then we can assume $\hat{\sigma}_s b^2 \approx \hat{\sigma}_{s+n,b}^2$ and continue the estimation as above.

VI. Experimental Results

A. Data set

For evaluation of the proposed algorithm we created a clean speech file containing 10 utterances randomly taken from 10 different speakers in TIMIT [15] database and mixed them with vehicle noise recorded in a moving car in proportion to create set of files with SNR of 0, 10, 20, 30, 40, and 50 dBC. Two sets of files were created for the evaluation of the proposed algorithm - one for training and one for evaluation. Different speakers and noise segments were selected for the two sets. Both sets also include the clean speech files and the noise only files. All files are with sampling rate of 16 kHz.

B. Evaluation parameters

As evaluation parameters we selected the error rate per frame and per bin. The error rate per frame is estimated as:

$$E_f = \sqrt{\frac{1}{N} \sum_n \left(P^{(n)}\left(H_1\right) - G^{(n)}\left(H_1\right)\right)}, \tag{20}$$

where $G^{(n)}\left(H_1\right)$ is the ground truth, a binary mask obtained by comparing the frame RMS of the clean speech signal with a fixed, very low threshold. The ground truth per bin $G_b^{(n)}\left(H_1\right)$ and the error rate per bin E_b are obtained and estimated in the same way. Identical approach was used to compute the error rates for the binary decisions.

C. Implementation and parameters

The values of the distributions parameters, assumed constant, were obtained using the noise only and clean speech only files from the training set. The prior probability proportions and optimal thresholds were obtained by minimizing the error rates against the training set. The values of the constant parameters are shown in Table I.

The implementation of the off-line VAD was done in Matlab. The histograms were estimated in $M = 100$ points. The frame size for conversion to frequency domain was 512 samples with 50% overlapping, Hann window. For the numerical estimation of the convolution in equation 1 was used grid ten times denser than the one used for the histogram. For solving the optimization problem in equation 9 was used the Matlab function for non-constrained optimization $fminunc()$, and the optimization process was constrained with setting minimal and maximal values for each optimization parameter and applying quadratic punishing functions, added to the optimization criterion. Proper measures were taken to prevent overflows and divisions by zero with adding small numbers to the denominators.

D. Results

All of the results presented in this section were obtained using the evaluation data set. Table II shows the results for three VAD structures. The first is the VAD per frame, as described in section IV, for the soft and binary decisions. The second is the per-bin VAD from section V with direct estimation of the variances of the noise and the speech signals according to equations 18 and 19. The third VAD is the per-bin VAD from section V with estimation of the noise and speech variances by fitting the distributions.

VII. Conclusions

In this paper we presented an approach for off-line VAD which utilizes the prior knowledge of the PDFs of the speech and noise signals, components of the mixture.

TABLE I
VALUES OF THE CONSTANT PARAMETERS

β_t	k_{nf}	k_{sf}	ε_f	η_f	k_{sb}	ε_b	η_b
0.2403	4.2416	0.3244	0.0215	0.9	0.571	0.0236	0.25

TABLE II
RESULTS

Algorithm	SNR, dB	0	10	20	30	40	50	Average
VAD per frame	soft	0.3898	0.2149	0.1355	0.0578	0.0380	0.0338	0.1449
	binary	0.3974	0.2061	0.1178	0.0524	0.0308	0.0283	0.1388
VAD per bin, direct	soft	0.3523	0.3110	0.2674	0.2557	0.2526	0.2509	0.2833
	binary	0.3233	0.2890	0.2478	0.2424	0.2312	0.2301	0.2606
VAD per bin, fit	soft	0.2699	0.2007	0.1611	0.1355	0.1166	0.1118	0.1659
	binary	0.2667	0.1947	0.1508	0.1236	0.1033	0.0985	0.1562

One of the core ideas in this approach is the numerical estimation of the mixed signal PDF from the PDFs of the noise and the speech signal. In time domain we model the noise with Gaussian distribution and the clean speech signal with Generalized Gaussian distribution. In frequency domain we model the magnitudes of the audio frames with Weibull distribution for both noise only and clean speech only audio frames using different values of the shape parameter. In frequency domain we model the magnitudes of the frequency bins with Rayleigh distribution for noise only bins and with Weibull distribution for the speech only bins. In all cases we assume that we have constant values of the shape parameters, which makes the PDFs function only of the variations. We estimate these variations by fitting the estimated PDF to the actual PDF using Jensen-Shannon divergence as minimization criterion.

The algorithm is evaluated against a small data set with various signal-to-noise ratios (SNRs) and shows comparable performance with the real-time VAD algorithms of the same class. The proposed approach handles much better signals with very high SNRs where, surprisingly enough, the classic VADs do not perform very well. This make it suitable for building the baseline for testing and evaluating another VAD algorithms.

Potential next steps are adding a hangover scheme, proven to be very effective for improving the precision of the VAD decisions for both per-frequency bin and per frame. Implementing the ideas from [9] for combining the per-bin decisions to form a better per-frame decision is also a factor for improving the performance.

REFERENCES

[1] "Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," 1997.

[2] J. Sohn and W. Sung, "A voice activity detection employing soft decision based noise spectrum adaptation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

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

[4] Ivan Tashev, Andrew Lovitt, and Alex Acero, "Unified framework for single channel speech enhancement," in *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2009.

[5] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, pp. 204–207, 2003.

[6] Ivan Tashev and Alex Acero, "Statistical modeling of the speech signal," in *Proceedings of International Workshop on Acoustic, Echo, and Noise Control (IWAENC)*, 2010.

[7] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[8] T. Lotter, *Speech Enhancement*, chapter Single- and Multi-Microphone Spectral Amplitude Estimation Using a Super-Gaussian Speech Model, pp. 67–95, Springer-Verlag, 2005.

[9] Ivan Tashev, Andrew Lovitt, and Alex Acero, "Dual stage probabilistic voice activity detector," in *NOISE-CON 2010 and 159th Meeting of the Acoustical Society of America*, 2010.

[10] Hinrich Schtze and Christopher D. Manning, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[11] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[12] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *Journal of the Acoustical Society of America*, vol. 86, pp. 1404–1415, 1989.

[13] W. Weibull, "A statistical distribution function of wide applicability," *Journal of Applied Mechchanics - Transactions of ASME*, vol. 18, pp. 293–297, 1951.

[14] Athanasios Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processe*, McGraw-Hill Science Engineering, 2001.

[15] John S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993.