Comparison of Statistical Model-Based Voice Activity Detectors for Mobile Robot Speech Applications

Ivan Marković, Hrvoje Domitrović and Ivan Petrović

University of Zagreb
Faculty of Electrical Engineering and Computing

September 5, 2012

1. Introduction
2. Statistical Model-Based VADs
   - Gaussian distribution
   - Generalized Gaussian distribution
   - Rayleigh-Rice distribution
3. Noise spetrum estimation
4. Experiments
5. Conclusion

Motivation

- A necessary front-end for robotic speech applications

Motivation

- A necessary front-end for robotic speech applications
- Speaker localization, speaker identification or speech recognition

Motivation

- A necessary front-end for robotic speech applications
- Speaker localization, speaker identification or speech recognition
- Focus on statistical model-based voice activity detectors

- A two hypothesis scenario:

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

where $\mathbf{X}$, $\mathbf{N}$ and $\mathbf{S}$ are the DFT coefficients of a K-point DFT of the noisy speech, noise, and clean speech

- A two hypothesis scenario:

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

where $\mathbf{X}$, $\mathbf{N}$ and $\mathbf{S}$ are the DFT coefficients of a K-point DFT of the noisy speech, noise, and clean speech

- Model distributions $p(\mathbf{X}|H_0)$ and $p(\mathbf{X}|H_1)$

- A two hypothesis scenario:

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

where $\mathbf{X}$, $\mathbf{N}$ and $\mathbf{S}$ are the DFT coefficients of a K-point DFT of the noisy speech, noise, and clean speech

- Model distributions $p(\mathbf{X}|H_0)$ and $p(\mathbf{X}|H_1)$
- Likelihood ratio

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)}$$

- A two hypothesis scenario:

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

where **X**, **N** and **S** are the DFT coefficients of a K-point DFT of the noisy speech, noise, and clean speech

- Model distributions $p(\mathbf{X}|H_0)$ and $p(\mathbf{X}|H_1)$
- Likelihood ratio

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)}$$

- Geometric mean

$$\log \Lambda = \frac{1}{K} \sum_{k=0}^{K-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Gaussian distribution [Sohn et al., 1999]

- A DFT coefficient $S_k = S_{R,k} + jS_{I,k}$

Gaussian distribution [Sohn et al., 1999]

- A DFT coefficient $S_k = S_{R,k} + jS_{I,k}$
- Independent zero-mean gaussian random variables with variance of $\lambda_{s,k}/2$

$$p(S_{R,k}) = \frac{1}{\sqrt{\pi\lambda_{s,k}}} \exp\left\{-\frac{S_{R,k}^2}{\lambda_{s,k}}\right\}$$

$$p(S_{I,k}) = \frac{1}{\sqrt{\pi\lambda_{s,k}}} \exp\left\{-\frac{S_{I,k}^2}{\lambda_{s,k}}\right\}$$

Gaussian distribution [Sohn et al., 1999]

- A DFT coefficient $S_k = S_{R,k} + jS_{I,k}$
- Independent zero-mean gaussian random variables with variance of $\lambda_{s,k}/2$

$$p(S_{R,k}) = \frac{1}{\sqrt{\pi\lambda_{s,k}}} \exp\left\{-\frac{S_{R,k}^2}{\lambda_{s,k}}\right\}$$
$$p(S_{I,k}) = \frac{1}{\sqrt{\pi\lambda_{s,k}}} \exp\left\{-\frac{S_{I,k}^2}{\lambda_{s,k}}\right\}$$

- Joint distribution

$$\begin{aligned} p(S_k) = p(S_{R,k})p(S_{I,k}) &= \frac{1}{\pi\lambda_{s,k}} \exp\left(-\frac{S_{R,k}^2 + S_{I,k}^2}{\lambda_{s,k}}\right) \\ &= \frac{1}{\pi\lambda_{s,k}} \exp\left(-\frac{|S_k|^2}{\lambda_{s,k}}\right) \end{aligned} \tag{1}$$

Gaussian distribution [Sohn et al., 1999]

- Two hypotheses

$$p(X_k|H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}$$

$$p(X_k|H_1) = \frac{1}{\pi(\lambda_{n,k} + \lambda_{s,k})} \cdot \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k} + \lambda_{s,k}}\right\}$$

Gaussian distribution [Sohn et al., 1999]

- Two hypotheses

$$p(X_k|H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}$$

$$p(X_k|H_1) = \frac{1}{\pi(\lambda_{n,k}+\lambda_{s,k})} \cdot \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}+\lambda_{s,k}}\right\}$$

- Likelihood ratio for GD VAD

$$\Lambda_k^{\mathrm{GD}} = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\},$$

where $\xi_k = \lambda_{s,k}/\lambda_{n,k}$ is the *a priori* SNR, and $\gamma_k = |X_k|^2/\lambda_{n,k}$ is the *a posteriori* SNR

Generalized Gaussian distribution [Chang et al., 2004]

- Joint distribution

$$p(S_k) = \frac{\nu^2 \alpha^2(\nu)}{4\lambda_{s,k}\Gamma^2(1/\nu)} \cdot \exp\left\{-\alpha^\nu(\nu)\left[\left|\frac{S_{R,k}}{\sqrt{\lambda_{s,k}}}\right|^\nu + \left|\frac{S_{I,k}}{\sqrt{\lambda_{s,k}}}\right|^\nu\right]\right\}$$

with

$$\alpha(\nu) = \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}}$$

Generalized Gaussian distribution [Chang et al., 2004]

- Joint distribution

$$p(S_k) = \frac{\nu^2 \alpha^2(\nu)}{4\lambda_{s,k}\Gamma^2(1/\nu)} \cdot \exp\left\{ -\alpha^\nu(\nu) \left[\left| \frac{S_{R,k}}{\sqrt{\lambda_{s,k}}} \right|^\nu + \left| \frac{S_{I,k}}{\sqrt{\lambda_{s,k}}} \right|^\nu \right] \right\}$$

with

$$\alpha(\nu) = \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}}$$

- Likelihood ration for GGD VAD

$$\begin{aligned} \Lambda_k^{GGD} =& \frac{1}{1+\xi_k} \cdot \frac{\nu_{s,k}^2 \alpha^2(\nu_{s,k}) \Gamma^2(1/\nu_{n,k})}{\nu_{n,k}^2 \alpha^2(\nu_{n,k}) \Gamma^2(1/\nu_{s,k})} \exp \Bigg\{ \\ & - \alpha^{\nu_{s,k}}(\nu_{s,k}) \left[\frac{|X_{R,k}|^{\nu_{s,k}} + |X_{I,k}|^{\nu_{s,k}}}{\left(\sqrt{\lambda_{n,k}(1+\xi_k)}\right)^{\nu_{s,k}}} \right] + \\ & + \alpha^{\nu_{n,k}}(\nu_{n,k}) \left[\frac{|X_{R,k}|^{\nu_{n,k}} + |X_{I,k}|^{\nu_{n,k}}}{\left(\sqrt{\lambda_{n,k}}\right)^{\nu_{n,k}}} \right] \Bigg\} \end{aligned} \tag{2}$$

Rayleigh-Rice distribution [Mumolo et al., 2003]

- Model the signal envelope $|X_k| = \sqrt{X_{R,k}^2 + X_{I,k}^2}$

Rayleigh-Rice distribution [Mumolo et al., 2003]

- Model the signal envelope $|X_k| = \sqrt{X_{R,k}^2 + X_{I,k}^2}$
- Under hypothesis H_0 we have Rayleigh distribution

$$p(X_k|H_0) = \frac{2|X_k|}{\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}$$

Rayleigh-Rice distribution [Mumolo et al., 2003]

- Model the signal envelope $|X_k| = \sqrt{X_{R,k}^2 + X_{I,k}^2}$
- Under hypothesis H_0 we have Rayleigh distribution

$$p(X_k|H_0) = \frac{2|X_k|}{\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}$$

- Under hypothesis H_1 we have Rice distribution

$$p(X_k|H_1) = \frac{2|X_k|}{\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}} - \xi_k\right\} \cdot I_0\left\{2\sqrt{\xi_k \frac{|X_k|^2}{\lambda_{n,k}}}\right\}$$

Rayleigh-Rice distribution [Mumolo et al., 2003]

- Model the signal envelope $|X_k| = \sqrt{X_{R,k}^2 + X_{I,k}^2}$
- Under hypothesis H_0 we have Rayleigh distribution

$$p(X_k|H_0) = \frac{2|X_k|}{\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}$$

- Under hypothesis H_1 we have Rice distribution

$$p(X_k|H_1) = \frac{2|X_k|}{\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}} - \xi_k\right\} \cdot I_0\left\{2\sqrt{\xi_k \frac{|X_k|^2}{\lambda_{n,k}}}\right\}$$

- Likelihood ratio for RRD VAD

$$\Lambda_k^{\mathrm{RRD}} = \exp\{-\xi_k\}\, I_0\left\{2\sqrt{\xi_k \gamma_k}\right\}$$

- The three VADs require estimates of $\lambda_{n,k}$ and ξ_k

- The three VADs require estimates of $\lambda_{n,k}$ and ξ_k
- $\lambda_{n,k}$ is estimated using minima controlled recursive averaging [Cohen, 2003]

$$\lambda_{n,k}(l) = \alpha\lambda_{n,k}(l-1) + (1-\alpha)|X_k(l)|^2$$

- The three VADs require estimates of $\lambda_{n,k}$ and ξ_k
- $\lambda_{n,k}$ is estimated using minima controlled recursive averaging [Cohen, 2003]

$$\lambda_{n,k}(l) = \alpha\lambda_{n,k}(l-1) + (1-\alpha)|X_k(l)|^2$$

- ζ_k is calculated via decision directed a-priori SNR estimation [Ephraim and Malah, 1984]

$$\xi_k(l) = f(\xi_k(l-1), \gamma_k(l-1), \gamma_k(l))$$

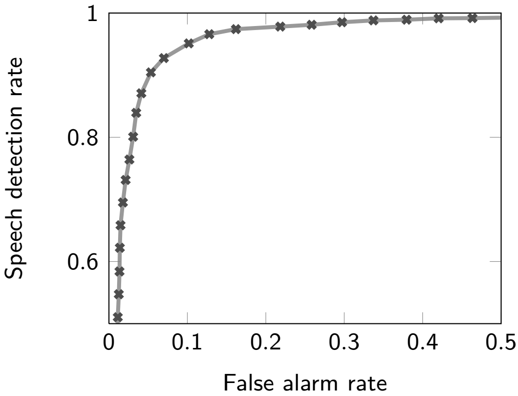
- NOIZEUS [Hu and Loizou, 2007] speech corpus (sound-proof booth, various noise added at different SNR levels, ...)

- NOIZEUS [Hu and Loizou, 2007] speech corpus (sound-proof booth, various noise added at different SNR levels, . . .)
- We used car, babble, and white noise at three different SNR levels (15, 10, 5 dB)

- NOIZEUS [Hu and Loizou, 2007] speech corpus (sound-proof booth, various noise added at different SNR levels, ...)
- We used car, babble, and white noise at three different SNR levels (15, 10, 5 dB)
- With 50% overlap we had 50000 examples, out of which 61.28% contained speech

Receiver operating characteristics

- depict relationship between speech detection rate and false alarm rate



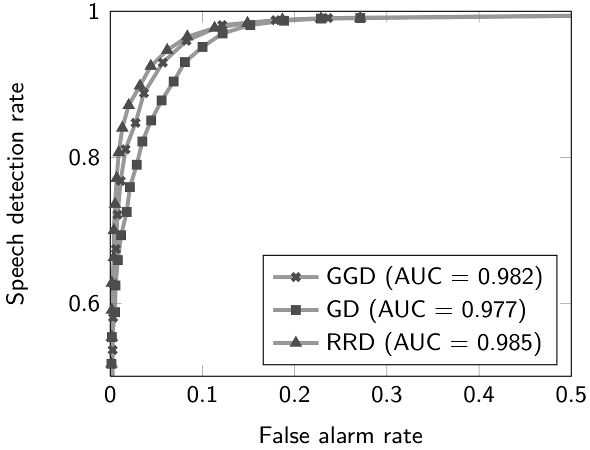


Figure: Clean speech signal

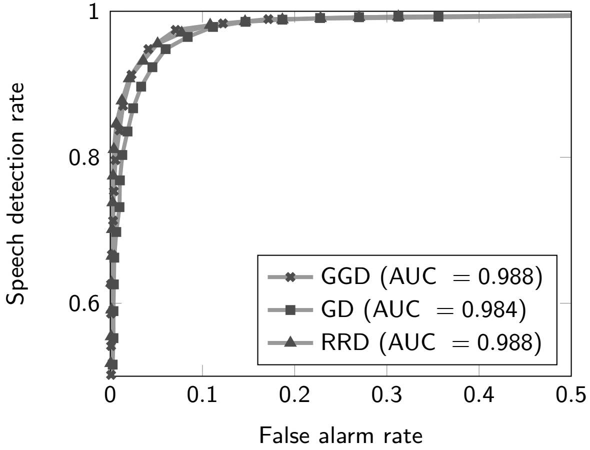


Figure: Speech corrupted with white Gaussian noise at SNR 15 dB

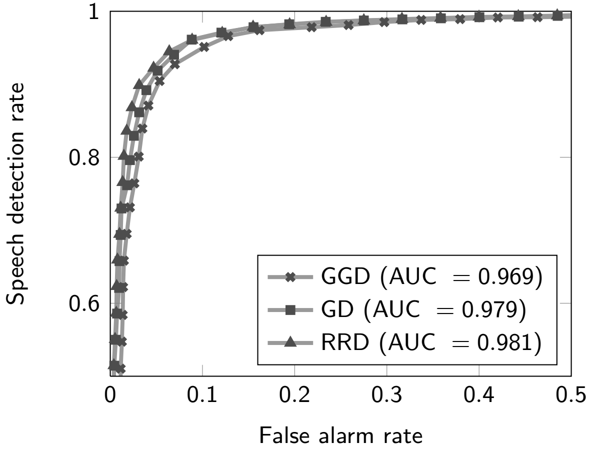


Figure: Speech corrupted with car noise at 10 dB

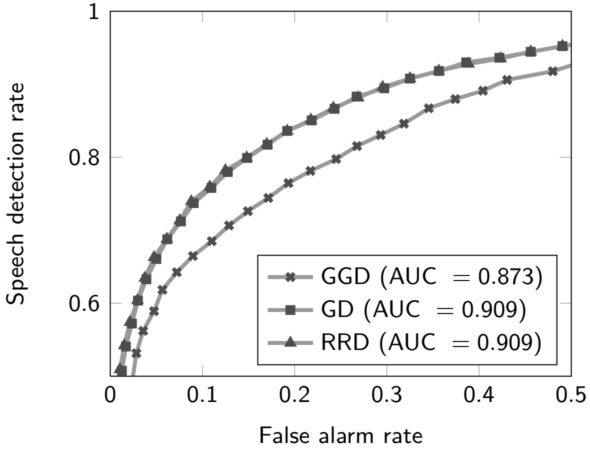


Figure: Speech corrupted with babble noise at 5 dB

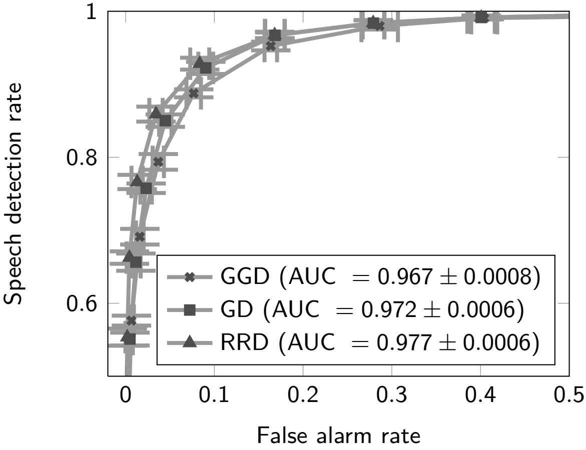


Figure: Threshold averaged ROC curves with area under curve scores

Discussion

- Execution time: GGD 9.70 ms, RRD 0.37 ms, GD 0.21 ms

Discussion

- Execution time: GGD 9.70 ms, RRD 0.37 ms, GD 0.21 ms
- All three VADs showed consistent performance

Discussion

- Execution time: GGD 9.70 ms, RRD 0.37 ms, GD 0.21 ms
- All three VADs showed consistent performance
- By AUC score RRD showed the highest result

Discussion

- Execution time: GGD 9.70 ms, RRD 0.37 ms, GD 0.21 ms
- All three VADs showed consistent performance
- By AUC score RRD showed the highest result

Summary

- Three different statistical model-based VADs: GGD, RRD, GD

Summary

- Three different statistical model-based VADs: GGD, RRD, GD
- Decision based on geometric mean of a likelihood ratio

Summary

- Three different statistical model-based VADs: GGD, RRD, GD
- Decision based on geometric mean of a likelihood ratio
- Experimental analysis on NOIZEUS speech corpus

Summary

- Three different statistical model-based VADs: GGD, RRD, GD
- Decision based on geometric mean of a likelihood ratio
- Experimental analysis on NOIZEUS speech corpus
- Evaluation done with ROC curves and the AUC score

Summary

- Three different statistical model-based VADs: GGD, RRD, GD
- Decision based on geometric mean of a likelihood ratio
- Experimental analysis on NOIZEUS speech corpus
- Evaluation done with ROC curves and the AUC score
- RRD based VAD is the method of choice

Future work

- Combine likelihood ratio with ‘weaker detectors’

Future work

- Combine likelihood ratio with ‘weaker detectors’
- Utilize machine learning algorithms for decision making (SVM, Neural networks, Boost)

Future work

- Combine likelihood ratio with ‘weaker detectors’
- Utilize machine learning algorithms for decision making (SVM, Neural networks, Boost)
- Perform input variable analysis

Thank you for your attention

Questions?

Chang, J.-H., Shin, J. W., and Kim, N. S. (2004).
Voice Activity Detector Employing Generalised Gaussian Distribution.
Electronics Letters, 40(24):25–26.

Cohen, I. (2003).
Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging.
IEEE Transactions on Speech and Audio Processing, 11(5):466–475.

Ephraim, Y. and Malah, D. (1984).
Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator.
IEEE Transactions on Acoustics Speech and Signal Processing, 32(6):1109–1121.

Hu, Y. and Loizou, P. C. (2007).
Subjective Comparison and Evaluation of Speech Enhancement Algorithms.
Speech Communication, 49(7):588–601.

Mumolo, E., Nolich, M., and Verchelli, G. (2003).
Algorithms for Acoustic Localization Based on Microphone Array in Service Robotics.
Robotics and Autonomous Systems, 42(2):69–88.

Sohn, J., Kim, N. S., and Sung, W. (1999).
A Statistical Model-Based Voice Activity Detection.
IEEE Signal Processing Letters, 6(1):1–3.