A DIFFERENTIAL SPECTRAL VOICE ACTIVITY DETECTOR

Philip N. Garner, Toshiaki Fukada and Yasuhiro Komori

Canon Inc.
3-30-2 Shimomaruko, Ota-ku, Tokyo 146-8501, Japan.
Email: philip.garner@canon.co.jp, fukada.toshiaki@canon.co.jp, komori.yasuhiro@canon.co.jp

ABSTRACT

The Voice Activity Detection (VAD) problem is placed into a decision theoretic framework, and the Gaussian VAD model of Sohn *et al.* is then shown to fit well with the framework. It is argued that the Gaussian model can be made more robust to correlation and expected spectral shapes of speech and noise by using a differential spectral representation. Such a model is formulated theoretically. The differential spectral VAD is then shown by experiment to compare favourably with the basic Gaussian VAD in a speech recognition setting, especially for noisy environments.

1. INTRODUCTION

Voice Activity Detection (VAD) is important in various applications involving speech. Perhaps the most common application is in telecommunications, where the main reason for the VAD is to save bandwidth by not transmitting non-speech portions of the input signal. Marzinzik and Kollmeier [1] present a useful recent review of the subject.

We are interested in (VAD) for the purpose of Automatic Speech Recognition (ASR) in general, and noise robust ASR in particular. VAD is important in ASR because it distinguishes the non-speech portions at the beginning and end of an utterance from the utterance itself. In doing this, the VAD ensures that the decoder, which is computationally intensive, only runs when necessary. This point is particularly important in embedded applications, where processing power is limited. The main difference between a VAD used in telecommunications and one used in ASR is that the latter typically uses a state machine in order to avoid false detections and to remain active during speech pauses.

A VAD working in the spectral domain, and with an appealing statistical basis, has been introduced by Sohn *et al.* [2, 3]. This spectral VAD has been shown to be superior to three standard VADs (QCELP, EVRC and G.729B) in a telecommunications environment. That result is reinforced in a comparison by Stadermann *et al.* [4], who find the spectral VAD to be superior to baselines based on frame energy and spectral entropy. The work has also been extended by Cho *et al.* [5], who show that smoothing can alleviate problems with errors in the end of speech region.

The spectral VAD of Sohn *et al.* is based on a simple Gaussian assumption. This basic Gaussian model has two distinct problems:

1. The model has no knowledge of the spectral shape of either the speech or the noise. It is well known, however, that speech has distinct spectral peaks (formants). Conversely, many types of noise have a smooth spectral shape.
2. From a purely statistical point of view, the model assumes that adjacent spectral bins are uncorrelated. This is not true, especially for speech, and even more especially for observations from the overlapped triangular mel-spaced filterbank typical of current ASR systems.

In this paper, we evaluate the basic Gaussian VAD algorithm described above in an ASR context. We then argue that a differential spectral representation can minimize the effects of the two problems described above. We derive a differential spectral version of the Gaussian VAD, and show that it leads to improved performance.

2. BACKGROUND THEORY

2.1. Decision theoretic framework

Define a boolean variable or hypothesis $\mathcal{H}$, which can take values 0 and 1. $\mathcal{H} = 0$ indicates non-speech and $\mathcal{H} = 1$ indicates the presence of speech. A VAD produces an estimate (or choice), $\hat{\mathcal{H}}$, given some observation. For this derivation, assume that the observation is the (complex) spectrum $\mathfrak{s}$.

The above leads to a simple decision theoretic formulation: Define a loss or cost function, $C\left(\mathcal{H}, \hat{\mathcal{H}}\right)$, that attaches a cost to each combination of $\mathcal{H}$ and $\hat{\mathcal{H}}$. Typically, the cost should be low for a correct classification, and high for an incorrect one. The expected costs of the two possible classifications are then

$$E\left(C\left(\mathcal{H},0\right) \mid \mathfrak{s}\right) = \sum_{\mathcal{H}} C\left(\mathcal{H},0\right) P\left(\mathcal{H} \mid \mathfrak{s}\right), \tag{1}$$

$$E\left(C\left(\mathcal{H},1\right) \mid \mathfrak{s}\right) = \sum_{\mathcal{H}} C\left(\mathcal{H},1\right) P\left(\mathcal{H} \mid \mathfrak{s}\right). \tag{2}$$

We can now choose the classification, $\hat{\mathcal{H}}$, that has the smaller expected cost; that is: Choose $\hat{\mathcal{H}} = 1$ if

$$\sum_{\mathcal{H}} C\left(\mathcal{H},1\right) P\left(\mathcal{H} \mid \mathfrak{s}\right) < \sum_{\mathcal{H}} C\left(\mathcal{H},0\right) P\left(\mathcal{H} \mid \mathfrak{s}\right). \tag{3}$$

Expanding the summations and rearranging,

$$\frac{P\left(\mathcal{H}=1 \mid \mathfrak{s}\right)}{P\left(\mathcal{H}=0 \mid \mathfrak{s}\right)} > \frac{C\left(0,1\right) - C\left(0,0\right)}{C\left(1,0\right) - C\left(1,1\right)}. \tag{4}$$

Given that we will assume a model for the generation of $\mathfrak{s}$, it is useful to apply Bayes's theorem to the conditional probabilities in equation 4. Notice that the evidence (denominator of Bayes's

theorem) term cancels, giving

$$\underbrace{\frac{f(\mathfrak{s} \mid \mathcal{H}=1)}{f(\mathfrak{s} \mid \mathcal{H}=0)}}_{\text{Likelihood ratio, } L(\mathfrak{s})} \cdot \underbrace{\frac{P(\mathcal{H}=1)}{P(\mathcal{H}=0)}}_{\text{Prior ratio}} > \underbrace{\frac{C(0,1)-C(0,0)}{C(1,0)-C(1,1)}}_{\text{Cost ratio}}, \quad (5)$$

where we refer to the terms as indicated.

The prior ratio and cost ratio can be set to unity given the following broad assumptions:

- The likelihood of an observation $\mathfrak{s}$ being speech is as likely as it being non-speech.
- The cost of an accurate classification is zero, and the costs of the two inaccurate classifications are identical.

Of course, the above assumptions may not be true for a given scenario, in which case the terms can be set accordingly. The combination of cost ratio and prior ratio into a single threshold term yields the likelihood ratio test used by Sohn *et al.*. The advantage of the decision theoretic approach is that it gives some insight into what the threshold should be.

2.2. Gaussian model

Broadly following Sohn *et al.* [3], but with a minor change of notation to allow subscripts to refer to vector elements, assume that both the speech and noise can be modeled by Gaussian distributions (more accurately, the real and imaginary components of each spectral bin are i.i.d. Gaussian). This is identical to the assumption made in the Ephraim Malah formulation for speech enhancement [6]. We define two probability distributions:

$$f(\mathfrak{s} \mid \mathcal{H}=0) = \prod_{k=1}^{S} \frac{1}{\pi \mu_k} \exp\left(-\frac{s_k^2}{\mu_k}\right), \quad (6)$$

$$f(\mathfrak{s} \mid \mathcal{H}=1) = \prod_{k=1}^{S} \frac{1}{\pi(\lambda_k+\mu_k)} \exp\left(-\frac{s_k^2}{\lambda_k+\mu_k}\right), \quad (7)$$

where $\mathfrak{s}$ is the S dimensional complex spectrum observation, s_k is the magnitude of the k^{th} element of $\mathfrak{s}$, λ_k is the variance of the k^{th} dimension of the speech signal and μ_k is the variance of the k^{th} dimension of the noise signal. All of the above is for a single frame, although the f subscript is omitted for clarity. Equation 7 follows from that fact that the sum of two Gaussian random variates is Gaussian with variance equal to the sum of the individual variances.

Substituting equations 6 and 7 into equation 5 gives a VAD likelihood ratio of

$$L(\mathfrak{s}) = \prod_{k=1}^{S} \frac{\mu_k}{\lambda_k+\mu_k} \exp\left(\frac{\lambda_k}{\lambda_k+\mu_k} \cdot \frac{s_k^2}{\mu_k}\right). \quad (8)$$

Notice that equation 8 is defined in terms of spectral power measures, eventhough the assumptions so far are based on complex spectrum.

2.3. Correction for correlation

When taking a product of probabilities known to be correlated, it is normal to make a simple correction for the correlation in the form of a weighted geometric mean,

$$f(\mathfrak{s}) = \prod_{k=1}^{S} f(\mathfrak{s}_k)^{\frac{1}{\kappa S}}, \quad (9)$$

where κ is an optimised constant analogous to the language model match factor in ASR. Sohn *et al.* do this implicitly by taking the unweighted geometric mean ($\kappa = 1$), although in this framework that is an extreme solution and represents absolute correlation between bins. $\kappa = 1/S$ represents complete independence.

3. DIFFERENTIAL SPECTRAL VAD

We suggest that the single zero high-pass filter (HPF),

$$s_k^{2\prime} = s_{k+1}^2 - s_k^2 \qquad 1 \le k < S, \quad (10)$$

applied in the frequency dimension of each power spectral frame will tackle the problems highlighted in section 1 as follows:

1. The HPF will map the smooth spectrum associated with noise, especially the flat spectrum of white noise or impulse noise, to a flatter spectrum centered around zero. This is much closer to the spectrum of silence.

2. The subtraction will reduce or eliminate the correlation between adjacent spectral bins.

In fact, the decorrelation effect has been demonstrated in the context of robust ASR by Nadeu *et al.* [7], who show that such a filter can be used in place of the cosine transform normally used in ASR.

In the VAD context, however, we require a probability distribution associated with the filter. This is derived as follows:

First, notice that the single zero filter of equation 10 corresponds to a probabilistic change of variable with $1 \le k < S$ and an integral over s_S^2. This integral turns out to be highly non-trivial. Instead, we decimate the above substitution as follows, allowing the problem to be solved as $S/2$ identical and much simpler integrals:

$$s_k^{2\prime} = s_{2k}^2 - s_{2k-1}^2 \qquad 1 \le k \le S/2. \quad (11)$$

In this case, the length of the resulting feature vector is $S/2$ instead of $S-1$. For the rest of the derivation in this section, as the integrals are identical, we simply consider the case where $k = 1$.

Second, given that the distribution of the complex spectrum is Gaussian, it can be shown by change of variable that the distribution of spectral power is the exponential distribution,

$$f(s^2 \mid \upsilon) = \frac{1}{\upsilon} \exp\left(-\frac{s^2}{\upsilon}\right), \quad (12)$$

where υ is a variance parameter to be substituted later. It follows that the joint distribution of two exponentially distributed observations is

$$f(s_1^2, s_2^2 \mid \upsilon_1, \upsilon_2) = \frac{1}{\upsilon_1 \upsilon_2} \exp\left(-\frac{s_1^2}{\upsilon_1} - \frac{s_2^2}{\upsilon_2}\right). \quad (13)$$

The PDF of the filtered signal arises from changing one of the variables to $z = s_2^2 - s_1^2$ and integrating out the other variable. To perform the integral, notice that in the case where $z \ge 0$, $s_2^2 \ge z$ and $s_1^2 \ge 0$. Also, when $z \le 0$, $s_1^2 \ge -z$ and $s_2^2 \ge 0$. This suggests the use of two different integrals:

In the case where $z \geq 0$,

$$\begin{aligned} f\left(z \mid \upsilon_1, \upsilon_2\right) &= \int_0^\infty ds_1^2 \, f\left(s_1^2\right) f\left(z + s_1^2\right) \\ &= \int_0^\infty ds_1^2 \, \frac{1}{\upsilon_1 \upsilon_2} \exp\left(-\frac{s_1^2}{\upsilon_1} - \frac{z + s_1^2}{\upsilon_2}\right) \\ &= \frac{1}{\upsilon_1 \upsilon_2} \exp\left(-\frac{z}{\upsilon_2}\right) \\ &\quad \times \int_0^\infty ds_1^2 \, \exp\left(-s_1^2 \left[\frac{1}{\upsilon_1} + \frac{1}{\upsilon_2}\right]\right) \\ &= \frac{1}{\upsilon_1 \upsilon_2} \exp\left(-\frac{z}{\upsilon_2}\right) \cdot \frac{\upsilon_1 \upsilon_2}{\upsilon_1 + \upsilon_2}. \end{aligned} \tag{14}$$

Similarly, in the case where $z \leq 0$,

$$\begin{aligned} f\left(z \mid \upsilon_1, \upsilon_2\right) &= \int_0^\infty ds_2^2 \, f\left(s_2^2\right) f\left(s_2^2 - z\right) \\ &= \frac{1}{\upsilon_1 \upsilon_2} \exp\left(\frac{z}{\upsilon_1}\right) \cdot \frac{\upsilon_1 \upsilon_2}{\upsilon_1 + \upsilon_2}. \end{aligned} \tag{15}$$

Substituting back for z, and combining the two results,

$$f\left(s_2^2 - s_1^2 \mid \upsilon_1, \upsilon_2\right) = \begin{cases} \dfrac{1}{\upsilon_1 + \upsilon_2} \exp\left(-\dfrac{s_2^2 - s_1^2}{\upsilon_2}\right) & \text{if } s_2^2 \geq s_1^2, \\ \dfrac{1}{\upsilon_1 + \upsilon_2} \exp\left(-\dfrac{s_1^2 - s_2^2}{\upsilon_1}\right) & \text{if } s_2^2 \leq s_1^2. \end{cases} \tag{16}$$

Note that both expressions are identical when $s_1^2 = s_2^2$.

The likelihood ratio follows easily from equation 16 by substituting for υ_1 and υ_2. Assuming for simplicity that $s_2^2 > s_1^2$, the likelihood ratio is the ratio of the following two equations:

$$f\left(s_2^2 - s_1^2 \mid \mathcal{H} = 1\right) = \frac{1}{\mu_1 + \lambda_1 + \mu_2 + \lambda_2} \exp\left(-\frac{s_2^2 - s_1^2}{\mu_2 + \lambda_2}\right), \tag{17}$$

and,

$$f\left(s_2^2 - s_1^2 \mid \mathcal{H} = 0\right) = \frac{1}{\mu_1 + \mu_2} \exp\left(-\frac{s_2^2 - s_1^2}{\mu_2}\right), \tag{18}$$

which evaluates to

$$\begin{aligned} L\left(s_2^2 - s_1^2\right) &= \frac{\mu_1 + \mu_2}{\mu_1 + \lambda_1 + \mu_2 + \lambda_2} \\ &\quad \times \exp\left(-\frac{s_2^2 - s_1^2}{\mu_2 + \lambda_2} + \frac{s_2^2 - s_1^2}{\mu_2}\right), \\ &= \frac{\mu_1 + \mu_2}{\mu_1 + \lambda_1 + \mu_2 + \lambda_2} \\ &\quad \times \exp\left(\frac{s_2^2 - s_1^2}{\mu_2} \cdot \frac{\lambda_2}{\mu_2 + \lambda_2}\right). \end{aligned} \tag{19}$$

The full likelihood ratio is the product of this expression applied to each pair of spectral bins,

$$L\left(\mathfrak{s}\right) = \prod_{k=1}^{S/2} L\left(s_{2k}^2 - s_{2k-1}^2\right). \tag{20}$$

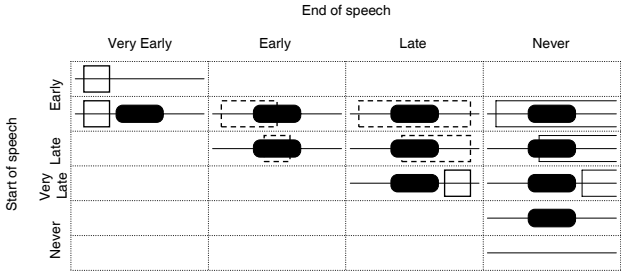

Fig. 1. Classification of VAD start and end times. The dark portion represents speech, the box represents the VAD result.

4. EVALUATION

4.1. Testing data

The VADs were evaluated using an in-house database, some aspects of which were designed specifically for VAD evaluation. The database consists of 14 speakers (7 male and 7 female) each speaking 40 utterances in each of 6 different environments. This is 3360 utterances in total. The utterances are isolated Japanese city names, but are each 5 seconds in length. Typically, the first 2 seconds are background noise, the utterance itself is one second or less, and the final 2 seconds are background noise. The data have been manually marked up with the speech start and end times. The data were recorded on a portable (PDA-like) device using an ear-mounted microphone, the actual microphone being close to the speaker's cheek.

Five of the six environments were chosen to be representative of those where the portable device might be used:

1. The laboratory sound-proof room.
2. A large open-plan office with carpets and fans.
3. A reverberant but open and quiet company lobby.
4. A cafeteria at lunch-time with constant babble noise.
5. A busy suburban street with occasional traffic.
6. A quieter, more open, outdoor area on a windy day.

The average signal to noise ratio for each environment is shown in table 1.

4.2. Evaluation metric

The main evaluation metric consisted of a classification of each utterance into one of the states indicated in figure 1. These are based on a combination of the speech start and end times, and can be thought of as a variation of the classes used by Rosca *et al.* [8]. The four classifications drawn with dashed lines represent the VAD working well, or in such a way that can be corrected using wide margins. The bottom result in the right-most column represents a correct non-detection of an empty utterance, one of which exists (accidentally) in our database. The seven other classifications are certainly errors, being either insertions, deletions or the offset time not being detected in the recording.

The right-most column of figure 1 provides a useful metric for optimizing κ: too small a value leads to large likelihoods and

Table 1. Error rate (%) for four VAD configurations. Also shown is the SNR for each environment, and the optimized value of κ for each VAD configuration.

		Power		Mel	
	SNR (dB)	Gauss	Diff.	Gauss	Diff.
1 (clean)	28.5	0.4	0.4	0.2	0.2
2 (office)	24.7	1.8	1.8	1.3	1.0
3 (lobby)	24.1	0.7	0.5	0.4	0.2
4 (cafe)	16.6	9.8	9.6	4.6	3.8
5 (street)	15.8	6.3	4.1	3.6	3.4
6 (outside)	21.4	6.3	5.2	8.9	5.5
κ		2.5	3.0	1.0	1.0

missing end times. Too large a value, however, leads to deletions. For ASR, we favour insertions over deletions as insertions can be handled using garbage modeling. Missing end times, however, are particularly bad as they cause the recogniser to "hang" and ultimately give an errorful recognition.

4.3. VAD construction

The VAD is inserted into the spectral part of the normal signal processing chain used in ASR. In this case, the signal is sampled at 11.025 KHz and pre-emphasized. Overlapping frames of 256 samples are then taken every 10 ms to form a 128 bin power spectrum (the bin at π is discarded). The power spectrum is transformed into 32 mel spaced bins using half overlapping triangular filters.

The noise variance from the previous frame, $\boldsymbol{\mu}_{f-1}$, is used in the likelihood calculation, and is then updated using a slightly modified version of the estimator described by Sohn *et al.* [2],

$$\hat{\boldsymbol{\mu}}_f = \frac{1-\rho_\mu}{1+L(\mathfrak{s})}\boldsymbol{s}_f^2 + \frac{\rho_\mu + L(\mathfrak{s})}{1+L(\mathfrak{s})}\hat{\boldsymbol{\mu}}_{f-1}, \tag{21}$$

where $\rho_\mu = 0.95$, and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_S)^T$. The speech variances, λ_k, are estimated using power spectral subtraction as suggested in [2], except with the usual over-subtraction and flooring. The VAD was found not to be sensitive to the over-subtraction and flooring values, but note that the flooring means that equation 8 does not reach it's minimum value of 1. For this reason, the cost ratio of equation 5 was set a little above 1 (actually 2.5).

The actual start and end points of the speech were determined using a simple state machine that requires at least 10 frames indicated to be speech in a 1 second window in order to transition to the speech state, and 40 frames of contiguous non-speech to transition into the non-speech state.

4.4. Results

The original Gaussian based VAD, and the differential spectral VAD were tested in both power spectral and mel spectral domains. In each case, the parameter κ was adjusted manually to minimize the number of deletions and missing end times as described in section 4.2. The results are shown in table 1.

The first 3 environments are relatively noise-free; there is no significant difference resulting from the choice of VAD. There is, however, a slight bias in favour of using the mel domain. The latter 3 environments are comparatively noisy, and show more variation. In particular, the mel domain is more robust to the babble noise of the cafeteria. The power spectral domain, however, is more suited to the outdoor wind noise.

Broadly, the differential VAD produces fewer errors than the equivalent non-differential formulation. This confirms the utility of the differential spectral approach. We have also confirmed that this improved VAD performance leads directly to improved speech recognition performance on the same database.

Finally, one obvious difference between the Gaussian and differential VADs is that the former uses a probability for each spectral bin, whereas the latter uses a probability for each pair of bins. In order to confirm that the advantage of the differential VAD is not simply from using half the parameters, we constructed a comparable Gaussian VAD by averaging adjacent bins. This approach only contributed detrimentally to the performance.

5. CONCLUSION

We have placed a spectral VAD into a rigorous decision theoretic framework, and evaluated it in an ASR environment. In order to optimize it to an ASR feature space, and make it robust to noises with smooth spectra, we have re-formulated it as a differential spectral VAD, again in a rigorous statistical manner. We have shown the differential spectral formulation to be superior to the basic Gaussian for an ASR application.

6. REFERENCES

[1] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, February 2002.

[2] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998, pp. 365–368.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, January 1999.

[4] J. Stadermann, V. Stahl, and G. Rose, "Voice activity detection in noisy environments," in *Proceedings of EUROSPEECH*, Scandinavia, 2001.

[5] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[7] C. Nadeu, D. Macho, and J. Hernando, "Time & frequency filtering of filter bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, pp. 93–114, April 2001.

[8] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proceedings of EUSIPCO*, Toulouse, France, September 2002.