音声と騒音の密度比推定を用いた音声区間検出法*

◎太刀岡勇気, 花沢利行, 成田知宏, 石井純 (三菱電機・情報総研)

1 はじめに

音声区間検出は、音声認識・強調において須要の前処理である。最も基本的なものは、音声のパワーが騒音のそれよりも大きいことを活用するものである [1]。この方法は音声が騒音に埋もれる高騒音環境下で有効でないため、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が提案されている [2]。また、これの発展として、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音声モデルと観測された騒音のモデルの尤度比による方法も提案されている [3]。上記方法に共通して、フレームごとに音声モデルの尤度と騒音モデルの尤度を算出し、それらの比をとることで音声・非音声を判断している。

ところが、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、それぞれ尤度は必要ない。近年機械学習の分野では、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推定する手法が提案されている [4]。本報では本手法を上記音声区間検出における尤度比の算出問題に応用した。2.1 で従来法に関して述べたのち、2.2 で確率密度比推定を音声区間検出に用いる方法を提案する。

また音声区間検出器は設定した閾値と比較し音声・非音声を判断するため、閾値により性能が左右されその決定が難しいという問題がある。2.3 でクラスタリング分析を応用した閾値の自動決定法を提案する。

2 音声区間検出法

2.1 Sohn の方法

尤度比を用いる代表的な手法 [2] 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_0 と音声区間 H_1 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、は Eqs. (1),(2) のように表される。

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

$$H_0 : \ \mathbf{X} = \mathbf{N}, \ H_1 : \ \mathbf{X} = \mathbf{N} + \mathbf{S}. \tag{2}$$

ここで H_0, H_1 それぞれの FFT 係数の確率密度関数が、Eq. (3) のように各次元で独立なガウス分布で表せると仮定する。

$$p(\mathbf{X}|H_0) = \prod_{k=1}^{L} \frac{1}{\pi\lambda_k^N} \exp\left(-\frac{|X_k|^2}{\lambda_k^N}\right), \tag{3}$$

$$p(\mathbf{X}|H_1) = \prod_{k=1}^{L} \frac{1}{\pi[\lambda_k^N + \lambda_k^S]} \exp\left(-\frac{|X_k|^2}{[\lambda_k^N + \lambda_k^S]}\right).$$

ここで λ_k^N, λ_k^S は N_k, S_k の分散を表す。すると k 次元目の音声・非音声の尤度比は Eq. (4) で表される。

$$\Lambda_k(X_k) = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left(\frac{\gamma_k \xi_k}{1+\xi_k}\right), \tag{4}$$

$$\xi_k = \lambda_k^S / \lambda_k^N, \gamma_k = |X_k|^2 / \lambda_k^N. \tag{5}$$

ここで ξ_k, γ_k はそれぞれ事前、事後 SN 比と呼ばれる。それぞれの次元ごとの尤度比の幾何平均により、音声・非音声を判断できる。

$$\log \Lambda(\mathbf{X}) = \frac{1}{L} \sum_{k=1}^{L} \log\left(\Lambda_k(X_k)\right) \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6}$$

$\log \Lambda(\mathbf{X})$ が閾値 η よりも大きければ H_1、小さければ H_0 となる。ここで λ_k^N は観測された騒音の分散を集めた騒音モデルであり、事前に推定しておく。λ_k^S は音声モデルであり、なんらかの基準で推定する。ML 基準により推定すると、$\xi_k^{(ML)}$ 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。

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

$$\log \Lambda^{(ML)}(\mathbf{X}) = \frac{1}{L} \sum_{k=1}^{L} \big(\gamma_k - \log \gamma_k - 1\big). \tag{8}$$

2.2 確率密度比の推定法 (KLIEP) の音声区間検出への応用

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 $p(\mathbf{X}|H_0), p(\mathbf{X}|H_1)$ を求め、それらの比 Λ_k を求めることで、音声区間検出を行うことができる。これを Fig. 1(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 [2] と、音声モデルをクリーンな音声から学習しておく方法 [3] がある。ところで尤度比 Λ を求めるためには、確率密度関数 $p(\mathbf{X}|H_0)$ および $p(\mathbf{X}|H_1)$ 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

*Voice activity detection using density ratio estimation of speech and noise, by TACHIOKA, Yuuki, HANAZAWA, Toshiyuki, NARITA, Tomohiro, ISHII, Jun (Mitsubishi Electric Corp.).

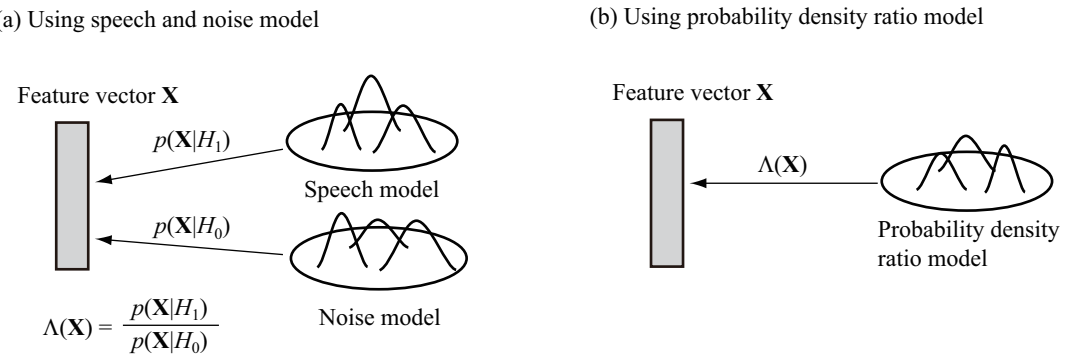

Fig. 1　Using speech and noise model and using probability density ratio model.

れが Fig. 1(b) である。このモデルを直接推定する枠組みが密度比推定 [4] であり、モデル推定がより単純化されロバスト性が向上する。

特徴量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、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 $(1 \leq k \leq M)$ 次元目の特徴量の標本を、騒音のデータ (H_0 からの標本) を $\mathbf{X_k^0} = \{X_k^0(i)\}_{i=1}^{n_0}$、音声のデータ ($H_1$ からの標本) を $\mathbf{X_k^1} = \{X_k^1(i)\}_{i=1}^{n_1}$ と表す。

以下、密度比の推定に用いた KLIEP (Kullback-Leibler Importance Estimation Procedure) に関して述べる [4]。密度比は $\Lambda_k(X_k) = p(X_k|H_1)/p(X_k|H_0)$ となるが、これを Eq. (9) の線形モデルでモデル化したものを $\hat{\Lambda_k}(X_k)$ とする。

$$\hat{\Lambda_k}(X_k) = \sum_{l=1}^{b} \alpha_l \varphi_l(X_k). \tag{9}$$

α_l は非負の混合重みであり、φ_l は非負の基底関数である。φ_l が定まったとき、α_l は以下のように求める。線形モデル (Eq. (9)) により、分子の密度 $\hat{p}(X_k|H_1)$ を $\hat{\Lambda_k}(X_k)p(X_k|H_0)$ で推定できる。KLIEP では α_l を、$p(X_k|H_1)$ から $\hat{p}(X_k|H_1)$ への KL 情報量を最小にするように決定する。ある標本 x に対する KL 情報量 I は Eq. (10) で表される。

$$I\left(p(x|H_1); \hat{p}(x|H_1)\right) = \int_{\mathcal{D}} p(x|H_1) \log \frac{p(x|H_1)}{p(x|H_0)} dx - \int_{\mathcal{D}} p(x|H_1) \log \hat{\Lambda_k}(x) dx. \tag{10}$$

ここで $\mathcal{D}$ はデータの定義域である。第 1 項目は α_l に関して定数であるため無視できる。$\hat{p}(x|H_1)$ は確率密度関数であるから、Eq. (11) の拘束条件を満たす必要がある。

$$\int_{\mathcal{D}} \hat{p}(x|H_1) dx = \int_{\mathcal{D}} \hat{\Lambda_k}(x) p(x|H_0) dx = 1. \tag{11}$$

よって KL 情報量を最小にするためには Eq. (10) の第 2 項を、Eq. (11) 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. (12) 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。最適解は疎になる傾向にあるため、いくつかの α_l は 0 である。

$$\underset{\{\alpha_l\}_{l=1}^{b}}{\arg\max} \left[\sum_{i=1}^{n_1} \log \left(\sum_{l=1}^{b} \alpha_l \varphi_l(X_k^1(i)) \right) \right]. \tag{12}$$

離散化した拘束条件は Eq. (13) である。

$$\sum_{l=1}^{b} \alpha_l \left[\frac{1}{n} \sum_{i=1}^{n_0} \varphi_l(X_k^0(i)) \right] = 1. \tag{13}$$

これより、Eq. (9) の尤度比モデルを得る。

密度比関数は H_1 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、カーネル φ_l には Eq. (14) で表されるガウシアンカーネル $K_\sigma(X_k, X_l^{ce})$ を用いる。

$$\varphi_l(X_k) = K_\sigma(X_k, X_l^{ce}) = e^{-\frac{|X_k - X_l^{ce}|^2}{2\sigma^2}}. \tag{14}$$

K_σ の中心 X_l^{ce} は、$\mathbf{X_k^1}$ から b 個ランダムに選ぶことにすれば、カーネルの幅 σ は n-fold 交差確認法で決定できる。(モデル学習の実装は [5] の Matlab コードを参考にした。)

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 N_N 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それらと等しくなるように正規化を行う。

2.3　閾値の自動決定法

音声区間検出には適切な閾値 η の設定が不可欠である。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が、これを自動的に決定するアルゴリズムに関する研究は見られない。以下では、閾値の設定をクラスタリング分析により自動的に行う手法を提案する。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 η_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 $\log \Lambda$) をクラスタリングする。例

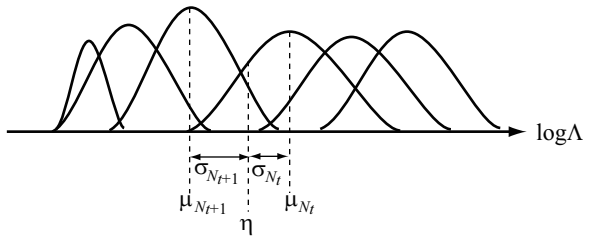

Fig. 2　Determination of threshold η using N_{cl} clusters.

えば K-means アルゴリズムを用いて、Fig. 2 のように $N_{cl}(\geq 3)$ 個のクラスターに分ける。この場合 $\log\Lambda$ 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。例えばそれぞれのクラスターの平均値 μ_j と分散 σ_j $(1 \leq j \leq N_{cl})$ を計算し、μ_j の大きさでソートし、N_t 番目と N_t+1 番目の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。

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[6] により評価する。評価環境は、人声、歩行音が主な環境 (RESTAURANT(学生食堂)) と、道路交通騒音が主な環境 (STREET(高速道路)) の 2 つで実収録している。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。被験者は、男女 5 名 (計 10 名) である。(20 歳前後男女各 3 名, 30 歳前後男女各 1 名, 40 歳以上男女各 1 名) 1 被験者に対して、各騒音環境および各 SN 比につき 1〜12 桁の連続数字を 8〜10 回、約 2 秒間の間隔で発声した音声を 1 つのファイルとして、計 4 ファイルで評価した。(総発話数：38〜39 発話) 付属の集計スクリプトにより、以下の Eqs.(15),(16) に示す Correct と Accuracy を算出した。

$$Corr. = N_c/N_u \times 100\ [\%], \tag{15}$$

$$Acc. = (N_c - N_f)/N_u \times 100\ [\%]. \tag{16}$$

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。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 であり、対称性を考えて $L = M = 129$ とした。Sohn による方法は、始めの 10 フレームから騒音モデルを学習した。提案法は特徴量に log パワーを用い、始めの $N_N(= 10)$ フレームで環境適応を行った。

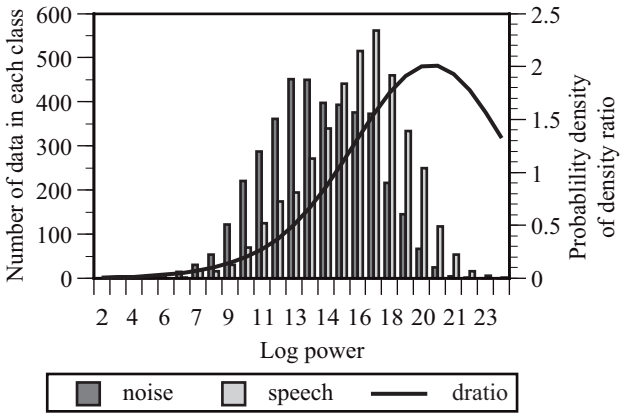

Fig. 3　Histogram of log power of noise and speech and probability density ratio (dratio) of training data. Feature is a logarithm of the fifteenth dimension of power spectrum.

3.2　実験条件 (学習データ)

提案法の密度比関数は、CENSREC-4[6] の 8 種類の残響と騒音を SN 比 {5, 10, 20, 25, 30} [dB] で重畳した音声より学習した。ただし 8 kHz にダウンサンプリングした。学習データは上記の音声からランダムにピックアップして、n_0, n_1 は 16000 フレーム (160 秒分) とした。使用する φ_l の最大数 b は 20 とした。ただし上述の通り、スパース性によりいくつかの φ_l の混合重み α_l は 0 となる。K_σ の幅 σ は 5-fold 交差確認法で決定した。

学習に用いた log パワー (15 次元目) 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. 3 に示す。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とがわかる。この場合、非零の α_l は 13 個であった。

3.3　結果と考察 (密度比モデルによる音声区間検出)

提案法 (proposed) と従来法 2 種 (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)) を比較した。proposed、sohn ともに閾値の自動決定法を用い、N_{cl} は 10、N_t は 3 とした。結果を Fig. 4 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。これは尤度比モデルを用いたことで、騒音モデルの誤推定に対して、sohn よりもロバストになったためと思われる。

3.4　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 STREET (LOW) における sohn による

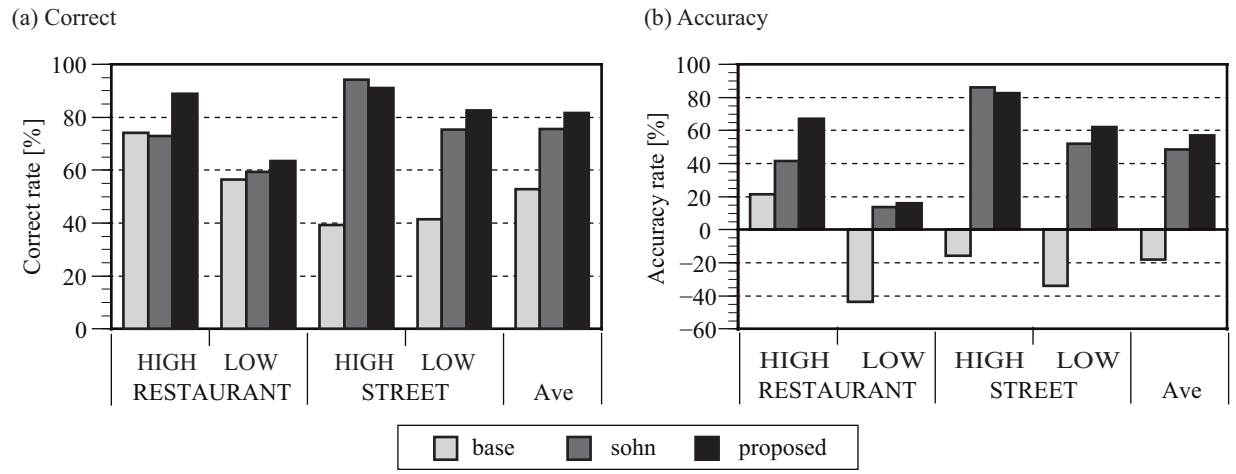

Fig. 4　Correct and accuracy rate[%] of proposed method (proposed) compared with those of CENSREC-1-C baseline method (base) and Sohn's method (sohn).

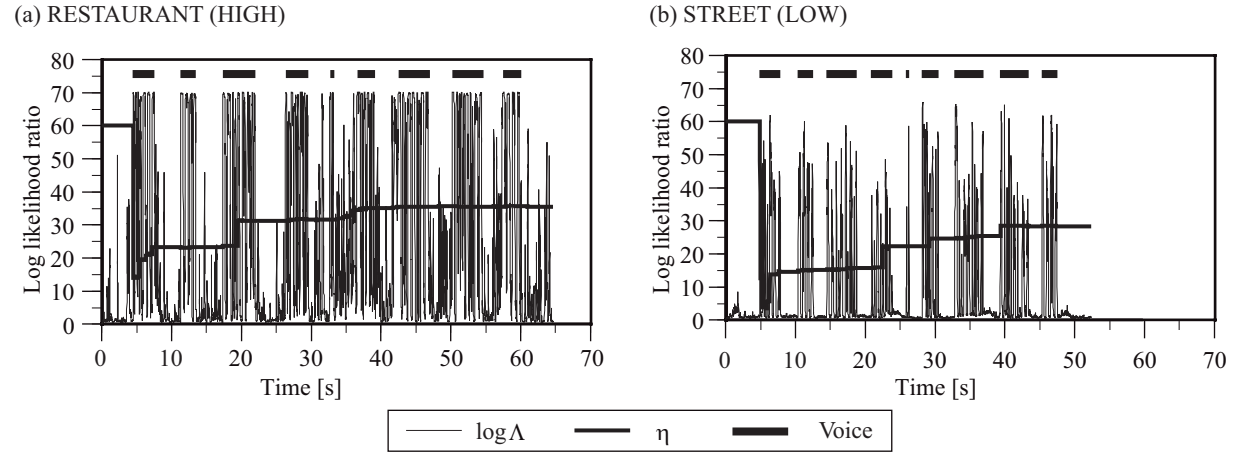

Fig. 5　Log likelihood ratio calculated by Sohn's method ($\log\Lambda$) and automatically determined threshold η.

尤度比と閾値の関係を Fig. 5(a)(b) にそれぞれ示す。閾値の初期値 η_0 は 60 とした。背景騒音が非定常な (a) では、数発話経ると閾値が高めで安定するため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、比較的低めで安定するため、誤棄却を減らすのに有効である。

4　まとめと今後の課題

確率密度比推定による密度比モデルを用いた音声区間検出法を提案した。非定常な騒音環境下において、特に従来法よりも効果的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た。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。今後の課題は、音声の性質を使った特徴量 (MFCC 等) やそれらとの組み合わせにより、さらに性能を改善することである。

参考文献

[1] L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech.*, **54**, 297–315 (1975).

[2] J. Sohn *et al.*, "Statistical model-based voice activity detection," *IEEE Signal Processing Letters*, **6**, 1–3 (1999).

[3] M. Fujimoto *et al.*, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," *in Proceedings of ICASSP*, 4441–4444 (2008).

[4] M. Sugiyama *et al.*, "A density-ratio framework for statistical data processing," *IPSJ Transactions on Computer Vision and Applications*, **1**, 183–208 (2009).

[5] http://sugiyama-www.cs.titech.ac.jp/~sugi/software/KLIEP/.

[6] http://research.nii.ac.jp/src/eng/list/.