VOICE ACTIVITY DETECTION USING HARMONIC FREQUENCY COMPONENTS IN LIKELIHOOD RATIO TEST

Lee Ngee Tan, Bengt J. Borgstrom and Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles
{tleengee, jonas, alwan}@ee.ucla.edu

ABSTRACT

This paper proposes a new statistical model-based likelihood ratio test (LRT) VAD to obtain reliable speech / non-speech decisions. In the proposed method, the likelihood ratio (LR) is calculated differently for voiced frames, as opposed to unvoiced frames: only DFT bins containing harmonic spectral peaks are selected for LR computation. To evaluate the new VAD's effectiveness in improving the noise-robustness of ASR, its decisions are applied to pre-processing techniques such as non-linear spectral subtraction, minimum mean square error short-time spectral amplitude estimator, and frame dropping. From the ASR experiments conducted on the Aurora2 database, the proposed harmonic frequency-based LRTs give better results than conventional LRT-based VADs and the standard G.729B and ETSI AMR VADs.

Index Terms— Voice activity detection, statistical model, harmonic frequency, robust speech recognition

1. INTRODUCTION

The performance of an automatic speech recognition (ASR) system degrades with mismatched training and test data. To improve the noise-robustness of an ASR system trained using clean data, techniques such as speech enhancement [1-4], noise-robust feature extraction feature enhancement [5][6], and model-based noise adaptation [7] can be applied. Most of these techniques require a reliable voice activity detector (VAD) to identify non-speech segments for noise estimation. ASR performance can also improve with a good VAD alone by dropping non-speech segments.

Spectral harmonicity has been utilized in noise-robust applications operating in the frequency domain because harmonic peaks are usually preserved in noisy speech. Non-linear spectral subtraction (NSS) in [3] defined a smaller subtraction factor at harmonic peaks which resulted in higher ASR accuracy at low SNRs. In [4], regeneration of harmonic structures improved the quality of denoised speech. A harmonic model is used in a generalized LRT for robust voiced/unvoiced detection in [8], and a periodic-to-aperiodic component ratio used for speech/non-speech detection in [9] showed promising results with aperiodic noise interferences.

One popular VAD is the statistical model-based LRT VAD first proposed in [10]. Variants of this noise-robust VAD to increase weak speech onset and offset detection have been proposed. For example, [11] used a smoothed LR, while [12] used multiple observations of the short-time DFT feature vector to replace the hangover scheme in [10]. For these VADs, the high LRs of strong speech frames aid the detection of weak neighboring speech frames. However, under low signal-to-noise ratio (SNR) conditions, the LRs of the stronger speech frames are not high enough to boost the detection of weaker speech frames. This paper presents a new way for calculating LR to tackle the above issue in LRT-based VADs and is organized as follows. Section 2 reviews the technical background of LRT-based VADs developed in [10] and [12], while Section 3 describes the proposed method built on these VADs. The proposed VAD is then compared with the referenced VADs [10][12] and standardized VADs: ITU's G.729B [13] and ETSI AMR VAD options 1 and 2 [14], in regards to speech detection accuracy and their influence on ASR performance in Sections 4 and 5, respectively.

2. LRT-BASED VAD

For existing LRT-based VADs [10], the decision rule is formulated by taking the geometric mean of the LRs of all K DFT bins to get the current LR, $\Lambda(t)$ of frame t. This is equivalent to taking the arithmetic mean of their logarithmic versions and comparing it with a threshold η to decide whether speech is present at frame t.

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

where $\Lambda_k(t)$ is the LR of the kth DFT bin. H_0 and H_1 denote speech absence and speech presence, respectively. Since speech spectrum is symmetric, $K = NDFT/2+1$ where NDFT is the number of DFT points taken.

To increase detection of weak speech tails, a hidden Markov model (HMM)-based hangover scheme is

implemented in [10], while a multiple observation likelihood ratio test (MOLRT) VAD is proposed in [12]. The MOLRT decision rule is established by summing consecutive log LRs from 2M+1 frames to make the decision for frame *n*. The use of multiple observations eliminates the need for the hangover scheme and reduces the variance of the LRT, giving rise to a more noise-robust VAD with threshold η'.

$$L(n) = \sum_{t=n-M}^{n+M} \log \Lambda(t) \underset{H_0}{\overset{H_1}{\gtrless}} \eta' \tag{2}$$

3. LRT-BASED VAD USING HARMONIC BINS

When speech is present, it does not always manifest energy in all DFT bins. This is especially true for high-pitched voiced utterance, where most of the energy is located in the harmonic frequency bins. Under low SNR conditions, most of the speech spectrum is masked by noise, thus the LR $\Lambda(t)$ tends to be low after taking the geometric mean of $\Lambda_k(t)$'s of all DFT bins. This increases the probability of missed speech detections. To boost the LR score in such situations, a new method for evaluating the LR for voiced frames is proposed: select only the DFT bins where the harmonic peaks reside for LR computation, since harmonic spectral peaks are stronger and more resilient to noise interferences than other speech spectral components. This LR formulation can be applied to both LRT and MOLRT-based VADs. For the method to work well, the frame size has to be large enough to reveal harmonic structures in the DFT spectrum for low pitch utterances. Since an adult's pitch frequency commonly ranges from 50 to 400 Hz (corresponding to pitch period from 2.5 to 20 ms), a frame size of 50 ms is selected. To identify voiced frames and pitch frequencies, the signal *x(n)* is down-sampled to a frequency F_D of 2 kHz before the normalized autocorrelation function *R(m)* is applied.

$$R(m) = \left[\sum_{n=0}^{N-m-1} d_s(n) d_s(n+m) \right] \Big/ \left[\sum_{n=0}^{N-1} d_s{}^2(n) \right] \tag{3}$$

where $d_s(n)$ is the down-sampled version of *x(n)*, *m* is the autocorrelation lag and *N* is the number of samples in $d_s(n)$. If the magnitude of the maximum peak of *R(m)* exceeds 0.3 (empirically chosen), and its corresponding autocorrelation lag $m_{\max}$ falls within the designated pitch range, the frame is classified as voiced, else it is classified as unvoiced. For every voiced frame, the harmonic bin separation (rounded off to the nearest integer) is calculated as:

$$h_{sep} = round\left(\frac{NDFT}{m_{\max} \times F_s / F_D} \right) \tag{4}$$

where F_s is the sampling frequency of the signal $x(n)$. Vector $\mathbf{H}_{idx}$ stores the DFT bin indices of the harmonic peaks which are obtained using the following iterations

Initialization:

$p = 0$, Next harmonic bin index: $h_{next} = h_{sep}$ (5)

Iterate (6-8) while $h_{next} < K$,

$$\mathbf{H}_{idx}(p) = \underset{h_{next}-1 \le k \le h_{next}+1}{\arg\max} (|X_k(t)|^2) \tag{6}$$

$$h_{next} = \mathbf{H}_{idx}(p) + h_{sep} \tag{7}$$

$$p = p+1 \tag{8}$$

where $X_k(t)$ is the kth DFT coefficient for noisy speech. Finally, the log LR for voiced frame is computed as follows

$$\log \Lambda_v(t) = \frac{1}{p} \sum_{n=0}^{p-1} \log \Lambda_{\mathbf{H}_{idx}(n)}(t) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

As for unvoiced frames, the log LR $\Lambda_{uv}(t)$ is still calculated using (1), with all DFT bins. This is done instead of picking only the stronger energy bins or higher frequency bands because the energy of unvoiced speech tends to be weak and might be dominated by noise interferences. Also, unvoiced consonants such as plosives usually have energy distributed across all frequency bins.

In this paper, the noise variance $\lambda_{N,k}(t)$ of $X_k(t)$ is estimated using a secondary energy-based VAD. The noise variances are updated during the first 10 frames or when the frame energy $E(t)$ falls below the adaptive energy threshold $E_{thres}(t)$. $E_{thres}(t)$ is computed from the mean $\mu(\mathbf{E}_{buf})$ and standard deviation $\sigma(\mathbf{E}_{buf})$ of the values stored in the circular buffer $\mathbf{E}_{buf}$ of dimension 10. The frame energies are stored in the buffer whenever the noise variances are updated.

Noise variances update:

$$\hat{\lambda}_{N,k}(t) = 0.9\hat{\lambda}_{N,k}(t-1) + 0.1|X_k(t)|^2, k = 0,1,...,K-1 \tag{10}$$

Initial noise variances:

$$\hat{\lambda}_{N,k}(0) = |X_k(0)|^2 \quad , \quad k = 0,1,...,K-1 \tag{11}$$

Energy threshold update:

$$E_{thres}(t) = \mu(\mathbf{E}_{buf}) + \sigma(\mathbf{E}_{buf}) \tag{12}$$

4. VAD RESULTS

Reference speech and non-speech segments for the Aurora2 database's Test Set A are obtained through manual labeling the clean version every 10 ms. The receiver operating characteristic (ROC) curves are used to evaluate the accuracy of the proposed VAD. In Section 5, non-speech frames are used for noise estimation in speech enhancement algorithms. Hence, the probability of detection, P_d is defined as the percentage of correctly detected reference non-speech frames, while the probability of false alarm, P_f is the percentage of reference speech frames wrongly identified as non-speech. Fig. 1 shows the ROC curves of LRT, MOLRT and the proposed improved versions of these LRT-based VADs, abbreviated as Hmfreq-LRT and Hmfreq-MOLRT respectively for subway and babble noise-corrupted data. M = 8 is used in MOLRT VADs because it is reported to yield the best performance [12]. The receiver operating points for the standardized VADs are also plotted.

At SNR of 20 dB, it is observed that the P_d of MOLRT-based VADs is higher than single observation LRTs at P_f lower than 20 %, while slightly poorer P_d is observed at P_f higher than 20 %. It is also observed that the ROC curves of the Hmfreq-based VADs are very close to their generic LRT

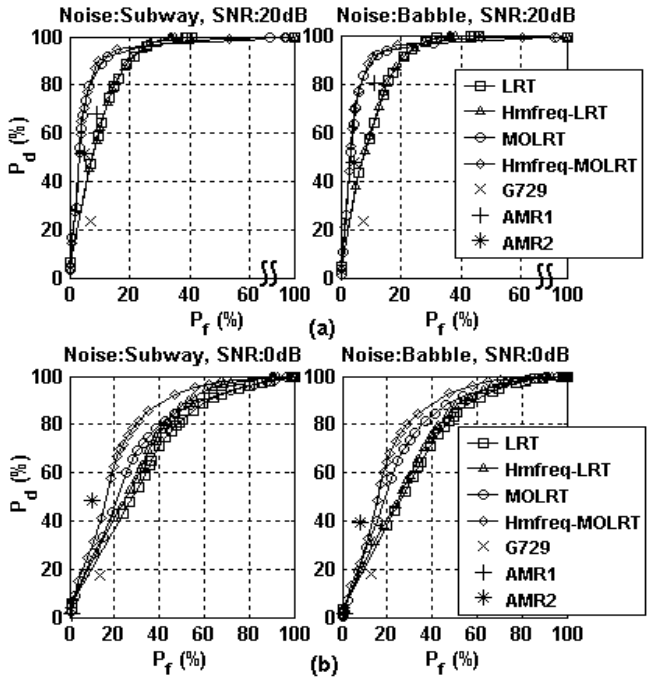

Fig. 1. Comparison of ROC curves for LRT-based and Standardized VADs for Subway and Babble Noise-Corrupted Data in Test Set A at SNR of (a) 20 dB, (b) 0 dB

versions. This result is expected because the LR scores of vowel frames are already very high at SNR of 20 dB. Thus, boosting these scores further through utilizing harmonic frequency bins does not have much effect on the VADs' accuracy. In contrast, at a low SNR of 0 dB, the Hmfreq versions outperform their generic LRT counterparts. There is a larger increase in P_d when Hmfreq-MOLRT is compared to MOLRT than when Hmfreq-LRT is compared to LRT. This is because boosted LR scores for weak vowel frames have a greater influence on the scores of surrounding speech frames in the multiple observations scheme than the single observation hangover scheme. The improvements in P_d are lower for the babble noise-corrupted data as some harmonic spectral structures are present in babble. Hence, there is a higher tendency of misclassifying these harmonic noise interferences as speech. In general, MOLRT-based VADs have better non-speech detection performance than the standardized VADs as well, with the exception of AMR2 at SNR of 0 dB. However, AMR2's receiver operating point is located at the lower P_d and P_f region, which may result in poor noise estimation for non-stationary noise.

5. ASR EXPERIMENTS

The Aurora experimental framework [15] is used to conduct the ASR experiments. A HMM-based recognizer is implemented using HTK version 3.4.1. HMMs are trained using the Aurora2 clean training set and tested on Test Set A. Two well-established speech enhancement algorithms are used to evaluate the effectiveness of the proposed Hmfreq VADs in improving ASR accuracy: Berouti et al's NSS [1] and Ephraim and Malah's minimum mean square error short-time spectral amplitude (STSA) estimator [2]. In NSS, the spectral floor parameter β is set at 0.02 and the subtraction factor α is linearly dependent on SNR as calculated in (4) of [1]. The noise estimation procedure used in the speech enhancement algorithms is similar to that described in Section 3. The noise power spectrum is initialized with the first frame's data (11) and then recursively updated (10) during non-speech frames determined by the VAD. The frame size and shift used for speech enhancement are 25 and 10 ms respectively, same as those used for extracting the 39-element MFCC_D_A_E feature vector [15] from the post-enhanced signal. Besides speech enhancement, frame dropping (FD) is another pre-processing technique selected to evaluate the VAD's performance. In this paper, FD is implemented by replacing the time samples of non-speech frames with those from a typical silence waveform. When combined with speech enhancement, FD is performed on the post-enhanced speech signal. For simplicity, a constant decision rule threshold is applied in LRT-based VADs. The threshold values for LRT, Hmfreq-LRT, MOLRT and Hmfreq-MOLRT are set to 2, 4, 6 and 20, respectively. These values give the best overall ASR results for the pre-processing schemes investigated.

ASR word accuracies, averaged across all the 4 noise types in Test Set A, for the VADs and pre-processing techniques investigated, are presented in Tables 1A – 1C. Tables 1A, 1B and 1C contain the averaged results for 20, 0 and 0-20 dB SNRs respectively. From Table 1C, the overall ASR results obtained with Hmfreq-based VADs outperform the standardized and referenced LRT-based VADs for all the pre-processing techniques investigated. Without FD, the highest averaged word accuracy is achieved by the Hmfreq-LRT-based VAD. On the other hand, Hmfreq-MOLRT-based VAD gives the best ASR results when FD is involved. The overall recognition accuracy of 74.21 % obtained with Hmfreq-MOLRT-based VAD is closest to the 75.35 % obtained using manually labeled VAD decisions.

A possible explanation for higher word accuracies achieved when NSS or STSA is operated with Hmfreq-LRT-based VAD than with Hmfreq-MOLRT-based VAD or manually labeled decisions is single observation LRT-based VADs' decisions result in better noise reductions at in-between-words regions. This improves the word recognition performance in cases where the words are not bounded by consonants. Even in cases where consonants are missed by Hmfreq-LRT-based VAD, the vowel itself is sometimes sufficient for correct word recognition, since this is a pure digit ASR task. Analysis also reveals that the Hmfreq-MOLRT-based VAD and manual-labeled decisions lead to more substitution errors than the Hmfreq-LRT-based VAD when NSS or STSA is performed. This could be caused by poorly suppressed noise that alters the spectral characteristics of weak speech regions detected by

Table 1A. Averaged Word Accuracy (%) for Test Set A at 20 dB. NSS: Non-linear Spectral Subtraction, STSA: Short-Time Spectral Amplitude Estimator, FD: Frame-Dropping

VADs	NSS	STSA	FD	NSS+FD	STSA+FD	Overall
G729	95.77	96.00	92.98	94.40	94.77	94.78
AMR1	95.72	95.63	92.76	93.80	93.58	94.30
AMR2	95.22	95.77	94.39	94.67	95.41	95.09
LRT	95.87	**96.12**	93.21	94.49	95.02	94.94
Hmfreq-LRT	**96.05**	96.08	93.25	94.38	94.82	94.92
MOLRT	95.72	95.86	96.24	95.75	**96.01**	95.92
Hmfreq-MOLRT	95.98	95.95	**96.59**	**96.02**	95.98	**96.10**
Manual labeling	96.28	96.26	97.45	96.55	96.58	96.62

Table 1B. Averaged Word Accuracy (%) for Test Set A at 0 dB

VADs	NSS	STSA	FD	NSS+FD	STSA+FD	Overall
G729	27.98	26.11	16.71	25.82	25.13	24.35
AMR1	25.59	23.94	16.72	25.74	23.92	23.18
AMR2	36.28	37.20	18.59	32.79	34.78	31.93
LRT	38.13	36.97	17.33	26.08	31.25	29.95
Hmfreq-LRT	**41.37**	39.62	18.76	27.27	31.96	31.77
MOLRT	38.29	38.45	20.56	33.66	37.19	33.63
Hmfreq-MOLRT	40.76	**39.95**	**23.01**	**36.51**	**38.81**	**35.81**
Manual labeling	40.18	40.00	21.01	43.10	44.24	37.71

Table 1C. Averaged Word Accuracy (%) for Test Set A from 0 to 20 dB

VADs	NSS	STSA	FD	NSS+FD	STSA+FD	Overall
G729	70.66	70.13	59.95	67.67	68.09	67.30
AMR1	68.60	68.24	60.30	67.52	67.24	66.38
AMR2	73.22	74.08	61.96	71.18	72.50	71.19
LRT	75.88	75.24	59.56	67.92	70.39	69.80
Hmfreq-LRT	**77.41**	**76.26**	60.22	68.29	70.46	70.53
MOLRT	75.39	75.34	65.77	73.55	75.11	73.03
Hmfreq-MOLRT	76.76	76.00	**67.64**	**74.94**	**75.72**	**74.21**
Manual labeling	76.79	75.97	67.79	78.17	78.04	75.35

Hmfreq-MOLRT-based VAD and manual-labeled decisions, resulting in more misclassifications.

When FD is combined with speech enhancement, there is a much sharper decrease in the average word accuracies for systems with single observation LRT-based VADs than those with MOLRT-based VADs. The contributing factor is the increasing number of deletion errors with decreasing SNRs for these VADs. This is evident from the larger decrease in ASR accuracies before and after FD is applied in Table 1B when compared to Table 1A. It is also observable from Tables 1A and 1B that the proposed Hmfreq scheme, when compared to the LRT and MOLRT-based VADs it is built upon, shows larger improvements in ASR accuracies at 0 dB than 20 dB. This corresponds to the VADs' performances shown via the ROC curves in Section 4.

6. CONCLUSION

This paper presents a novel way to improve the noise-robustness of existing LRT-based VADs by selecting the harmonic DFT components for computing the LR scores of voiced frames. ROC curves show an increase in correct VAD decisions for this implementation. Higher ASR accuracies are obtained when Hmfreq-based VADs' decisions are applied to well-known pre-processing techniques compared to conventional LRT-based VADs as well standardized VADs, especially in low SNR conditions.

7. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *in Proc. Int. Conf. Acoustic, Speech and Signal Processing (ICASSP)*, pp. 208–211, Apr. 1979.

[2] Y. Ephraim, and D. Malah, "Speech Enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. ASSP-32, pp. 1109–1125, Dec. 1984.

[3] J. Beh, and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech", *ICASSP Proc.*, pp. 684–687, Apr. 2003.

[4] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration", *ICASSP Proc*., pp. 157–160, 2005.

[5] H. Hermansky, and N. Morgon, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing (TSAP)*, vol. 2, no. 4, pp. 578-589, Oct. 1994.

[6] Q. Zhu, and A. Alwan, "Non-linear feature extraction for robust recognition in stationary and non-stationary noise", *Computer, Speech and Language*, vol. 17, pp. 381-402, 2003.

[7] T. Kristjansson, B. Frey, L. Deng and A. Acero, "Towards non-stationary model-based noise adaptation for large vocabulary speech recognition", *ICASSP Proc*., pp. 337–340, May. 2001.

[8] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model", *IEEE TSAP*, vol. 14, pp. 502–510, Dec. 1984.

[9] K. Ishizuka, and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio", *in Proc. ISCA Tutorial and Research Workshop on Statistical and Perception Audition*, pp. 65-70, Sep. 2006.

[10] 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, Jan. 1999.

[11] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," *ICASSP Proc.*, vol. 2, pp. 737–740, May 2001.

[12] J. Ramírez, J. C. Segura, et al, "Statistical voice activity detection using a multiple observation likelihood ratio test", *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689-692, Oct. 2005.

[13] ITU-T Rec. G.729, Annex B, pp. 44–59, 2007.

[14] ETSI TS 126 094 V7.0.0, Chapter 3–4, pp 5–24, 2007.

[15] H. G. Hirsch, and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *in Proc. ISCA ITRW ASR2000*, vol. ASSP-32, pp. 181–188, Sep. 2000.