Multichannel Voice Detection in Adverse Environments

J. Rosca, R. Balan, N.P. Fan
Siemens Corporate Research
Department of Multimedia and Video Technology
755 College Road East
Princeton, NJ 08540
rosca,rvbalan,fan@scr.siemens.com

C. Beaugeant, V. Gilg
Siemens AG
ICM - Mobile Phones
Grillparzerstr. 10
D-81675 Munich , Germany
Christophe.Beaugeant@mch.siemens.de

ABSTRACT

Detecting when voice is or is not present is an outstanding problem for speech transmission, enhancement and recognition. Here we present a novel multichannel source activity detector that exploits the spatial localization of the target audio source. The detector uses an array signal processing technique to maximize the signal-to-interference ratio for the target source thus decreasing the activity detection error rate. We compare our two-channel voice activity detector (VAD) with the AMR voice detection algorithms on real data recorded in a noisy car environment. The new algorithm shows improvements in error rates of 55-70% compared to the state-of-the-art adaptive multi-rate algorithm AMR2 used in present voice transmission technology.

1 Introduction

The voice (and more generally acoustic source) activity detection (VAD) is a cornerstone problem in signal processing practice and often it has a stronger influence on the overall performance of a system than any other component. Speech coding, multimedia communication (voice and data), speech enhancement in noisy conditions and speech recognition are important applications where a good VAD can substantially increase the performance of the respective system. The role of a VAD is basically to extract features of the signal that emphasize differences between speech and noise and then classify them to take a final VAD decision. The variety and the varying nature of speech and background noises makes the VAD problem challenging.

Traditionally, VADs use energy criteria such as SNR estimation based on long-term noise estimation [1]. Improvements proposed use a statistical model of the signal and derive the likelihood ratio [2] or compute the kurtosis [3]. Alternatively, methods attempt to extract robust features (e.g. the presence of a pitch [4], the formant shape [5], or the cepstrum [6]) and compare them to a speech model. Recently, multiple channel VAD algorithms have been investigated [7, 8, 9] to take advantage of the extra information provided by additional sensors.

In this article we focus on a multi channel VAD algorithm. Spatial localization is the key underlying our scheme, which can be used equally for voice and non-voice signals of interest. We assume the following scenario: the target source (such as a person speaking) is located in a noisy environment, and two or more microphones record the audio mixture. Noise is assumed diffuse, but not necessarily uniform, i.e. the sources of noise are not spatially well-localized, and the spectral coherence matrix may be time-varying. Under this scenario we propose an algorithm that blindly identifies the mixing model and outputs a signal with the largest signal-to-interference-ratio (SIR) possibly obtainable through linear filtering. Although the output signal contains large artifacts and is unsuitable for signal estimation it is ideal for signal activity detection.

In the next section we present the mixing model and main statistical assumptions. Section 3 shows the filter derivations and presents the overall VAD architecture. Section 4 addresses the blind model identification problem. Section 5 discusses the evaluation criteria used and section 6 discusses implementation issues and experimental results on real data.

2 Mixing Model and Statistical Assumptions

The time-domain mixing model assumes D microphone signals $x_1(t), \ldots, x_D(t)$, which record a source $s(t)$ and noise signals $n_1(t), \ldots, n_D(t)$:

$$x_i(t) = \sum_{k=0}^{L_i} a_k^i s(t - \tau_k^i) + n_i(t), \quad i = 1, ..., D. \tag{1}$$

where (a_k^i, τ_k^i) are the attenuation and delay on the k^{th} path to microphone i.

In frequency domain, convolutions become multiplications. Furthermore, since we are not interested in balancing the channels, we redefine the source so that the first channel becomes unity:

$$\begin{aligned} X_1(k,\omega) &= S(k,\omega) + N_1(k,\omega) \\ X_2(k,\omega) &= K_2(\omega) S(k,\omega) + N_2(k,\omega) \\ &\cdots \\ X_D(k,\omega) &= K_D(\omega) S(k,\omega) + N_D(k,\omega) \end{aligned} \tag{2}$$

where k is the frame index, and ω the frequency index.

More compactly, this model can be rewritten as:

$$X = KS + N \tag{3}$$

where X, K, N are complex vectors.

We make the following assumptions: (1) The source signal $s(t)$ is independent of the noise signals $n_i(t)$, for all i; (2) The mixing parameters $K(\omega)$ are either time-invariant, or slowly time-varying; (3) $S(\omega)$ is a zero-mean stochastic process with spectral power $\rho_s(\omega) = \mathbf{E}[|S|^2]$; (4) $(N_1, N_2, \ldots, N_D)$ is a zero-mean stochastic signal with spectral covariance matrix $R_n(\omega)$.

3 Algorithm Design

In this section we obtain the optimal-gain filter, and then we present the overall system architecture.

A linear filter A applied on X produces:

$$Z = AX = AKS + AN$$

We look for the linear filter that maximizes the SNR (SIR). The (output) SNR achieved by A is:

$$oSNR = \frac{\mathbf{E}[|AKS|^2]}{\mathbf{E}[|AN|^2]} = \frac{\rho_s AKK^* A}{AR_n A^*} \tag{4}$$

Maximizing $oSNR$ over A results in a generalized eigenvalue problem: $AR_n = \lambda AKK^*$, whose maximizer can be obtained based on the Rayleigh quotient theory [10]:

$$A = \mu K^* R_n^{-1}$$

where μ is an arbitrary nonzero scalar. This expression suggests to run the output Z through an energy detector with an input dependent threshold in order to decide whether the source signal is present or not in the current data frame. The detection decision becomes:

$$VAD(k) = \begin{cases} 1 & if \quad |Z|^2 \geq B|X|^2 \\ 0 & if \quad otherwise \end{cases} \tag{5}$$

where $B > 0$ is a constant boosting factor. Since on the one hand A is determined up to a multiplicative constant, and on the other hand we want to maximize the output energy when the signal is present, we choose $\mu = R_s$, the estimated signal spectral power. The filter we use becomes:

$$A = \rho_s K^* R_n^{-1} \tag{6}$$

Now we can present the overall architecture of our VAD, as in Figure 1. The VAD is based on equations 5 and 6. We assumed that K, ρ_s, R_n are estimated from data, as will be described next.

4 Mixing Model Identification

Here we present estimators for the transfer function ratios K and spectral power densities ρ_s and R_n. We also use the most recently available VAD signal.

4.1 Adaptive Model-based Estimator of K

The adaptive estimator of K makes use of the *direct path* mixing model to reduce the number of parameters:

$$K_l(\omega) = a_l e^{i\omega\delta_l} \ , \ l \geq 2 \tag{7}$$

We choose parameters (a_l, δ_l) that best fit into

$$R_x(k,\omega) = \rho_s(k,\omega) KK^* + R_n(k,\omega) \tag{8}$$

Fitting uses the Frobenius norm. Thus we have to minimize:

$$I(a_2, .., a_D, \delta_2, .., \delta_D) = \sum_{\omega} trace\{(R_x - R_n - \rho_s KK^*)^2\} \tag{9}$$

Summation above is across frequencies because the same parameters $(a_l, \delta_l)_{2\leq l\leq D}$ should explain all frequencies. The gradient of I evaluated on the current estimate $(a_l, \delta_l)_{2\leq l\leq D}$ is:

$$\frac{\partial I}{\partial a_l} = -4\sum_{\omega} \rho_s \cdot \mathrm{real}(K^* E v_l) \tag{10}$$

$$\frac{\partial I}{\partial \delta_l} = -2a_l \sum_{\omega} \omega\rho_s \cdot \mathrm{imag}(K^* E v_l) \tag{11}$$

where $E = R_x - R_n - \rho_s KK^*$ and v_l the D-vector of zeros everywhere except on the l^{th} entry where it is $e^{i\omega\delta_l}$, $v_l = [0 \ \cdots \ 0 \ \ e^{i\omega\delta_l} \ \ 0 \ \cdots \ 0]^T$. Then the updating rule is given by:

$$a_l' = a_l - \alpha\frac{\partial I}{\partial a_l} \tag{12}$$

$$\delta_l' = \delta_l - \alpha\frac{\partial I}{\partial \delta_l} \tag{13}$$

with $0 \leq \alpha \leq 1$ the learning rate.

4.2 Estimation of Spectral Power Densities

The estimation of R_n is done based on the VAD signal simply by:

$$R_n = \begin{cases} (1-\beta)R_n^{old} + \beta XX^* & \text{if} \quad \text{voice present} \\ R_n^{old} & \text{if} \quad \text{otherwise} \end{cases} \tag{14}$$

The signal spectral power ρ_s is estimated through spectral subtraction. The estimate we use is:

$$\rho_s = \begin{cases} R_{x;11} - R_{n;11} & \text{if} \quad R_{x;11} > \beta_{SS} R_{n;11} \\ (\beta_{SS} - 1)R_{n;11} & \text{if} \quad \text{otherwise} \end{cases} \tag{15}$$

where $\beta_{SS} > 1$ is a floor-dependent constant.

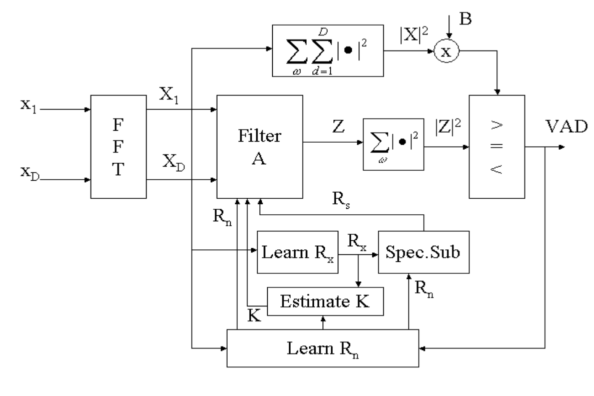

Figure 1: Two-channel VAD block scheme.

5 VAD Performance Criteria

We first define the possible errors that can be obtained when comparing the VAD signal with the true source presence signal. Errors take into account the "context" of the VAD prediction, i.e. the true VAD state (desired signal present or absent) before and after the state of the present data frame as follows (see Figure 2): (1) Noise detected as useful signal (e.g. speech); (2) Noise detected as signal before the true signal actually starts; (3) Signal detected as noise in a true noise context; (4) Signal detection delayed at the beginning of signal; (5) Noise detected as signal after the true signal subsides; (6) Noise detected as signal in between frames with signal presence; (7) Signal detected as noise at the end of the active signal part, and (8) Signal detected as noise during signal activity.

The literature is mostly concerned with four error types showing that speech is misclassified as noise (types 3,4,7,8 above). Some only consider errors 1,4,5,8: these are called "noise detected as speech" (1), "front-end clipping" (2), "noise interpreted as speech in passing from speech to noise" (5), and "mid-speech clipping" (8) in [11].

Our evaluation aims at assessing VAD in three problems (1) Speech transmission/coding, where error types 3,4, 7, and 8 should be as small as possible so that speech is rarely if ever clipped and all data of interest (voice but noise) is transmitted; (2) Speech enhancement, where error types 3,4, 7, and 8 should be as small as possible, nonetheless errors 1,2,5 and 6 are also weighted in depending on how noisy and non-stationary noise is in common environments of interest; and (3) Speech recognition (SR), where all errors are taken into account. In particular error types 1,2,5 and 6 are important for non-restricted SR. A good classification of background noise as non-speech allows SR to work effectively on the frames of interest.

6 Experimental Results

We compare three VAD algorithms: (1-2) Implementations of two adaptive multi-rate (AMR) algorithms, as described in [4], targeting discontinuous transmission of voice; (3) Two-Channel (TwoCh) VAD following the approach described in this paper. We evaluated the algorithms on real data recorded in a car environment in two setups, where the two sensors are either closeby or distant. For each case car noise while driving was recorded separately and additively superimposed on car voice recordings from static situations. The average input SNR for the "medium noise" test suite was zero dB for the closeby case, and -3dB for the distant case. In both cases, we also considered a second test suite "high noise" where the input SNR dropped another 3dB.

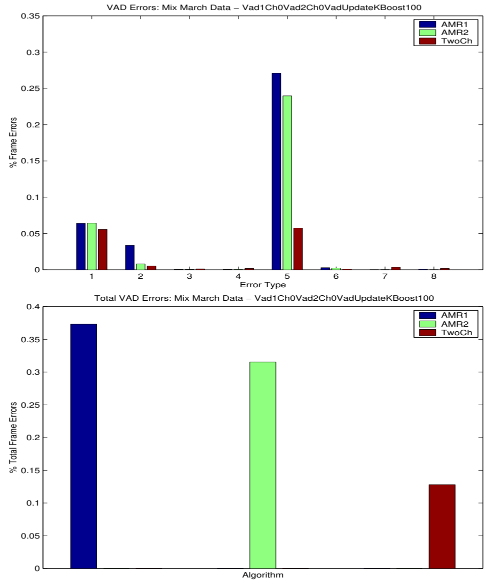

Figure 3: Frame error rates by error type and total error for medium noise, distant microphone scenario.

6.1 Algorithm Implementation

The implementation of the AMR1 and AMR2 algorithms is based on the GSM AMR speech encoder version 7.3.0 [12]. The VAD algorithms use results calculated by the encoder, which may depend on the encoder input mode, therefore a fixed mode of MRDTX was used here. The algorithms indicate whether each 20 ms frame (160 samples frame length at 8kHz) contains signals that should be transmitted, i.e. speech, music or information tones. The output of the VAD algorithm is a boolean flag indicating presence of such signals.

We have implemented the TwoCh VAD based on the MaxSNR filter, adaptive model-based K estimator and spectral power density estimators as presented before (5,10,11,14,15) . We used a boost factor $B = 100$, the learning rates $\alpha = 0.01$ (in K estimation), $\beta = 0.2$ (for R_n), and $\beta_{SS} = 1.1$ (in Spectral Subtraction). Processing was done block wise with a frame size of 256 samples and a time step of 160 samples.

6.2 Results

We obtained "ideal" VAD labeling on car voice data only with a simple power level voice detector. Then we obtained overall VAD errors with the three algorithms under study. Errors represent the average percentage of frames with decision different from ideal VAD relative to the total number of frames processed.

Figures 3 and 4 present individual and overall errors obtained with the three algorithms in the medium and high noise scenarios. Table 1 summarizes average results obtained when comparing the TwoCh VAD with AMR2. Note that in the described tests, the mono AMR algorithms utilized the best (highest SNR) of the two channels (which was chosen by hand).

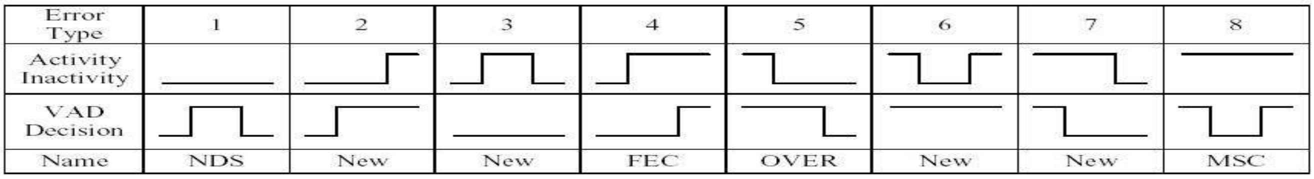

Figure 2: Types of errors considered for evaluating VAD algorithms.

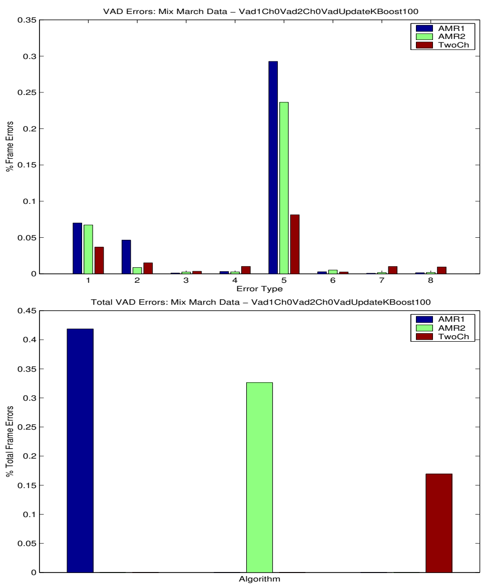

Figure 4: Frame error rates by error type and total error for high noise, distant microphone scenario.

TwoCh VAD is superior to the other approaches when comparing error types 1,4,5, and 8, used for example in [11] and other reports. In terms of errors of type 3,4,7, and 8 only, AMR2 has a slight edge over the TwoCh solution which really uses no special logic or hangover scheme to enhance results. However, with different settings of parameters (particularly the boost factor) TwoCh VAD becomes competitive with AMR2 on just this subset of errors. We expect it to perform better with the suggested improvements. Nonetheless, in terms of overall error rates, TwoCh VAD was clearly superior to the other approaches. This indicates the two channel VAD is a viable detector particularly for speech recognition or speech enhancement scenarios.

Data	Med.Noise	High Noise
Best mic (closeby)	54.5	25
Worst mic (closeby)	56.5	29
Best mic (distant)	65.5	50
Worst mic (distant)	68.7	54

Table 1: Percentage improvement in overall error rate over AMR2 for the two-channel VAD across two data and microphone configurations.

7 Conclusions

The paper presented a novel multichannel source activity detector that exploits the spatial localization of the target audio source. The implemented detector maximizes the signal-to-interference ratio for the target source and uses two channel input data. We compare our two channel VAD with the AMR VAD algorithms on real data recorded in a noisy car environment. The two channel algorithm shows improvements in error rates of 55-70% compared to the state-of-the-art adaptive multi-rate algorithm AMR2 used in present voice transmission technology. Future enhancement of the algorithm will explore parameter optimization and post-processing decision enhancement based on a VAD dependent state.

References

[1] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. Of the IEEE Speech Coding Workshop*, Oct 1993, pp. 85–86.

[2] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proceedings ICASSP*. 2001, IEEE Press.

[3] R.Goubran E.Nemer and S. Mahmoud, "Snr estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Processing Letters*, vol. 6, no. 7, pp. 171–174, July 1999.

[4] "Digital cellular telecommunication system (phase 2+); voice activity detector for adaptive multi-rate (amr) speech traffic channels," ETSI Report, DEN/SMG-110694Q7, 2000.

[5] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, pp. 237–240.

[6] J.A. Haigh and J.S. Mason, "Robust voice activity detection using cepstral features," *IEEE TEN-CON*, pp. 321–324, 1993.

[7] P. Naylor N. Doukas and T. Stathaki, "Voice activity detection using source separation techniques," in *Proceedings Eurospeech*, 1997, pp. 1099–1102.

[8] J. F. Chen and W. Ser, "Speech detection using microphone array," *Electronics Letters*, vol. 36(2), pp. 181–182, 2000.

[9] Q. Zou, X. Zou, M. Zhang, and Z. Lin, "A robust speech detection algorithm in a microphone array teleconferencing system," in *Proceedings ICASSP*. 2001, IEEE Press.

[10] G.H. Golub and C.F.van Loan, *Matrix Computations*, The John Hopkins University Press, 1989.

[11] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and compatrison of itu-t/etsi voice activity detectors," in *Proceedings ICASSP*. 2001, IEEE Press.

[12] "3gpp ts mandatory speech codec speech processing functions, amr speech codec, voice activity detector (vad)," Tech. Rep. Release 4, 26.094 v4.0.0, March 2001.