Voice activity detection based on density ratio estimation and system combination

Yuuki Tachioka*, Toshiyuki Hanazawa*, Tomohiro Narita*, and Jun Ishii*
*Information Technology R & D Center, Mitsubishi Electric Corporation, Kanagawa, Japan
E-mail: Tachioka.Yuki@eb.MitsubishiElectric.co.jp

***Abstract*—We propose a robust voice activity detection (VAD) based on density ratio estimation. In highly noisy environments, the likelihood ratio test (LRT) is effective. Conventional LRT estimates both speech and noise models, calculates the likelihood of each model, and uses ratios of such likelihood to detect speech. However, in LRT, the likelihood ratio of speech and noise models is required, whereas likelihood of individual models is not necessarily required. The framework of the density ratio estimation models likelihood ratio functions by a kernel and directly generates a likelihood ratio. Applying density ratio estimation to VAD requires that feature selection and noise adaptation must be considered. This is because the density ratio estimation constrains the shape of the likelihood ratio functions and speech is dynamic. This paper addresses these problems. To improve accuracy, the proposed method is combined with conventional LRT. Experimental results using CENSREC-1-C show that the proposed method is more effective than conventional methods, especially in non-stationary noisy environments.**

I. Introduction

Voice activity detection (VAD) is an essential pre-process in speech processing. Determining the speech area effectively reduces error recognition and the adjustment of the strength of noise suppression in noisy environments. The most basic VAD, which assumes that the power of speech is usually greater than that of noise [1], is ineffective in highly noisy environments where speech is masked by noise. The use of the characteristics of speech, e.g., the periodic structure of speech [2], is susceptible to noise [3]. The use of decoder output is effective but computational costs are high [4].

A simple and effective model-based method called the likelihood ratio test (LRT) is effective in highly noisy environments. Even if the power of speech is lower than that of noise, the likelihood of the speech model is greater than that of noise model because the characteristics of speech are available. Sohn *et al.* proposed using the likelihood ratio of speech and noise models after estimating both models from observation to detect speech [5].

Among methods [3], [6], [7] that improve Sohn's method, Fujimoto *et al.* proposed constructing speech models by synthesizing *a priori* clean speech and observed noise at each frame and constructs a noise model by using observed noise to calculate the likelihood ratio of these models [3]. This outperforms Sohn's method, especially in noisy environments, mainly by on-line estimation of models. However, Sohn's method remains an important benchmark of LRT-based VAD and, currently, many comparisons have been made with Sohn's method.

The common point of the above methods is calculating the likelihoods of speech and noise models, respectively, and using the ratio of likelihood to determine whether individual frames are speech or noise. The noise model is estimated from observation and the speech model is estimated by maximum likelihood [5] or by clean speech in advance [3]. In LRT, however, if the likelihood ratio of speech and noise model is estimated directly, the likelihood of individual models is not required.

In the field of machine learning, Sugiyama *et al.* have recently proposed estimating the probability density ratio of two probability distributions directly without estimating their probability densities [8]. This directly models the density ratio function by using a kernel and estimating its parameters from training data, and calculates the likelihood ratio directly, which is effective in change-point detection tasks [9].

There are two problems in applying density ratio estimation to VAD: feature selection and noise adaptation. This is because density ratio estimation puts constraints on feasible features due to the shape of the kernel and speech is dynamic. This paper addresses these problems and proposes a method that directly estimates the likelihood ratio for VAD. To use the advantages of conventional LRT and the proposed method, the systems of different features and models are combined. Conventional LRT is introduced in Section II and density ratio estimation in Section III. The proposed method is described in Section IV.

II. Conventional LRT (Sohn's method)

One of the simplest and most effective conventional LRT methods [5] is described here. The K_X-dimensional short-time Fourier transform coefficients of observation is $\boldsymbol{X} = \{X_k\}_{k=1}^{K_X}$. The likelihood of the power spectrum $|X_k|^2$ conditioned on the speech model λ^S and the noise model λ^N are assumed to be represented as independent Gaussian distributions, as in Eq. (1):

$$\begin{aligned} p(\boldsymbol{X}|\lambda^S) &= \prod_{k=1}^{K_X} \frac{1}{\pi[v_k^S + v_k^N]} e^{-\frac{|X_k|^2}{v_k^S + v_k^N}}, \\ p(\boldsymbol{X}|\lambda^N) &= \prod_{k=1}^{K_X} \frac{1}{\pi v_k^N} e^{-\frac{|X_k|^2}{v_k^N}}, \end{aligned} \tag{1}$$

where v_k^S and v_k^N are the variance of speech and noise spectrum, respectively. The log likelihood ratio of speech and

noise at the k^{th} dimension is represented as

$$\Lambda_k(X_k|\lambda^S,\lambda^N)=\ln\frac{p(X_k|\lambda^S)}{p(X_k|\lambda^N)}. \tag{2}$$

The geometric mean of the likelihood ratio is used to determine whether individual frames are speech or noise as

$$\Lambda(\boldsymbol{X}|\lambda^S,\lambda^N)=\frac{1}{K_X}\sum_{k=1}^{K_X}\Lambda_k(X_k|\lambda^S,\lambda^N)\underset{H_N}{\overset{H_S}{\gtrless}}\eta, \tag{3}$$

where if $\Lambda(\boldsymbol{X}|\lambda^S,\lambda^N)$ is greater than the threshold η, this frame is considered to be H_S ((noisy) speech state), and otherwise H_N (noise state). The noise model is estimated in advance by observed noise, and the speech model is estimated by maximum likelihood estimation, i.e., $\partial\Lambda_k(X_k)/\partial\lambda_k^S=0$, which results in relationship $v_k^S=|X_k|^2-v_k^N$. This shows that speech model λ_k^S is estimated assuming that speech and noise power are additive.

III. Density ratio estimation (KLIEP)

Probability density ratio q for sequential data y is defined as

$$q(y|\lambda^n,\lambda^d)=\frac{p(y|\lambda^n)}{p(y|\lambda^d)}, \tag{4}$$

where p is the probability density function of y conditioned on numerator model λ^n and denominator model λ^d, respectively. Here, we assume that training data are labeled as $\boldsymbol{y^n}=\{y^n(i)\}_{i=1}^I$ and $\boldsymbol{y^d}=\{y^d(j)\}_{j=1}^J$ for models λ^n and λ^d, respectively. It is known that simple kernel density estimation, which estimates the density ratio function using statistics of $\boldsymbol{y^n}$ and $\boldsymbol{y^d}$ separately[1], results in low estimation accuracy [9].

The Kullback-Leibler Importance Estimation Procedure (KLIEP) [8], in contrast, directly models density ratio model λ^r instead of λ^n and λ^d. This improves the robustness of density ratio calculation. The density ratio is modeled as linear model $\hat{q}(y)$ which consists of M mixture kernels φ_m, as in Eq. (5):

$$\hat{q}(y|\lambda^r)=\frac{\hat{p}(y|\lambda^r,\lambda^d)}{p(y|\lambda^d)}=\sum_{m=1}^M\alpha_m\varphi_m(y)=\sum_{m=1}^M\alpha_m e^{-\frac{|y-\mu_m^r|^2}{2v^r}}, \tag{5}$$

where α_m is a non-negative mixture weight and φ_m is a Gaussian kernel whose parameters are μ_m^r and v^r, which are the center and width of a kernel, respectively. A Gaussian kernel requires that the density ratio function takes larger values at the point where many samples from $\boldsymbol{y^n}$ converge, but otherwise takes smaller values close to zero.

Here, μ_m^r, v^r, and α_m are unknown variables that are estimated in the following four steps:

1) Some kernel widths v^r are set arbitrarily.
2) M samples from $\boldsymbol{y^n}$ are picked as $\{\mu_m^r\}_{m=1}^M$.
3) Mixture weight α_m is obtained by solving the optimization problem shown below.
4) The appropriate value of v^r is determined by n-fold cross validation.

In KLIEP, α_m is determined as the KL divergence of a sample y from $p(y|\lambda^n)$ to $\hat{p}(y|\lambda^r,\lambda^d)$ is minimized, where $\hat{p}(y|\lambda^r,\lambda^d)$ is the numerator estimated density represented by $\hat{q}(y|\lambda^r)p(y|\lambda^d)$. KL divergence L is represented as

$$\begin{aligned}L\left(p(y|\lambda^n);\hat{p}(y|\lambda^r,\lambda^d)\right)&=\int_{\mathcal{D}}p(y|\lambda^n)\ln\frac{p(y|\lambda^n)}{p(y|\lambda^d)}dy\\&\quad-\int_{\mathcal{D}}p(y|\lambda^n)\ln\hat{q}(y|\lambda^r)dy,\end{aligned} \tag{6}$$

where $\mathcal{D}$ is a data domain. Since $\hat{p}(y|\lambda^r,\lambda^d)$ is a probability density function, constraint must be satisfied as

$$\int_{\mathcal{D}}\hat{p}(y^n|\lambda^r,\lambda^d)dy^n=\int_{\mathcal{D}}\hat{q}(y^d|\lambda^r)p(y^d|\lambda^d)dy^d=1. \tag{7}$$

To minimize KL divergence, the second term of Eq. (6) is minimized under the constraint in Eq. (7) because the first term on the right side of Eq. (6) is constant for α_m. The optimization problem in Eq. (8) is obtained by substituting a sample mean for an expectation of the second terms of Eq. (6) and Eq. (7). Solving the optimization problem requires only labeled features $\boldsymbol{y^n}$ and $\boldsymbol{y^d}$ and thus does not require information on λ^n and λ^d. This problem is a convex optimization problem because α_m is non-negative and reaches global optimization by gradient descent and constraint satisfaction. Optimized solutions tend to be sparse, that is, some α_m values are zero. This property is effective in reducing computational costs.

$$\begin{aligned}&\underset{\{\alpha_m\}_{m=1}^M}{\arg\min}\left[-\sum_{i=1}^I\ln\left(\sum_{m=1}^M\alpha_m\varphi_m(y^n(i))\right)\right],\\&\text{subject to }\sum_{m=1}^M\alpha_m\left[\frac{1}{J}\sum_{j=1}^J\varphi_m(y^d(j))\right]=1.\end{aligned} \tag{8}$$

IV. Application of density ratio estimation (KLIEP) for VAD

The density ratio estimation is applied to the VAD problem by substituting variables into Eq. (5) as

$$y\leftarrow Y_k,\hat{q}\leftarrow\hat{\Lambda}_k, \tag{9}$$

where Y_k is a component of the K_Y-dimensional feature vector $\boldsymbol{Y}=\{Y_k\}_{k=1}^{K_Y}$ and $\hat{\Lambda}_k$ is a likelihood ratio conditioned on density ratio model λ^r obtained by the above procedure[2]. Speech is detected as

$$\hat{\Lambda}(\boldsymbol{Y}|\lambda^r)=\frac{1}{K_Y}\sum_{k=1}^{K_Y}\hat{\Lambda}_k(Y_k|\lambda^r)\underset{H_N}{\overset{H_S}{\gtrless}}\eta. \tag{10}$$

There are two problems in applying the above KLIEP to VAD: the feature selection and the noise adaptation. This is because density ratio estimation puts constraints on feasible features due to the shape of the kernel and speech is dynamic. First, we consider feature selection. Features are assumed to

[1]For example, after assuming two Gaussian kernels and estimating these parameters from each sample $\boldsymbol{y^n}$ and $\boldsymbol{y^d}$, the ratio of these density functions is calculated [10].

[2]We refer to a Matlab® code [11] when implementing model learning.

be independent at each dimension. Features are certainly correlated across feature dimensions, but use of a full covariance matrix requires extremely large computational costs. Thus, the density ratio function is estimated at each dimension. Training data need to be labeled as speech and noise. The estimation performance of KLIEP is high when the variance of the denominator distribution v^d is greater than that of the numerator distribution v^n, because the value of the density ratio function is unstable when the denominator value is small and the numerator value is large. If, for example, denominator and numerator distributions are represented as Gaussians kernels ($\exp(-|y-\mu^d|^2/2v^d)$ and $\exp(-|y-\mu^n|^2/2v^n)$), the density ratio function is represented as $\exp(-|y-\mu^r|^2/2v^r)$ where

$$\mu^r = \frac{v^d\mu^n - v^n\mu^d}{v^d - v^n},\ v^r = \frac{v^n v^d}{v^d - v^n}.$$

In this case, estimation is only stable when v^d is greater than v^n. Otherwise, v^r is negative. Power often satisfies this requirement because the dynamics of noise is greater than that of speech in the long term whereas the MFCC feature, which is normally used for ASR, does not necessarily satisfy this requirement. In fact, in the case of MFCC, the estimation accuracy of the density ratio function is low as shown in Section V-B. We propose to use a log power spectrum as $\boldsymbol{Y} = \ln|\boldsymbol{X}|^2$ for the feature because the range of a 'raw' power spectrum is too large to be represented by a linear model.

Second, we consider the adaptation of a model. There is a mismatch between training and evaluation environments due to noise diversity. Adaptation of a model is effective because speech and noise are dynamic [12]. For both Sohn's method and the proposed method, it is necessary to adjust the mean of features because these methods assume a relative power difference between speech and noise. It is clearly shown that, even for the same speech, the boundary of speech and noise shifts if microphone gain changes. Sohn's method avoids this mean shift effect by using variance as a model. The proposed method equates the mean and variance of noise during first N_N frames with those of training noise to adapt noise. The on-line adaptation of a model, e.g., [3], [13], is a future work.

As [3] mentioned, combining different features and models is effective. Here, the proposed method is combined with Sohn's method to exploit the advantages of both. Two likelihood ratios are combined as

$$\Lambda'' = \gamma\Lambda'(\boldsymbol{X}|\lambda^S, \lambda^N) + (1-\gamma)\hat{\Lambda}'(\boldsymbol{Y}|\lambda^r), \qquad (11)$$

where Λ' and $\hat{\Lambda}'$ are likelihood ratios normalized by the maximum value of Λ and $\hat{\Lambda}$ during utterance and γ is the constant weight of the two systems, which weighs the importance on either system ($\gamma = 0$: Sohn's method and $\gamma = 1$: proposed method). VAD is performed using the obtained likelihood ratio Λ''

V. Experiments

A. Experimental setup

The proposed method was evaluated using the CENSREC-1-C database [14], [15], which is commonly used for evaluating VAD in noisy environments. Evaluation data were recorded in two real environments: 'RESTAURANT' (speech and foot noise: non-stationary) and 'STREET' (traffic noise: stationary), with two different SNRs ('HIGH' and 'LOW'). Each file consisted of 8-10 utterances, which were 1-12 digit numbers. The sampling frequency was 8 kHz and the dimension, the window length and the frame shift of short-time Fourier transform were 256, 25 ms and 10 ms, respectively. Feature dimensions K_X and K_Y were 129, considering symmetry. Performance was evaluated in terms of the correct and accuracy rate [%].

We compared results for the proposed method to those of two conventional methods: a power-based method attached to CENSREC-1-C as a baseline (similar to [1]) and Sohn's method in Section II. Some methods that use on-line adaptation for noise certainly outperform Sohn's method because, for this database, noise adaptation is effective due to the long files which contain multiple utterances with changing noise. However, Sohn's method is still an LRT based benchmark among methods without on-line adaptation and the proposed method does not use on-line adaptation.

The first 10 ($= N_N$) frames were used to construct a noise model for Sohn's method and to adapt the mean and the variance of a background noise for the proposed method. Thresholds η were optimized among some candidates. The density ratio model was trained using CENSREC-4 database [15], including eight types of reverberation and noise with SNRs {5, 10, 20, 25, 30} [dB], which were totally different from CENSREC-1-C. They were down-sampled to 8 kHz from 16 kHz. The number of training data was 16000 (160 seconds) for speech and noise data, respectively. The number of kernels was 20 ($= M$). The width v^r was determined by 5-fold cross-validation. The weight γ for system combination was 0.3 (turned on the preliminary experiments).

B. Results and discussions

Fig. 1 (a) shows the distributions of the 15^{th} dimension (which approximately equals to 500 Hz and includes rich information of speech) of the log power of speech and noise and a density ratio function obtained by KLIEP, where there are 13 non-zero α_m. This shows that KLIEP estimated a density ratio function for VAD. On the other hand, Fig. 1 (b) shows the distributions and a density ratio function of the 1^{st} dimension of MFCC. Here, because v^d is apparently much smaller than v^n, the density ratio function does not satisfy KLIEP requirements. The estimated function shape is flat and cannot discriminate between speech and noise.

Table I shows that the proposed method improves the average correct rate by 28.6% from the CENSREC baseline and 6.0% from Sohn's method, and improves the average accuracy rate by 74.8% and 8.5%, respectively. The proposed method

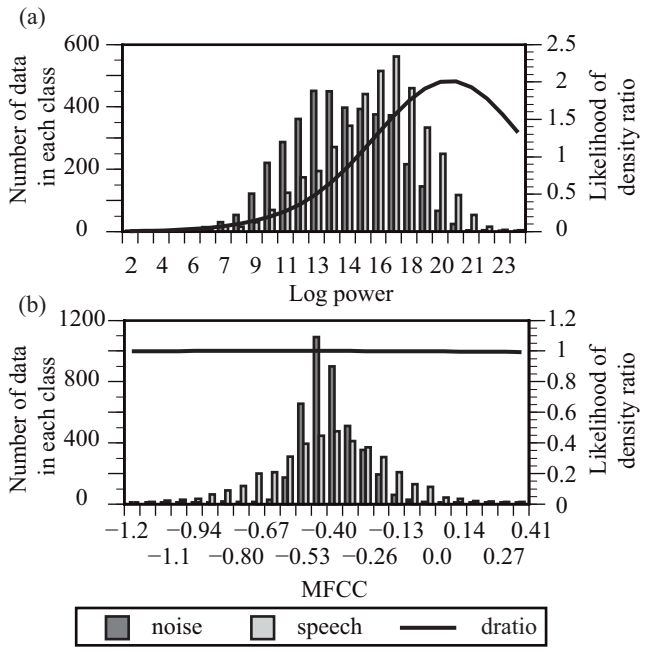

Fig. 1. Histogram of (a) the log power (15^{th} dimension) and (b) MFCC (1^{st} dimension) of speech and noise and probability density ratio (dratio).

TABLE I
CORRECT AND ACCURACY RATES[%] OF THE PROPOSED METHOD (PROP) AND SYSTEM COMBINATION (COMB) COMPARED TO THOSE OF THE CENSREC-1-C BASELINE (BASE) AND SOHN'S METHOD (SOHN) IN TERMS OF ENVIRONMENTS (RESTAURANT AND STREET) AND SNR (HIGH (H) AND LOW (L)).

		Correct				Accuracy			
		base	Sohn	prop	comb	base	Sohn	prop	comb
RESTRANT	H	74.2	73.0	**89.0**	81.2	21.5	41.5	**67.0**	57.1
	L	56.5	59.4	**63.5**	57.4	-43.5	13.9	15.9	**24.6**
STREET	H	39.4	94.2	91.0	**95.7**	-15.7	86.1	82.6	**92.5**
	L	41.5	75.4	82.6	**86.1**	-33.9	52.2	62.0	**74.8**
Average		52.9	75.5	**81.5**	80.1	-17.9	48.4	56.9	**62.3**

outperforms the conventional methods for 'RESTAURANT', which is non-stationary noise. This shows that the density ratio model is more robust in mis-estimating the model than Sohn's model. The system combination, moreover, improves the accuracy rate by 13.9% from Sohn's method. Sohn's method is effective in stationary noise, therefore the system combination exploits the advantages of both Sohn's method and the proposed method.

Fig. 2 (a) and (b) show the likelihood ratio calculated by Sohn's method and the proposed method, respectively, under the condition of RESTAURANT(HIGH). The likelihood ratio of the proposed method remains stable at low during the non-speech area. Because noise is non-stationary, Sohn's noise model obtained by using the first 10 frames mismatches actual noise and generates high likelihood ratios that lead to a false detection. The proposed method is more robust than Sohn's method for mis-estimation by using a density ratio model.

VI. CONCLUSIONS AND FUTURE WORK

We proposed a voice activity detection method based on the density ratio estimation. Experiments show that the proposed method is more effective than conventional methods, espe-

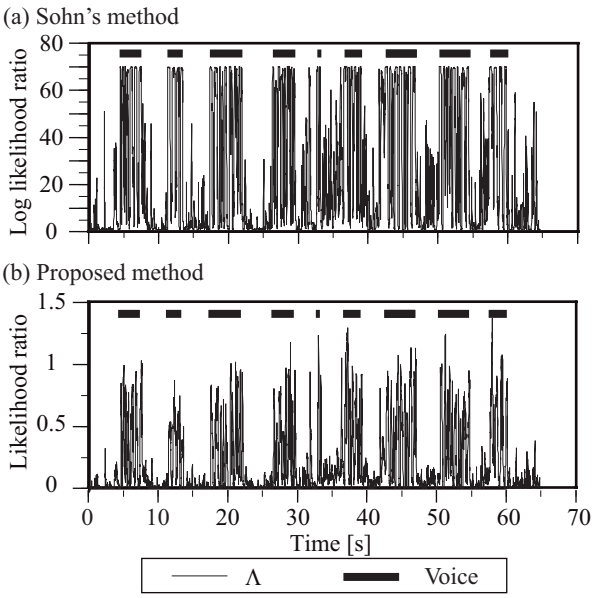

Fig. 2. Likelihood ratio of (a) Sohn's method and (b) the proposed method (RESTAURANT(HIGH)).

cially under non-stationary noisy environments. Future work lies in finding new features that are effective in density ratio estimation and on-line adaptation of a model.

REFERENCES

[1] L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell Sys. Tech. J.*, vol. 54, pp. 297–315, Feb 1975.
[2] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in *Proceedings SAPA*, pp. 65–70, Sep 2006.
[3] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in *Proceedings ICASSP*, pp. 4441–4444, 2008.
[4] T. Ohnish, P. Dixon, K. Iwano, and S. Furui, "Robust speech recognition using VAD-measure embedded decoder," in *Proceedings INTERSPEECH*, pp. 2239–2242, Sep 2009.
[5] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Proc. Lett.*, vol. 6, pp. 1–3, Jan 1999.
[6] J. Chang, N. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. on Sig. Proc.*, vol. 54, pp. 1965–1976, June 2006.
[7] O. Pernía, J.M. Górriz, J. Ramírez, C. Puntonet, and I. Turias, "An efficient VAD based on a generalized Gaussian PDF," in *Proceedings Int. Conf. on Adv. in Nonlinear Speech Proc.*, pp. 246–254, 2007.
[8] M. Sugiyama, T. Kanamori, T. Suzuki, S. Hido, J. Sese, I. Takeuchi, and L. Wang, "A density-ratio framework for statistical data processing," *IPSJ Trans. on Comp. Vision and Appl.*, vol. 1, pp. 183–208, Sep 2009.
[9] Y. Kawahara and M. Sugiyama, "Change-point detection in time-series data by direct density-ratio estimation," in *Proceedings SIAM Int. Conf. on Data Mining*, pp. 389–400, 2009.
[10] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and semiparametric models*, Springer Series in Statistics, 2004.
[11] http://sugiyama-www.cs.titech.ac.jp/˜sugi/software/kliep/, 2013.
[12] S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath, "Dynamic noise adaptation," in *Proceedings ICASSP*, pp. 1197–1200, 2006.
[13] R. Weiss and T. Kristjansson, "DySANA: Dynamic speech and noise adaptation for voice activity detection," in *Proceedings INTERSPEECH*, pp. 127–130, 2008.
[14] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoust Sci. & Tech.*, vol. 30, pp. 363–371, Sep 2009.
[15] http://research.nii.ac.jp/src/eng/list/, 2013.