電気学会論文誌 C（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.133 No.8 pp.1549–1555　DOI: 10.1541/ieejeiss.133.1549

論　文

# 音声と騒音の密度比推定を用いた音声区間検出法

正　員　太刀岡勇気*a)　非会員　花沢　利行*
非会員　成田　知宏*　非会員　石井　　純*

# Voice Activity Detection Using Density Ratio Estimation of Speech and Noise

Yuuki Tachioka*a), Member, Toshiyuki Hanazawa*, Non-member, Tomohiro Narita*, Non-member, Jun Ishii*, Non-member

（2012年7月4日受付，2013年2月20日再受付）

In this paper, we propose a robust voice activity detection (VAD) method that uses a density ratio model. For VAD under highly noisy environments, the likelihood ratio test (LRT) is effective. Conventional LRT constructs speech and noise models, calculates the likelihood of each model, and takes the ratio of those likelihoods to detect speech. Although some improved LRT have been proposed, in conventional LRT, it has not been taken into account that the likelihood ratio of speech and noise model is required, not the likelihood of each model. The proposed method directly estimates the likelihood ratio without calculating each likelihood using an density ratio model obtained in advance by density ratio estimation procedure. Moreover, there is the problem of determining thresholds, which are used for VAD and significantly affect its performance. We propose a method that automatically determines thresholds using discriminant analysis. The experiments show that the proposed method is more effective than conventional methods especially under non-stationary noisy environments.

キーワード：音声区間検出，尤度比検定，密度比推定，耐騒音性，判別分析による閾値の決定
**Keywords:** Voice activity detection, Likelihood ratio test, Density ratio estimation, Noise robustness, Threshold determination by discriminant analysis

## 1. はじめに

音声区間検出は，観測された音の中から音声の区間を特定する技術であり，音声認識・強調において須要の前処理である[(1)(2)]。最も基本的なものは，有声部では音声のパワーが騒音のそれよりも大きく，無声子音部ではゼロ交差数が多くなることを活用するものである[(3)]。この方法は必要な計算量が非常に小さいため，現在でもよく用いられている。比較的静かな環境では十分な性能を発揮するが，音声が騒音に埋もれる高騒音環境下では，誤受理もしくは誤棄却が多くなるため有効でない。(誤受理，誤棄却のいずれが多くなるかは閾値の設定に依存する。)

そのため，観測音から音声と騒音のモデル化を行い，それらの尤度比により音声区間検出を行う手法 (Likelihood ratio test, 以下 LRT と呼ぶ) が提案されている。LRT は，特に高騒音下で有効であることがよく知られている[(4)]。また，この手法を発展させたものとして，観測特徴量ベクトル列の LRT を用いた手法[(5)]や，複数の確率モデルを用いる手法[(6)]，事前に学習したクリーン音声と観測された騒音をオンライン合成した騒音が重畳した音声のモデルと観測された騒音のモデルの尤度比による手法も提案されている[(7)]。上記手法に共通しているのは，音声・騒音それぞれのモデルの各フレームにおける尤度を算出し，それらの比をとることで音声・非音声を判定している点である。

ところが LRT において最終的に必要なのは音声モデルと騒音モデルの尤度比であり，途中経由しているそれぞれの尤度は必要ない。上述の従来法では，このことが考慮されていない。一般にある 2 つの確率変数が与えられたときに，それぞれのモデルの尤度を求めてから尤度比を算出するよりも，直接尤度比を求める方が，求める未知数が少ないため，易しい問題であるといえる。近年機械学習の分野では，この点に着目して，2 つの確率分布それぞれの確率密度を推定するのではなく，それらの確率密度比を直接推定することで推定のロバスト性を向上させた手法が提案されている[(8)]。本報では，本手法を上記音声区間検出における尤度

a) Correspondence to: Tachioka.Yuki@eb.MitsubishiElectric.co.jp
* 三菱電機 (株) 情報技術総合研究所
〒 247-8501　神奈川県鎌倉市大船 5-1-1
Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1, Ofuna, Kamakura, Kanagawa 247-8501, Japan

比の算出問題に応用する。〈2·1〉では，従来法として LRT の中でも代表的な Sohn の方法に関して概要を述べる。その後，〈2·2〉で従来法に触れ，〈2·3〉で確率密度比推定を音声区間検出に適用する方法を提案する。

ところで音声区間検出においては，音声・非音声を判断する際に設定したある閾値と比較するが，閾値の決定には試行錯誤が必要である。音声区間検出の性能は閾値により左右され，適切に閾値を定めることは難しい。これは LRT だけでなく音声区間検出法一般に通ずる課題である。一般には事前の検討で定めた閾値を用いることが多いが，本報では，〈2·4〉においてクラスタリング分析を用いた閾値の自動決定法を提案する。

## 2. 音声区間検出法

**〈2·1〉 LRT (Sohn の方法)** LRT の代表的な手法である Sohn の方法[4]に関して述べる。短時間フーリエ変換 (STFT) により，観測音の FFT 係数の $L$ 次元ベクトル $\mathbf{X}$ を求める。非音声区間 $H_n$ と音声区間 $H_s$ での音声と騒音のベクトルをそれぞれ $\mathbf{S},\mathbf{N}$ とすると，$\mathbf{S},\mathbf{N},\mathbf{X}$ は Eqs (1),(2) のように表される。

$$\begin{aligned}&\mathbf{S}=(S_1,\ldots,S_k,\ldots,S_L),\\&\mathbf{N}=(N_1,\ldots,N_k,\ldots,N_L),\\&\mathbf{X}=(X_1,\ldots,X_k,\ldots,X_L),\end{aligned}\quad\cdots\cdots\cdots\cdots(1)$$

$$H_n:\ \mathbf{X}=\mathbf{N},\ H_s:\ \mathbf{X}=\mathbf{N}+\mathbf{S}.\quad\cdots\cdots\cdots\cdots(2)$$

ここで $H_n, H_s$ はそれぞれ騒音だけの状態 (“speech absent”)，音声と騒音が混ざった状態 (“speech present”) である。FFT 係数の確率密度関数が，Eq. (3) のように次元ごとに独立なガウス分布で表せると仮定する。

$$p(\mathbf{X}|H_n)=\prod_{k=1}^{L}\frac{1}{\pi\lambda_k^N}\exp\left(-\frac{|X_k|^2}{\lambda_k^N}\right),\quad\cdots\cdots\cdots\cdots(3)$$

$$p(\mathbf{X}|H_s)=\prod_{k=1}^{L}\frac{1}{\pi\left[\lambda_k^N+\lambda_k^S\right]}\exp\left(-\frac{|X_k|^2}{\left[\lambda_k^N+\lambda_k^S\right]}\right).$$

ここで $\lambda_k^N, \lambda_k^S$ は $N_k, S_k$ の分散であり，それぞれのパワーに対応している。すると $k$ 次元目の音声・非音声の尤度比は Eqs (4), (5) で表される。

$$\Lambda_k(X_k)=\frac{p(X_k|H_s)}{p(X_k|H_n)}=\frac{1}{1+\xi_k}\exp\left(\frac{\gamma_k\xi_k}{1+\xi_k}\right),\quad\cdots\cdots(4)$$

$$\xi_k=\frac{\lambda_k^S}{\lambda_k^N},\ \gamma_k=\frac{|X_k|^2}{\lambda_k^N}.\quad\cdots\cdots\cdots\cdots(5)$$

ここで $\xi_k, \gamma_k$ はそれぞれ事前，事後 SN 比と呼ばれる[9]。それぞれの次元ごとの尤度比の幾何平均により，Eq. (6) に従って，音声・非音声を判定でき，$\log\Lambda(\mathbf{X})$ が閾値 $\eta$ よりも大きければ $H_s$，小さければ $H_n$ となる。

$$\log\Lambda(\mathbf{X})=\frac{1}{L}\sum_{k=1}^{L}\log\left(\Lambda_k(X_k)\right)\underset{H_n}{\overset{H_s}{\gtrless}}\eta.\quad\cdots\cdots\cdots\cdots(6)$$

ここで $\lambda_k^N, \lambda_k^S$ はそれぞれ騒音，音声のモデルに当たり，

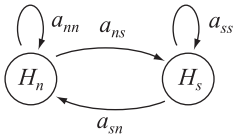



Fig. 1. State transition between speech and noise.

なんらかの方法で推定する必要がある。$\lambda_k^N$ は発話冒頭の騒音区間から推定しておく。$\lambda_k^S$ は尤度最大基準によりフレームごとに推定する。文献[4]には示されていないが，$\partial\Lambda_k(X_k)/\partial\lambda_k^S=0$ のとき，$\lambda_k^S=|X_k|^2-\lambda_k^N$ なので，Sohn の方法では騒音と音声のパワーが加法的であると仮定して音声モデル $\lambda_k^S$ を求めていることになる。$\xi_k^{(ML)}$ は Eq. (7) のようになるので，これを Eq. (6) に代入すると，最終的に音声・非音声の判別式は Eq. (8) のようになる。

$$\xi_k^{(ML)}=\gamma_k-1,\quad\cdots\cdots\cdots\cdots(7)$$

$$\log\Lambda^{(ML)}(\mathbf{X})=\frac{1}{L}\sum_{k=1}^{L}(\gamma_k-\log\gamma_k-1)\underset{H_n}{\overset{H_s}{\gtrless}}\eta.\quad\cdots\cdots(8)$$

Sohn の方法では，さらに HMM-Based Hang-Over と呼ばれる Fig. 1 に示される音声・非音声の状態遷移を考えることで，過去のフレームでの判別結果を考慮した音声区間検出を行っている。$a_{nn}, a_{ns}, a_{ss}, a_{sn}$ はそれぞれ図に示す遷移確率である。

$$a_{nn}+a_{ns}=1,\ a_{ss}+a_{sn}=1,$$

の関係がある。ここでは文献[4]に従い，$a_{ns}=0.2$，$a_{sn}=0.1$ とした。$t$ フレームにおける最終的な判別式は

$$\Gamma(t)=\frac{a_{ns}+a_{ss}\Gamma(t-1)}{a_{nn}+a_{sn}\Gamma(t-1)}\log\Lambda(\mathbf{X}),\quad\cdots\cdots\cdots\cdots(9)$$

$$\mathcal{L}(\mathbf{X})=\frac{a_{sn}}{a_{ns}}\Gamma(t)\underset{H_n}{\overset{H_s}{\gtrless}}\eta,\quad\cdots\cdots\cdots\cdots(10)$$

のようになる。

**〈2·2〉 確率密度比の推定法 (KLIEP)** LRT では〈2·1〉のように，騒音モデルと音声モデルそれぞれの確率密度 $p(\mathbf{X}|H_n), p(\mathbf{X}|H_s)$ を求め，それらの比 $\Lambda_k$ を求めることで，音声区間検出を行う。この方法を Fig. 2(a) に図示する。この場合，音声モデルを Eq. (7) により推定する方法[4]と，音声モデルをクリーンな音声から学習しておく方法[7]がある。ところで尤度比 $\Lambda$ を求めるためには，確率密度関数 $p(\mathbf{X}|H_n)$ および $p(\mathbf{X}|H_s)$ がそれぞれ求まる必要はなく，それらの比のモデルが直接推定できればよい。この方法を Fig. 2(b) に図示する。この密度比のモデルを直接推定する枠組みが密度比推定[8]であり，ロバスト性が向上するとされている。

特徴量 $\mathbf{Y}$ は各次元で独立であると仮定し，それぞれ密度比関数を独立に推定する。密度比推定では，密度比の分母・分子それぞれに対応する標本が必要である。すなわち，音声・非音声のラベル付けがされた特徴量の時系列データを学習に用いる。$k$ $(1\le k\le M)$ 次元目の特徴量の標本を，騒音のデータ ($H_n$ からの標本) を $\mathbf{Y_k^n}=\{Y_k^n(i)\}_{i=1}^{n^n}$，音声のデー

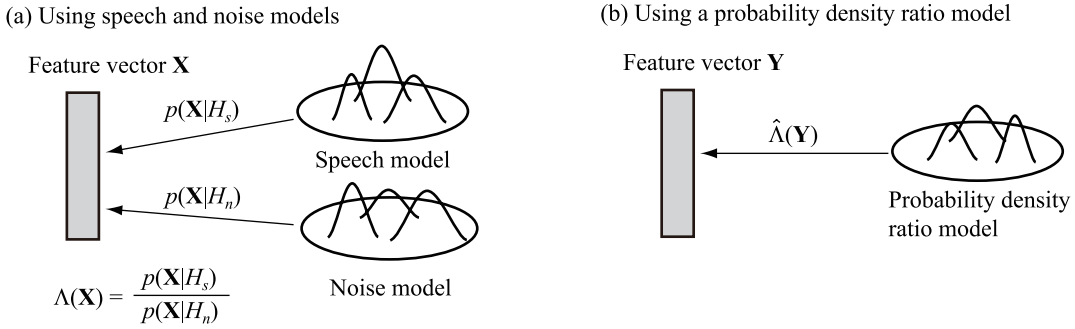



Fig. 2. Using speech and noise models and using probability density ratio model.

タ ($H_s$ からの標本) を $\mathbf{Y}^{\mathbf{s}}_{\mathbf{k}} = \{Y^s_k(i)\}^{n^s}_{i=1}$ と表す。この時，$\mathbf{Y}^{\mathbf{n}}_{\mathbf{k}}$, $\mathbf{Y}^{\mathbf{s}}_{\mathbf{k}}$ の統計量から，密度比関数を推定する方法は推定精度がよくないことが知られている[8]。例えばそれぞれの標本に対してガウシアンカーネルを仮定し，それぞれの確率密度関数を求めた後，それらの比を求めるものがそれにあたる。

ここでは直接密度比を推定する手法として有力な KLIEP (Kullback-Leibler Importance Estimation Procedure)[8] を用いた。KLIEP では，密度比

$$\Lambda_k(Y_k) = \frac{p(Y_k|H_s)}{p(Y_k|H_n)}, \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (11)$$

を線形モデルでモデル化する。これを $\hat{\Lambda}_k(Y_k)$ とし，Eq. (12) のように表す。

$$\hat{\Lambda}_k(Y_k) = \frac{\hat{p}(Y_k|H_s)}{p(Y_k|H_n)} = \sum_{l=1}^{b} \alpha_{k,l}\varphi_{k,l}(Y_k). \cdots\cdots\cdots (12)$$

$\alpha_{k,l}$ は非負の混合重みであり，$\varphi_{k,l}$ は非負の基底関数である。

密度比関数は $H_s$ からの標本が多いところで大きな値をとり，それ以外の場所でゼロに近い値を取る傾向にあるので，$\varphi_{k,l}$ には Eq. (13) で表されるガウシアンカーネル $K_{\sigma_k}(Y_k, Y^{ce}_{k,l})$ を用いるのが適当である。

$$\varphi_{k,l}(Y_k) = K_{\sigma_k}(Y_k, Y^{ce}_{k,l}) = \exp\left(-\frac{|Y_k - Y^{ce}_{k,l}|^2}{2\sigma^2_k}\right). \cdots\cdots\cdots\cdots\cdots\cdots (13)$$

$Y^{ce}_{k,l}$ と $\sigma_k$ はカーネルの平均と幅である。ここで未知数は $Y^{ce}_{k,l}$ と $\sigma_k$，$\alpha_{k,l}$ であり，特徴量の次元 $k$ ごとに以下の 4 手順で決定する。

（1） $\sigma_k$ を適当にいくつか決定する。
（2） $Y^{ce}_{k,l}$ は $\mathbf{Y}^{\mathbf{s}}_{\mathbf{k}}$ から $b$ 個ランダムに選ぶ。
（3） 以下に述べるプロセスで $\alpha_{k,l}$ を求める。
（4） $\sigma_k$ の適切な値は $n$-fold 交差確認法で決定する。

KLIEP では $\alpha_{k,l}$ を，ある標本 $y$ に対する Eq. (12) における分母の密度 $p(y|H_s)$ から分子の密度 $\hat{p}(y|H_s)$ への KL 情報量を最小にするように決定する。この際線形モデルにより，分子の密度 $\hat{p}(y|H_s)$ は $\hat{\Lambda}_k(y)p(y|H_n)$ で推定できることに注意すると，KL 情報量 $I$ は Eq. (14) で表される。

$$\begin{aligned} I(p(y|H_s); \hat{p}(y|H_s)) &= \int_{\mathcal{D}} p(y|H_s)\log\frac{p(y|H_s)}{\hat{p}(y|H_s)}dy \\ &= \int_{\mathcal{D}} p(y|H_s)\log\frac{p(y|H_s)}{p(y|H_n)}dy \\ &\quad - \int_{\mathcal{D}} p(y|H_s)\log\hat{\Lambda}_k(y)dy. \end{aligned} \cdots\cdots\cdots\cdots\cdots\cdots (14)$$

ここで $\mathcal{D}$ はデータの定義域である。第 1 項目は $\alpha_{k,l}$ に関して定数であるため無視できる。また $\hat{p}(y|H_s)$ は確率密度関数であるから，Eq. (15) の拘束条件を満たす必要がある。

$$\int_{\mathcal{D}} \hat{p}(y|H_s)dx = \int_{\mathcal{D}} \hat{\Lambda}_k(y)p(y|H_n)dy = 1. \cdots\cdots (15)$$

よって KL 情報量を最小にするためには Eq. (14) の第 2 項を，Eq. (15) の拘束条件の下で最大化すればよい。期待値操作を標本平均で近似することで，Eq. (16) の凸最適化問題が得られる。凸最適化問題は，勾配上昇と制約付加により大局的な最適解に至る[10]。最適解は疎になる傾向にあるため，いくつかの $\alpha_{k,l}$ は 0 である。このことはモデルの尤度を計算する際の計算負荷にとって有利な性質である。

$$\arg\max_{\{\alpha_{k,l}\}^b_{l=1}} \left[\sum_{i=1}^{n^s} \log\left(\sum_{l=1}^{b} \alpha_{k,l}\varphi_{k,l}(Y^s_k(i))\right)\right]. \cdots\cdots\cdots\cdots (16)$$

離散化した拘束条件は Eq. (17) である。

$$\sum_{l=1}^{b} \alpha_{k,l}\left[\frac{1}{n^n}\sum_{i=1}^{n^n} \varphi_{k,l}(Y^n_k(i))\right] = 1. \cdots\cdots\cdots\cdots\cdots (17)$$
$$(\alpha_1, \alpha_2, \ldots, \alpha_b \geq 0)$$

これより，Eq. (12) の尤度比モデルを得る。モデル学習の実装は[11]の Matlab® コードを参考にした。

これらより Eq. (6) と同様にして，Eq. (18) により音声・非音声を判定できる。

$$\hat{\Lambda}(\mathbf{Y}) = \frac{1}{L}\sum_{k=1}^{L} \hat{\Lambda}_k(Y_k) \underset{H_n}{\overset{H_s}{\gtrless}} \eta. \cdots\cdots\cdots\cdots\cdots\cdots (18)$$

**〈2·3〉 確率密度比の推定法 (KLIEP) の音声区間検出への応用** 上記の KLIEP を音声区間検出に応用するため

には，適切な特徴量の選択と環境適応化という課題がある。第 1 に特徴量の選択である。KLIEP ではカーネルの制約から，重なりが少なく値が小さいもしくは大きいところで確率密度が低い必要がある。音声認識には MFCC がよく用いられるが，これは音声と騒音の境界があいまいである。実際 MFCC を特徴量として使った場合〈3·3〉に示すようにカーネルの推定精度が低かった。また Sohn の方法のように，パワースペクトルを使うと特徴量のレンジが大きくなり，線形モデルの仮定から外れるためカーネルの推定精度が低下する。そこで提案法には，Sohn の方法と同じく音声区間で値が大きく，騒音区間で値の小さい特徴量である帯域別のパワーを用いるが，レンジを適切な範囲に収めるために，対数を取った。

第 2 に実際の環境では騒音が多様なため，学習時と評価時で騒音のミスマッチが起こるという問題がある。Sohn の方法や提案法は，音声と騒音の間に相対的な帯域別パワー差があることを前提とした手法なので，平均値の調整は必須である。例えば，同じ音声に対してマイクのゲインを変えただけで音声と騒音の境界がシフトしてしまうことからもこれはわかる。Sohn の方法では，分散を用いることで平均値がシフトする影響を排除している。分散に関しても Sohn の方法では冒頭から学習しており，これは帯域別の事前 SNR を求めていることに相当している。騒音や音声はダイナミクスが大きいので，騒音抑圧や音声認識のタスクにおいて分散の適応が有効であることはよく知られている [12][13]。提案法では学習時と評価時の環境の違いを考慮する (環境適応化) ために，評価データの始めの 10 フレームの騒音の特徴量の平均と分散が，学習データの騒音の平均と分散と等しくなるように正規化を行うことでこの影響を低減している。

**〈2·4〉 閾値の自動設定**　音声区間検出には適切な閾値 $\eta$ の設定が不可欠であるが，閾値は環境によって最適な値が異なるため設定が難しい。閾値の変動に対しては，音声と騒音のラベル付けされた前フレームまでの尤度比のクラスタの判別分析によって決定した。騒音だけの情報から閾値を計算することはできないので，始めは何らかの初期値 $\eta_0$ に従い音声区間検出を行う。音声区間が検出されたら過去の尤度比 (Sohn の方法では $\log\Lambda$，提案法では $\hat{\Lambda}$) をクラスタリングする。例えば $K$-means アルゴリズムを用いて，Fig. 3 のように $N_{cl}$ 個のクラスターに分ける。この場合尤度比の大きい方には，音声のデータが偏り，小さい方には騒音のデータが偏るため，音声データに該当するクラスタと騒音データに該当するクラスタの平均値の中間をとることで閾値を計算できる。ここではそれぞれのクラスターの平均値 $\mu_{n,s}$ と分散 $\sigma'_{n,s}$ を計算し，判別分析により Eq. (19) のようにクラスタの内分点を閾値として用いることで閾値を決定できる。

$$\eta = \frac{\mu_s\sigma'_n + \mu_n\sigma'_s}{\sigma'_n + \sigma'_s} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (19)$$

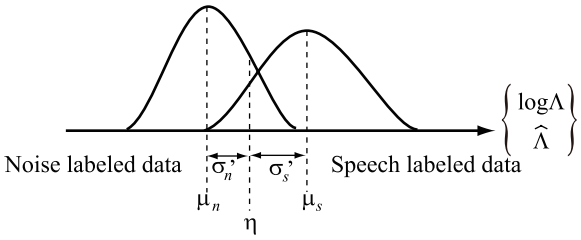



Fig. 3. Determination of threshold $\eta$ using discriminant analysis.

## 3. 実　　験

**〈3·1〉 実験条件 (評価データ)**　提案法を音声区間検出評価環境 CENSREC-1-C [14] により評価した。評価環境は，

（1） 人声・歩行音が主な環境 (RESTAURANT(学生食堂))

（2） 道路交通騒音が主な環境 (STREET(高速道路))

の 2 つである。それぞれの環境に対して，背景騒音が平均 60dBA の HIGH と 70dBA の LOW の 2 つの SN 比のデータがある。これらは実収録した音声である。発声内容は 1〜12 桁の連続数字であり，これを 8〜10 回，約 2 秒間の間隔で発声した。

サンプリング周波数は 8 kHz，STFT の窓幅は 25 ms，フレームシフトは 10 ms とした。FFT の次元は 256 とし，対称性を考えて $L = M = 129$ とした。Sohn による方法は，始めの 10 フレームから $\lambda_k^N$ を算出した。

**〈3·2〉 実験条件 (学習データ)**　提案法の密度比関数は，CENSREC-4 [14]の 8 種類の残響を畳み込んだのち，騒音を SN 比 {5, 10, 20, 25, 30} [dB] で重畳した音声より学習した。CENSREC-4 のサンプリング周波数は 16 kHz だが，評価環境である CENSREC-1-C に合わせて，8 kHz にダウンサンプリングした。学習データは上記データからの音声と騒音をランダムに 16000 フレーム分 (160 秒分) ピックアップした。(すなわち $n^n = n^s = 16000$ である。) カーネル $\varphi_{k,l}$ の最大数 $b$ は 20 とした。ただし上述の通り，ス

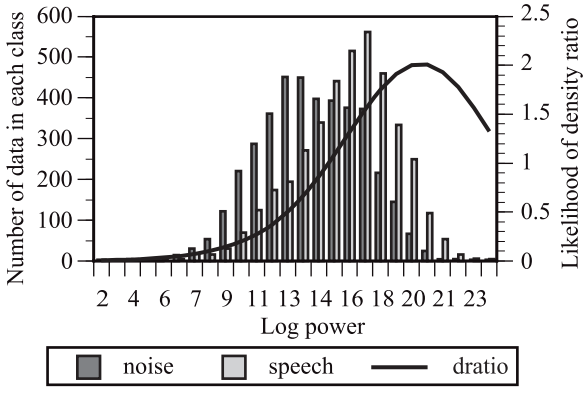



Fig. 4. Histogram of log power of noise and speech of training data and estimated probability density ratio (dratio). (Feature is the fifteenth dimension of logarithmic power spectrum.)

パース性によりいくつかの $\varphi_{k,l}$ の混合重み $\alpha_{k,l}$ は 0 となる。$K_{\sigma_k}$ の幅 $\sigma_k$ は 5-fold 交差確認法で決定した。

**〈3·3〉 結果と考察 (密度比モデル)** 学習に用いた log パワーの音声と，騒音区間での分布および KLIEP による学習の結果得られた密度比関数を Fig. 4 に示す。音声帯域である $k = 15$ (= 469 Hz) の場合である。音声と騒音のオーバーラップを考慮して密度比モデルが学習できている。この場合，非零の $\alpha_{15,l}$ は 13 個であった。残りの 7 個の $\alpha_{15,l}$ は 0 であり，この分の尤度計算を省略することができる。これは特徴量に用いた log パワーがレンジが小さく，音声と騒音の分離性がよい特徴量だからである。MFCC は音声区間で大きく，騒音区間で小さいわけではないので，音声と騒音の分離性能が悪く，密度比関数が全体的になだらかで識別性能が劣るものになってしまった。例えば MFCC の 1 次元目の密度比関数の推定結果を Fig. 5 に示す。このようにピーク位置がほぼ一致するため，密度比関数が全域でフラットで識別できない。

**〈3·4〉 結果と考察 (密度比モデルによる音声区間検出)** 音声区間検出の性能を比較するために，提案法 (proposed) と従来法 2 種を比較した。従来法としたのは，CENSREC-1-C にベースラインとして結果が付属している音声のパワーを用いる方法 (base) と，〈2·1〉に示した Sohn による方法 (sohn) である。proposed，sohn ともに〈2·4〉の閾値の自動決定法を用いた。様々な閾値設定で試したものの，最適な場合でも Accuracy で比較して若干性能がよくなるか同程度あったため，自動決定法を両手法でベースラインとして採用した。CENSREC-1-C 付属の集計スクリプトにより，以下の Eqs (20), (21) に示す Correct と Accuracy を算出した。

$$Correct = \frac{N_c}{N_u} \times 100\ [\%], \quad \cdots\cdots\cdots\cdots\cdots\cdots (20)$$

$$Accuracy = \frac{N_c - N_f}{N_u} \times 100\ [\%]. \quad \cdots\cdots\cdots\cdots\cdots (21)$$

ここで $N_u$ は総発話数，$N_c$ は正検出数 (300 ms のマージンをつけて正誤判定している)，$N_f$ は誤検出数である。

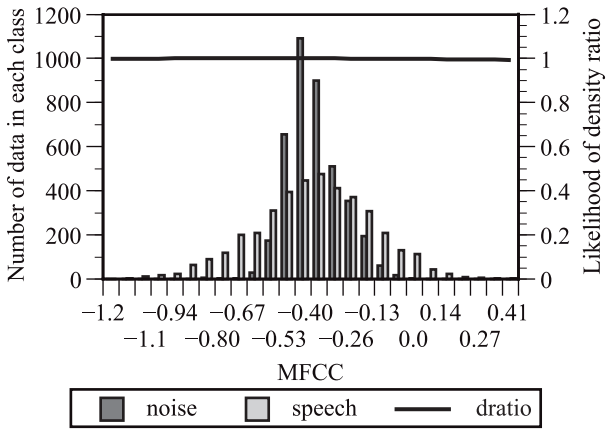



Fig. 5. Histogram of MFCC of noise and speech of training data and estimated probability density ratio (dratio). (Feature is the first dimension of MFCC.)

Correct と Accuracy の結果を Fig. 6 に示す。提案法は，平均の Correct を base に比べて 28.6%，sohn に比べて 6.0%向上させ，平均の Accuracy もそれぞれ 74.8%，8.5%向上させた。特に非定常な騒音である RESTAURANT の時に，提案法の利点がよく表れている。

Fig. 7 に sohn の場合の尤度比を，Fig. 8 に提案法の場合の尤度比を示す。これらは同じ発話内容だが，実環境で収録された発話のため発話長が異なる。STREET(LOW) の場合はおおむね同傾向であるが，RESTAURANT(HIGH) の場合は提案法の非音声区間における尤度比が小さいことがわかる。RESTAURANT は非定常性が高いため，sohn では初めの 10 フレームから学習した $\lambda_k^N$ が実際の環境とミスマッチを起こし騒音に対しても尤度が高くなってしまい，それがノイズとなったことが影響している。これに対して提案法では密度比モデルを用いたことで，騒音の誤推定に対して，sohn よりもロバストになったと考えられる。

**〈3·5〉 結果と考察 (閾値の自動決定法)** 閾値の自動決定法の検証のため，RESTAURANT (HIGH) と

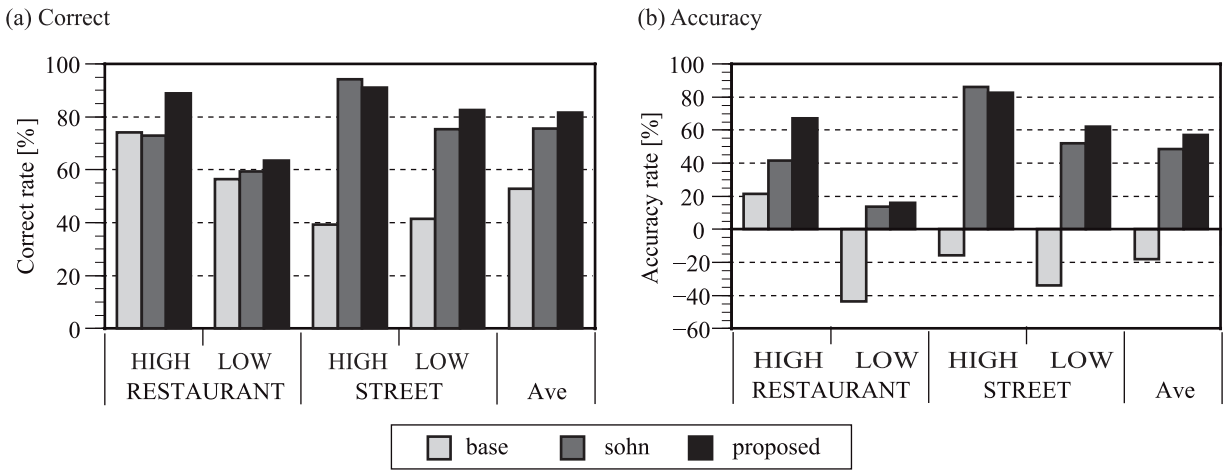



Fig. 6. Correct and accuracy rate[%] of proposed method (proposed) compared with those of CENSREC-1-C baseline method (base) and Sohn's method (sohn).

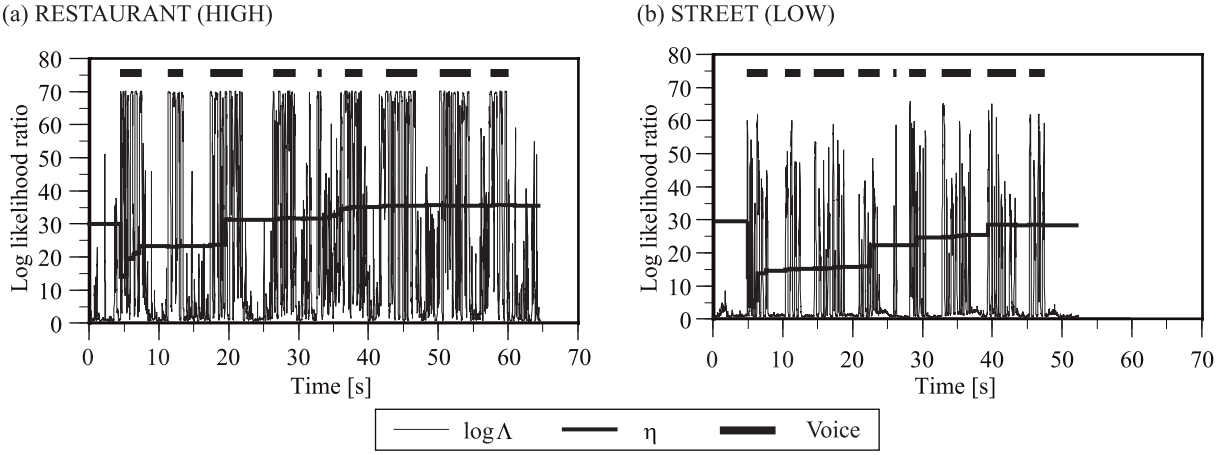



Fig. 7. Log likelihood ratio calculated by Sohn's method ($\log\Lambda$) and automatically determined threshold $\eta$.

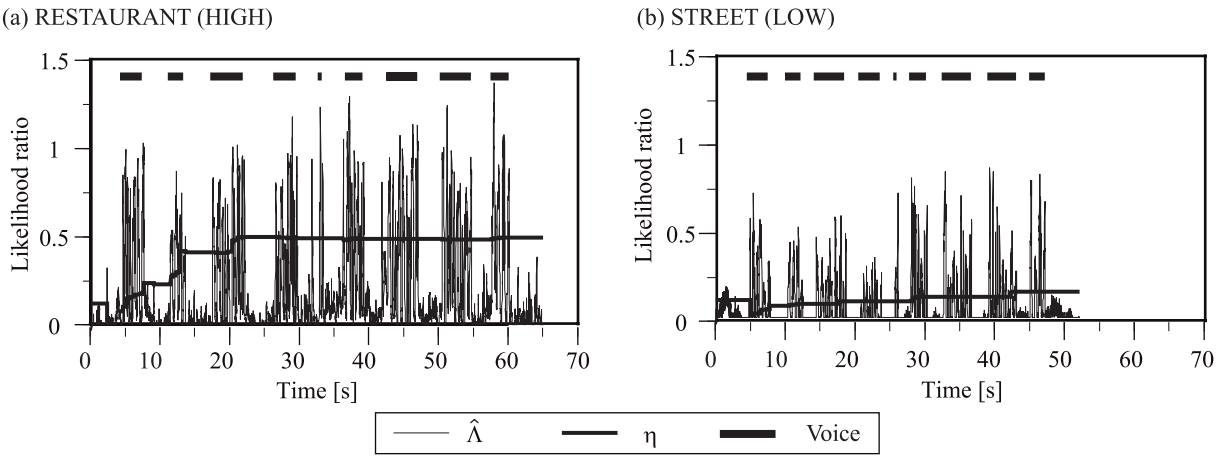



Fig. 8. Likelihood ratio calculated by the proposed method ($\hat{\Lambda}$) and automatically determined threshold $\eta$.

STREET (LOW) における sohn による尤度比と閾値の関係を Fig. 7 (a)(b) にそれぞれ示す。閾値の初期値 $\eta_0$ は 30 とした。(a) では最適値は 40 程度であり，よく推定できている。(b) では騒音のクラスタが 0 付近に固まっているため，22 秒付近で音声を騒音に誤推定したことが原因で最適な閾値 10 よりも高めになってしまっている。騒音のクラスタの分散が小さい場合に誤推定が大きな影響を与える問題に関しては，今後の検討が必要である。次に密度比モデルを用いた場合を Fig. 8 に示す。$\eta_0$ は 0.12 とした。こちらは (a) では最適値は 0.4 程度であり，よく推定できている。(b) でも上記と違って誤推定がなかったため，閾値は最適値 0.2 付近のまま維持されている。まとめると，背景騒音が学習時とミスマッチな (a) では，どちらの手法でも背景騒音が大きいため，数発話経ると閾値が高めで安定し，誤検出を減らすのに有効である。これに対して，背景騒音が定常的な (b) では，学習時とのミスマッチが小さく (a) に比べて比較的低めで安定し，誤棄却を減らすのに有効である。

## 4. まとめと今後の課題

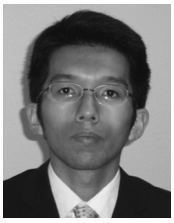
耐騒音性に優れる音声区間検出手法である LRT の性能改善を目的として，音声モデル・騒音モデルの尤度を算出したのちにそれらの尤度比を求めていた従来法に対して，確率密度比推定により得られた密度比モデルを用いることでそれらの尤度比を直接算出する音声区間検出法を提案した。音声区間検出評価環境である CENSREC-1-C を用いた実験の結果，提案法はモデルの誤推定にロバストになることで，特に非定常な騒音環境下において，従来法よりも有効であることが示された。また音声区間検出において問題となる閾値の自動決定法を提案し，背景騒音の特徴に合わせて，閾値が決定されることを示した。今後の課題としては，音声と騒音の識別性により優れた特徴量の検討や，騒音モデルの逐次推定法の検討が挙げられる。

## 文　献

（1） 石塚健太郎・藤本雅清・中谷智広：「音声区間検出技術の最近の研究動向」，音響誌，Vol.65, pp.537–543 (2009)
（2） 藤本雅清：「音声区間検出の基礎と最近の研究動向」，信学会技報，Vol.SP2010-23, pp.7–12 (2010)
（3） L. Rabiner and M. Sambur: "An algorithm for determining the endpoints of isolated utterances", *The Bell System Technical Journal*, Vol.54, pp.297–315 (1975)
（4） J. Sohn, N.S. Kim, and W. Sung: "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, Vol.6, pp.1–3 (1999)
（5） O. Pernía, J.M. Górriz, J. Ramírez, C.G. Puntonet, and I. Turias: "An effi-

cient VAD based on a generalized gaussian PDF", International Conference on Advances in Nonlinear Speech Processing, pp.246–254 (2007)
（6） J.H. Chang, N.S. Kim, and S.K. Mitra: "Voice activity detection based on multiple statistical models", *IEEE Trans. Signal Processing*, Vol.54, pp.1965–1976 (2006)
（7） 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", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2008), pp.4441–4444 (2008)
（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. Computer Vision and Applications*, Vol.1, pp.183–208 (2009)
（9） Y. Ephraim and D. Malah: "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.32, pp.1109–1121 (1984)
（10） S. Boyd and L. Vandenberghe: Convex Optimization, Cambridge Univ Press (2004)
（11） http://sugiyama-www.cs.titech.ac.jp/˜sugi/software/kliep/
（12） S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath: "Dynamic Noise Adaptation", IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP2006), pp.1197–1200 (2006)
（13） M. Delcroix, T. Nakatani, and S. Watanabe: "Static and Dynamic Variance Compensation for Recognition of Reverberant Speech With Dereverberation Preprocessing", *IEEE Trans. Audio, Speech, and Language Processing*, pp.324-334 (2009)
（14） http://research.nii.ac.jp/src/eng/list/

**太刀岡　勇　気**



（正員） 2006 年 東京大学工学部建築学科卒業。2008 年同大学大学院新領域創成科学研究科修士課程修了。在学中，数値音響解析に関する研究に従事。同年三菱電機 (株) 入社。以来，音声認識，マイクロフォンアレイ信号処理の研究開発に従事。現在，同社情報技術総合研究所音声・言語処理部研究員。2008 年日本建築学会優秀修士論文賞。

**花　沢　利　行**

（非会員） 1985 年東京都立大学理学部卒業。同年三菱電機 (株) 入社。以来，音声認識の研究開発に従事。1987～1990 年 ATR 音声翻訳通信研究所に出向。現在，三菱電機 (株) 情報技術総合研究所音声・言語処理部主任研究員。

**成　田　知　宏**

（非会員） 1995 年東京工業大学工学部卒業。1997 年同大学大学院修士課程修了。同年三菱電機 (株) 入社。以来，音声認識の研究開発に従事。現在，同社情報技術総合研究所音声・言語処理部 主任研究員。

**石　井　　純**

（非会員） 1988 年 新潟大学工学部卒業。1990 年同大学大学院修士課程修了。同年三菱電機 (株) 入社。1995～1997 年 ATR 音声翻訳通信研究所に出向。現在，三菱電機 (株) 情報技術総合研究所音声・言語処理部 音声処理グループマネージャー。