A pattern recognition approach to VAD using modified group delay

R. Padmanabhan*, Sree Hari Krishnan P.† and Hema A. Murthy*
*Department of Computer Science and Engineering
Indian Institute of Technology Madras
Email: padmanabhan@lantana.tenet.res.in, hema@lantana.tenet.res.in
† IDIAP Research Institute, Martigny, Switzerland
†École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Email: Hari.Parthasarathi@idiap.ch

***Abstract*—This paper explores the use of phase-based features (in particular, group delay) for voice activity detection (VAD). We establish via theoretical analysis the robustness of the group delay function in noise. Based on this, we extract group delay based features and pose the VAD problem as a two-class classification task. Two trained classifiers, namely Gaussian mixture models (GMM) and support vector machines (SVM) are evaluated for VAD. Both methods are compared with standard VAD algorithms and are found to perform better, particularly in low SNRs.**

I. Introduction

The voice activity detection (VAD) problem seeks to identify speech and non-speech regions in a given speech signal. A VAD subsystem is a preprocessor to several speech processing systems including speech coders and automatic speech recognisers. In these systems, channel utilisation and recogniser accuracy respectively are improved significantly with a good VAD front end. Traditional methods for VAD have used threshold-based measurements of features like short-term energy and zero crossing rate. These methods do not work well in low SNR conditions. Other approaches to the VAD problem exploit statistical properties of speech and non-speech [1],[2],[3],[4],[5].

Traditionally, most spectrum-based methods for VAD have used features derived from the Fourier transform magnitude spectrum. The group delay function, which is derived from the Fourier transform phase, has been used for various speech processing applications in [6],[7],[8] and has been found to be useful in VAD as well. The group delay function is not well behaved for signals that are not minimum phase. To address this condition, algorithms using group delay have used two techniques: minimum-phase group delay and modified group delay.

A VAD algorithm using minimum phase group delay was described in [9]. This algorithm, however, had a latency equal to signal length. To reduce latency, a buffering scheme was proposed in [10], which enabled the algorithm to work on short segments of the signal. VAD decisions were made after a certain number of frames were buffered, and the latency was reduced to about 150 ms.

This paper describes two VAD algorithms that use the modified group delay[1] (MODGD). The modified group delay is well behaved even for signals that are not minimum phase, and is described in [11]. Both algorithms make frame-wise VAD decisions, which was not the case of the algorithm described in [10].

Both algorithms pose the VAD problem as a two-class pattern classification task, the two classes being (noisy) speech and non-speech. MODGD features show good discrimination between speech and non-speech even at low SNRs [12]. A trained classifier can thus accurately distinguish between the two classes in the MODGD space. Since the MODGD has a size equal to the FFT size, we decorrelate the MODGD to reduce the dimension of the feature space. The two classifiers evaluated in the decorrelated MODGD space are support vector machines (SVM) and Gaussian mixture models (GMM). The performance of the proposed technique is compared against standard VAD algorithms.

The rest of this paper is organised as follows. Section II analyses the robustness of group delay in noise. Section III describes the feature extraction procedure and describes the two classifiers. The experiments done are described in section IV, and the results are discussed in section V. Finally, we conclude in section VI.

II. Analysis of group delay in noise

In this section, we analytically show why group delay functions are robust to noise. This section follows the work of one of the authors in [12].

Let $x[n]$ denote a clean speech signal degraded by uncorrelated, zero-mean, additive noise $v[n]$. Then, the noisy speech, $y[n]$, can be expressed as,

$$y[n] = x[n] + v[n] \tag{1}$$

Taking the Fourier transform, we have

$$Y(\omega) \quad = \quad X(\omega) + V(\omega) \tag{2}$$

Multiplying by corresponding complex conjugates and taking the expectation, we have the power spectrum

$$P_Y(\omega) = P_X(\omega) + \sigma^2(\omega) \tag{3}$$

[1]The modified group delay is also called the modified group delay function or modified group delay spectrum.

where we have used the fact that the expectation of noise is zero. The power spectra of the resulting noisy speech signal can be related to noise power and (clean) speech power in one of three mutually exclusive frequency regions: (i) the high noise power case where $P_X(\omega) \ll \sigma^2(\omega)$ (ii) the high signal power case where $P_X(\omega) \gg \sigma^2(\omega)$ and (iii) the equal power case where $P_X(\omega) \approx \sigma^2(\omega)$. Following the same notation as in [12], the power spectra of the noisy speech signal in each case are denoted respectively as $P_Y^n(\omega)$, $P_Y^s(\omega)$ and $P_Y^e(\omega)$. We analyse the group delay representation of noisy speech in the three cases mentioned above.

A. High noise power spectral regions ($P_Y^n(\omega)$)

In this subsection, we consider frequencies ω such that $P_X(\omega) \ll \sigma^2(\omega)$, i.e., regions where the noise power is higher than signal power. From Equation 3 we have

$$\begin{aligned} P_Y^n(\omega) &= P_Y(\omega) \quad \forall \omega \quad \text{s.t.} \quad P_X(\omega) \ll \sigma^2(\omega) \\ &= P_X(\omega) + \sigma^2(\omega) \\ &= \sigma^2(\omega)\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right) \end{aligned}$$

Taking logarithms on both sides, using the Taylor series expansion[2] of $\ln(1 + \frac{P_X(\omega)}{\sigma^2(\omega)})$, and ignoring the higher order terms,

$$\begin{aligned} \ln\left(P_Y^n(\omega)\right) &= \ln\left[\sigma^2(\omega)\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right)\right] \\ &= \ln\left(\sigma^2(\omega)\right) + \ln\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right) \\ &\approx \ln\left(\sigma^2(\omega)\right) + \frac{P_X(\omega)}{\sigma^2(\omega)} \end{aligned} \tag{4}$$

Expanding $P_X(\omega)$ as a Fourier series ($P_X(\omega)$ is a periodic, continuous, function of ω with a period $\omega_0 = 2\pi$),

$$\ln\left(P_Y^n(\omega)\right) \approx \ln\left(\sigma^2(\omega)\right) + \frac{1}{\sigma^2(\omega)}\left[\frac{d_0}{2} + \sum_{k=1}^{\infty} d_k \cos\left(\frac{2\pi}{\omega_0}\,\omega\, k\right)\right] \tag{5}$$

where, d_k are the Fourier series coefficients in the expansion of $P_X(\omega)$. Since $P_X(\omega)$ is an even function, coefficients of the sine terms are zero.

For a minimum phase signal, the group delay function can be computed in terms of the cepstral coefficients of the log-magnitude spectrum, as given in [7],

$$\begin{aligned} \log|X(\omega)| &= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(\omega\, k) \\ \tau(\omega) &= \sum_{k=1}^{\infty} k\, a_k \cos(\omega\, k) \end{aligned} \tag{6}$$

where, τ is the group delay function and a_k are the cepstral coefficients. From (6), it can be observed that the group delay function can be obtained from the log-magnitude response by ignoring the dc term, and by multiplying each coefficient with k. Applying this observation to Equation (5), we get the group delay function as:

$$\tau_{Y^n}(\omega) \approx \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \tag{7}$$

This expression shows that the group delay function is inversely proportional to the noise power ($\sigma^2(\omega)$) in regions where noise power is greater than the signal power.

B. High signal power spectral regions ($P_Y^s(\omega)$)

Now consider frequencies ω such that $P_X(\omega) \gg \sigma^2(\omega)$. Starting with Equation (3), and following the steps similar to those in previous subsection:

$$\ln\left(P_Y^s(\omega)\right) \approx \ln\left(P_X(\omega)\right) + \frac{\sigma^2(\omega)}{P_X(\omega)} \tag{8}$$

Since $P_X(\omega)$ is non-zero, continuous, and periodic in ω, $\frac{1}{P_X(\omega)}$ is also periodic and continuous. Consequently, $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ can be expanded using Fourier series, giving

$$\ln\left(P_Y^s(\omega)\right) \approx \frac{d_0 + \sigma^2(\omega)\, e_0}{2} + \sum_{k=1}^{\infty}\left(d_k + \sigma^2(\omega)\, e_k\right)\cos(\omega\, k)$$

Using the properties of group delay function listed in Equation (6), and following the steps in the previous case[3], we obtain the expression for the group delay function as,

$$\tau_{Y^s}(\omega) \approx \sum_{k=1}^{\infty} k\left(d_k + \sigma^2(\omega)\, e_k\right)\cos(\omega\, k) \tag{9}$$

where d_k and e_k are the Fourier series coefficients of $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ respectively. It is satisfying to observe that if $\sigma^2(\omega)$ is negligible, the group delay function can be expressed solely in terms of log-magnitude spectrum.

C. Signal power ≈ noise power regions ($P_Y^e(\omega)$)

For frequencies ω such that $P_X(\omega) \approx \sigma^2(\omega)$, we again start with Equation (3), and follow the steps similar to those in previous subsections, except in this case we do not need the Taylor series expansion:

$$\begin{aligned} P_Y^e(\omega) &\approx 2P_X(\omega) \\ \ln\left(P_Y^e(\omega)\right) &\approx \ln 2 + \ln\left(P_X(\omega)\right) \end{aligned} \tag{10}$$

Expanding $\ln\left(P_X(\omega)\right)$ as a Fourier series, since it is a periodic, continuous, function of ω with a period 2π, the group delay function can be computed as,

$$\tau_{Y^e}(\omega) \approx \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \tag{11}$$

where d_k are the Fourier series coefficients of $\ln(P_X(\omega))$.

[2]Taylor series expansion of $\ln(1+x)$ is: $\ln(1+x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} x^{n+1} \quad |x| < 1$

[3]Ignoring the dc term, and multiplying each coefficient with k

D. Behaviour of minimum phase group delay functions in noise

From Equations 7, 9, and 11, the estimated group delay functions are summarised respectively for the three cases:

$$\tau(\omega) \approx \begin{cases} \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \\ \sum_{k=1}^{\infty} k\, (d_k + \sigma^2(\omega)\, e_k) \cos(\omega\, k) \\ \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \end{cases} \tag{12}$$

where the first case is for $\forall\omega$ such that $P_X(\omega) \ll \sigma^2(\omega)$, the second for $\forall\omega$ such that $P_X(\omega) \gg \sigma^2(\omega)$, and the third for $\forall\omega$ such that $P_X(\omega) \approx \sigma^2(\omega)$. From Equation 12, we note that the group delay function of a minimum phase signal is *inversely* proportional to the noise power for frequencies corresponding to high noise regions in the power spectrum. Similarly, for low noise regions, from Equation 9, the group delay function becomes *directly* proportional to the signal power. In other words, its behaviour is similar to that of the magnitude spectrum. This shows that the group delay function of a minimum phase signal preserves the peaks and valleys in the magnitude spectrum well even in the presence of additive noise.

E. The modified group delay function

Practically, a frame of speech is typically non-minimum phase, due to the zeros introduced by nasals, pitch and the analysis window. Thus, the above analysis is directly applicable only to the minimum phase components derived from speech signals. To overcome this, we use the modified group delay (MODGD), which is an approximation to the minimum phase group delay. Using the modified group delay enables computation of the group delay even when the signal is not minimum phase [11].

F. The modified group delay feature

The modified group delay feature or MODGDF (also called modified group delay cepstra) is formed by converting the modified group delay (MODGD) into cepstral features using the discrete cosine transform [11]. This results in features that are linearly decorrelated. When compared to MODGD features, MODGDF features can be of considerably lower dimension. Experimental observations reveal that MODGDF features have different characteristics for speech and non-speech.

III. VAD as a pattern classification task

The previous section established that spectral features are retained for speech and non-speech regions in the MODGDF domain, even at low SNRs. This encourages the use of standard classifiers to discriminate between a frame of speech and a frame of non-speech.

A. Feature extraction

For each frame of speech, the MODGDF feature is extracted as given in [11]:

1) Compute the DFT of the speech signal $x[n]$ as $X[k]$.
2) Next, the DFT of the speech signal $n\,x[n]$ is computed as $Y[k]$.
3) Compute the cepstrally smoothed spectra of $X[k]$ and denote it as $S[k]$.
4) Compute the MODGD feature as:

$$\tau_m[k] = (\frac{\tau[k]}{|\tau[k]|})(|\tau[k]|)^{\alpha} \tag{13}$$

where $\tau[k] = \frac{X_R[k]Y_R[k]+X_I[k]Y_I[k]}{|S[k]|^{2\gamma}}$ with the parameters α and γ being set to 1.

5) Compute the MODGDF feature by taking the DCT:

$$c[n] = \sum_{k=0}^{k=N_f} \tau_m[k] \cos(n(2k+1)\pi/N_f) \tag{14}$$

where N_f is the DFT size. Also, for every frame, the feature vector is averaged with the feature vectors of the past two frames.

B. Using Gaussian mixture models

A simple classifier, namely Gaussian mixture models (GMMs) can be used to classify MODGDF features extracted from speech or non-speech. During the training phase, GMM models are built for speech and non-speech using training data. During the testing phase, MODGDF features extracted from a frame of speech are classified as speech or non-speech using the GMM models. A VAD decision is thus made for a given frame of speech. The above algorithm is called GMM-VAD.

C. Using support vector machines

Following [13], a support vector machine (SVM) was trained on labelled MODGDF features derived from training examples of the two classes. Similar to GMM-VAD, in the testing phase, MODGDF features extracted from a frame of speech are classified as speech or non-speech using the SVM models. This algorithm is called SVM-VAD.

IV. Experimental setup

To compare results, the same experimental setup as in [10] was used. 432 speech files (216 female, 216 male), were obtained by concatenating sets of three individual speech utterances from TIMIT [14]. To model the typical speech activity over a telephone conversation, silence was inserted so that the ratio of silence to speech is 60:40 [15]. Three different types of noise (babble, pink and white) from the NOISEX-92 [16] database were added resulting in twelve test-sets, each having 0, 5, 10, or 15 dB SNR respectively.

A. Evaluation of GMM-VAD

The number of cepstral coefficients used in the MODGDF was 32. The training data had an overall SNR of 10 dB. Speech–non-speech GMM pairs were built for each noise type and the test data was evaluated for each. The optimum number of mixtures for each GMM were determined experimentally. The results are given in Table III.

B. Evaluation of SVM-VAD

The SVM implementation used was SVM-Torch [17]. Three SVM models were built for the three different noise types. The training data had an overall SNR of 10 dB. Optimum VAD performance is obtained when speech frames corrupted with a particular type of noise is classified using that particular SVM model. The results are tabulated in Table III.

V. Results and discussions

The performance of the algorithms were evaluated by comparing the percentage of correct classifications (non-speech and speech) with manually marked decisions on all the 432 test utterances. The performance metrics P_{cn} (percentage of correct non-speech identification ie. insertion error), P_{cs} (percentage of correct speech identification ie. clipping error) and P_f (percentage of false detection) described in [15] are defined below.

$$P_{cn} = \frac{\text{No. of non-speech frames from algorithm} \times 100}{\text{No. of non-speech frames in manual VAD}}$$

$$P_{cs} = \frac{\text{No. of speech frames from algorithm} \times 100}{\text{No. of speech frames in manual VAD}}$$

$$P_f = \frac{\text{No. of misclassified frames by algorithm} \times 100}{\text{Total no. of frames}}$$

The performance of the modified group delay based GMM-VAD and SVM-VAD algorithms are tabulated in Table III. The performance of three standard VAD algorithms, namely G.729 Annexe B VAD, AMR VAD option 1 and AMR VAD option 2 are given in Table II.

The results are summarised below:

1) The group delay based VAD algorithms SVM-VAD and GMM-VAD outperform the standard algorithms G.729B, AMR VAD option 1 and AMR VAD option 2 in babble noise.
2) SVM-VAD performs comparably with AMR VAD option 2 in pink and white noise, and outperforms G.729B and AMR option 1.
3) GMM-VAD performs comparably with AMR VAD option 1 in pink and white noise and outperforms G.729B.

The average error for each method in each noise type and the relative computation speed is given in Table I. The time taken by the fastest method is denoted by 1. The other times are in multiples of this. The SVM based method requires computation of support vectors in the kernel space for each frame and hence requires more computation. At the time of this writing, it is not clear why group delay based features work well in babble noise. Future investigations will address this issue.

VI. Conclusion

In this paper, the VAD problem was viewed as a classification task in the modified group delay space. An analysis of the group delay based representation showed its robustness to noise. Two algorithms, using SVMs and GMMs respectively, were described and evaluated against standard VAD algorithms. In babble noise, the group delay based algorithms outperform the standard algorithms, whereas in pink and white noise their performances are better or comparable. This demonstrates the robustness of phase-based methods to noise for a feature discrimination task. The paper also shows that pattern classification approaches to VAD show encouraging results.

TABLE I
Summary of VAD schemes.

Method	VAD Error			Speed
	$E_{babble}\%$	$E_{white}\%$	$E_{pink}\%$	
G.729B	31.75	17.18	16.37	1.04
AMR1	22.01	10.97	9.58	1
AMR2	32.11	5.57	5.21	1.08
GMM	7.17	8.33	8.58	1.26
SVM	5.75	6.21	6.32	6.87

TABLE III
Performance of proposed VAD methods in different noise environments.

Method	GMM-VAD			SVM-VAD		
SNR (dB)	$P_{cn}\%$	$P_{cs}\%$	$P_f\%$	$P_{cn}\%$	$P_{cs}\%$	$P_f\%$
Babble noise environment						
0	95.86	79.26	10.88	96.19	85.49	8.15
5	95.93	88.43	7.11	96.00	92.82	5.29
10	95.04	93.85	5.43	95.40	96.48	4.15
15	93.43	96.65	5.26	92.03	98.35	5.39
Average	**95.07**	**89.55**	**7.17**	**94.90**	**93.29**	**5.75**
White noise environment						
0	90.57	85.77	11.37	96.68	80.64	9.83
5	91.95	90.22	10.11	96.74	88.57	6.57
10	92.65	94.11	6.84	96.35	94.62	4.34
15	90.59	97.86	6.44	94.59	97.84	4.08
Average	**91.44**	**91.99**	**8.33**	**96.10**	**90.42**	**6.21**
Pink noise environment						
0	91.31	80.41	13.11	96.55	80.59	9.93
5	94.26	85.22	9.41	97.14	87.74	6.67
10	95.15	91.05	6.52	96.96	93.18	4.57
15	94.44	95.09	5.29	95.39	96.58	4.12
Average	**93.79**	**87.94**	**8.58**	**96.51**	**89.59**	**6.32**
Average environment						
Avg	**93.43**	**89.82**	**8.03**	**95.83**	**91.10**	**6.09**

References

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, 1999.

[2] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett*, vol. 8, pp. 276–278, 2001.

[3] J. Sohn and W. Sung, "A Voice Activity Detector employing soft decision based noise spectrum adaptation," *ICASSP*, vol. 1, pp. 365–368, 1998.

[4] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, p. 478482, 2000.

[5] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech Audio Process*, vol. 13, 2005.

[6] H. A. Murthy and B. Yegnanarayana, "Formant extraction from minimum phase group delay function," *Speech Comm*, pp. 209–221, 1991.

[7] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 610–622, Jun. 1984.

TABLE II
PERFORMANCE OF STANDARD VAD METHODS IN DIFFERENT NOISE ENVIRONMENTS.

Method	G.729B VAD			AMR VAD 1			AMR VAD 2		
SNR (dB)	$P_{cn}\%$	$P_{cs}\%$	$P_f\%$	$P_{cn}\%$	$P_{cs}\%$	$P_f\%$	$P_{cn}\%$	$P_{cs}\%$	$P_f\%$
Babble noise environment									
0	65.33	57.32	37.87	47.82	95.75	33.01	44.59	95.66	34.98
5	65.45	68.28	33.42	56.9	97.5	26.86	42.78	99.13	34.68
10	65.33	77.47	29.81	72.27	97.26	17.73	45.49	99.7	32.83
15	65.87	86.43	25.91	84.52	97.15	10.43	57.09	99.55	25.93
Average	**65.5**	**72.38**	**31.75**	**65.38**	**96.92**	**22.01**	**47.49**	**98.51**	**32.11**
White noise environment									
0	89.54	54.27	24.57	87.53	67.58	20.45	94.76	89.05	7.52
5	89.47	66.82	19.59	90.29	84.66	11.96	93.69	96.43	5.21
10	89.33	77.47	15.41	92.4	93.45	7.18	93.23	97.95	4.88
15	93.11	87.43	9.16	95.65	95.78	4.3	93.00	98.79	4.68
Average	**90.36**	**71.5**	**17.18**	**91.47**	**85.37**	**10.97**	**93.67**	**95.56**	**5.57**
Pink noise environment									
0	89.50	60.52	22.09	89.64	69.24	18.52	95.07	91.40	6.4
5	88.84	71.48	18.1	92.64	86.27	9.91	93.67	97.23	4.91
10	89.38	81.43	13.8	94.79	93.08	5.89	93.46	98.31	4.6
15	88.40	88.65	11.5	95.61	96.63	3.98	92.27	99.28	4.93
Average	**89.03**	**75.52**	**16.37**	**93.17**	**86.31**	**9.58**	**93.62**	**96.56**	**5.21**
Average environment									
Average	**81.63**	**73.13**	**21.77**	**83.34**	**89.53**	**14.19**	**78.26**	**96.87**	**14.3**

[8] V. Prasad, T.Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Comm*, vol. 42, pp. 429–446, 2004.

[9] S. H. K. P., R. Padmanabhan, and H. A. Murthy, "Robust voice activity detection using group delay functions," *IEEE ICIT*, 2006.

[10] ——, "Voice activity detection using group delay processing on buffered short-term energy," *NCC*, 2007.

[11] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 190–202, 2007.

[12] S. H. K. P., "Voice activity detection using group delay functions," Master's thesis, Indian Institute of Technology, Madras, 2007.

[13] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," *Proc. 6th International Conference on Signal Process.*, vol. 2, pp. 1124–1127, 2002.

[14] NTIS, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

[15] B. F., C. S., and R. G., "Performance evaluation and comparison of itu-t/etsi voice activity detectors," *Proceedings of ICASSP*, vol. 3, pp. 1425–1428, 2001.

[16] "NOISEX-92," http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html, database of various type of noise available on CD-ROM.

[17] R. Collobert and S. Bengio, "SVMTorch: Support Vector Machines for Large-Scale Regression Problems," *Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.