情報処理学会論文誌 Vol.60 No.4 1162–1170 (Apr. 2019)

補助音声特徴量によるDNN適応を用いた音声区間検出

太刀岡 勇気[1,a)]

受付日 2018年6月5日, 採録日 2019年1月15日

概要：音声区間検出は，騒音環境下で音声認識を行う際には必須の前処理である．音声区間検出を行う際には，パワーに基づく方法がよく使われる．しかしながら，この方法は高騒音下において性能の低下が著しいため，近年ではスペクトルの形状を考慮するような方法が提案されている．とりわけ深層神経回路網（deep neural network; DNN）に基づく方法が性能が高いことが知られている．音声認識や音声強調の分野では，DNN を対象の環境に適応させて性能を向上させるために，補助特徴量が使われる．DNN に基づく音声区間検出の性能をさらに向上させるため，本論文では 2 つの音声のモデル化に基づく特徴量とそれらの結合を提案する．第 1 は非負値行列因子分解のアクティベーション，第 2 は音声認識の音響モデルの音響スコアを使うものである．騒音下音声区間検出の実験により，DNN に基づく手法は従来の方法を性能を上回り，2 つの補助特徴量は，フレーム別の音声区間検出精度，音声認識の単語正解精度の両観点から有効であることが分かった．

キーワード：DNN に基づく音声区間検出，非負値行列因子分解，音声認識，補助特徴量

Voice Activity Detection Using DNN Adaptation with Auxiliary Speech Features

Yuuki Tachioka[1,a)]

Received: June 5, 2018, Accepted: January 15, 2019

Abstract: Voice activity detection (VAD) is an essential pre-process for automatic speech recognition (ASR) in noisy environments. Power-based methods are widely used; however, because these methods are susceptible to noise, recently, methods that consider the shape of spectrum have been proposed. In particular, deep neural network (DNN) based methods have outperformed previous methods. In the fields of ASR and speech enhancement, to improve their performance by adapting DNNs to a target environment, auxiliary features are used. To improve the performance of DNN-based VAD further, this paper proposes two types of auxiliary features based on speech modelings and their combination. The first is activation of non-negative matrix factorization and the second is acoustic score of ASR acoustic models. Experimental results for noisy VAD tasks demonstrated that DNN-based methods outperformed one of the most effective conventional methods and that both auxiliary features improved performance in terms of both frame-level VAD accuracy and ASR word accuracy.

Keywords: DNN-based voice activity detection, non-negative matrix factorization, automatic speech recognition, auxiliary features

1. はじめに

音声認識技術の進展にともない，発話スイッチを使わない遠隔発話を許容する音声インタフェースが多く出回るようになってきた．このようなインタフェースを利用する機会が増えると，特に高騒音環境での音声区間検出の性能が重要になってくる[*1]．高騒音環境下では音声区間を正確に検出することが難しい．誤検出をすれば無音区間に対して認識を行うことになるので湧き出し誤りが発生し，誤棄却

[1] 株式会社デンソーアイティーラボラトリ
Denso IT Laboratory, Shibuya, Tokyo 150–0002, Japan
[a)] ytachioka@d-itlab.co.jp

[*1] 本論文では音声認識の前処理として使われる音声区間検出技術を対象とする．

をすれば発話を取り逃すことになるため音声認識性能の低下に直結するためである．従来，パワーに基づく方法 [1] が長らく用いられてきた．この方法は音声のパワーが騒音のパワーよりも大きいと仮定しているが，高騒音環境ではこのような仮定は成り立たないため，尤度比検定に基づく方法 [2], [3] が代わりに使われている．これは周波数ごとにスペクトルの特性をガウス分布でモデル化している．モデルパラメータは観測ノイズだけから求められるため，事前の学習が要らないという利点がある．パワーという 1 次元特徴量を使う手法と比べて，多次元の特徴量を使うためより詳細なモデル化が可能で，音声の調波構造といった特有のスペクトルのパターンをとらえることができる．

一方で，事前に音声のモデルを学習しておくことで性能を向上させることができる．文献 [4] では，クリーン音声モデルを使うことの有効性が示されている．運用時には，騒音音声モデルを，クリーン音声から事前に学習したクリーン音声モデルと観測騒音から構築したノイズモデルをオンラインで合成することで音声区間検出を行う．

音声認識において深層神経回路網（deep neural network; DNN）の有効性が示されると [5]，すぐに，DNN の音声区間検出における有効性が示された [6], [7]．DNN はスペクトル由来の多次元特徴量を使い，騒音音声の学習データでモデルパラメータを学習する．DNN に基づく方法は，従来法に比べて 2 つの利点を有する．1 点目は様々な音声や騒音のパターンを表現できる柔軟性の高いモデリングである点である．2 点目は非線形関数を用いた次元圧縮により，従来法では扱えなかった高次元の特徴量を扱え，次元間の相関にもあまり配慮する必要がない点である．

音声認識の分野において，DNN の音響モデルに対して話者性や環境の特性を表す i-vector [8] のような特徴量を補助的に入力することで，音声認識の性能が向上することが知られている [9], [10]．DNN に基づく音声合成においても，話者コード等の補助特徴量を使うことで，合成音声の話者性を制御できることが示されている [11]．これらは補助特徴量を使うことで，DNN を環境に合わせて適応化しているととらえられる．

さらに，DNN に基づく音声強調の場合にも，スペクトル特徴に加えて，音声認識の結果から得られた音素の情報といった補助特徴量により音声強調の性能が向上する [12], [13]．音素情報は，音素の特性に応じて音声強調をかける程度を調整できるため有用である．すなわち，騒音と混ざりやすい子音のような音素は注意深く扱い，調波構造が明確で騒音と区別しやすい母音に関しては強く音声強調をかける等の調整ができる．

音声区間検出は音声と騒音を判別する問題だが，音声も騒音も多様性が大きく，摩擦音のように音声を区別しにくいものもあるため，これを直接解くのは難しい．補助特徴量により音声のパターンを限定することで，音声の多様性に制約をかけられる．これを異なる視点から見ると，[9], [10], [11] 等と同様，各フレームで事前に推定された音素に対して DNN を適応していると考えられる．このような適応手法は，学習と運用時で環境に差異がある場合にも有効に働く．音声区間検出においても，騒音の種類は多種多様で，学習の時点ですべて考慮することはできないため，モデルの適応が有効である．音声強調と同じく，音声区間検出においても，子音等騒音とまざりやすい音素に対しては慎重に判定を行い，母音のような騒音と区別しやすい音素に関しては，積極的に判定することが有効である．

本論文では，2 種類の補助音声モデルから得られる補助特徴量により，DNN に基づく音声区間検出の性能がどの程度向上するか実験的に検討する．モデルには，非負値行列因子分解（non-negative matrix factorization; NMF）[14], [15] のモデルと，音声認識に用いる音響モデルを使う．これらのモデルにより，前者からは NMF アクティベーションを，後者からは音響モデルの音響スコアを出力させ，補助特徴量とする．

車内環境での音声データ [16] を用いて，提案法の有効性を明らかにする．この実験には 3 つの目的がある．第 1 は，DNN に基づく音声区間検出法と従来法の性能比較を行うことである．これらの実環境での比較はあまり見られない．第 2 は，本論文の主たる目的で，補助特徴量の有効性を検証することである．第 3 は，音声区間検出の性能向上が，どの程度音声認識性能の向上につながるかを示すことにある．

2 章では，DNN に基づく音声区間検出が提案される前によくつかわれていた尤度比検定に基づく方法 [2] を概観する[*2]．同じく学習が必要な方法であるガウス混合モデル（Gaussian mixture model; GMM）に基づく音声区間検出は，DNN ベースのものに比べて明らかに性能が低いことが，文献 [7] によって示されているため比較しなかった．デコーダを利用した音声区間検出 [17] もあるが，計算量が多くなりすぎ現実的でない．音声区間検出はつねに動いていることから計算負荷が小さいことが求められる．音声認識の計算負荷はけっして小さくないので，つねにデコードすることは現実的でない．デコーダのモデルサイズを小さくすることで計算負荷を下げられるが，それでは音声認識性能が低下してしまうため，音声区間検出とデコードは 2 段階にした方がよい．実際，近年サーバ側で音声認識を行う状況が多くなっており，そのような場合も音声区間検出をクライアント側で行うことで通信量を少なくすることができる．また，すべてデコードすると湧き出し誤りが増加する可能性があるため，必ずしもすべてデコードするのがよいわけではない．3 章では，提案法のベースラインとなる DNN に基づく音声区間検出について述べる．4 章では，2

[*2] 文献 [6] でもベースラインとして使われている．

種類の補助特徴量を提案する．最後に 5 章において，騒音環境下で音声区間検出の実験を行う．

2. Sohn の方法

ここでは DNN 以前の従来法として一般的な，周波数ごとのスペクトル特徴を利用して音声区間検出を行う Sohn の方法 [2] を概観する．短時間フーリエ変換（short-time Fourier transform; STFT）により，観測音の F 次元 T フレームの STFT 係数 $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T] \in \mathbb{C}^{F\times T}$ の時刻 t $(1 \leq t \leq T)$ における $\boldsymbol{x}_t = [x_1, \ldots, x_f, \ldots, x_F]^\top \in \mathbb{C}^F$ を求める（$\top$ は転置で，時間のインデックスは省略した）．非音声区間 $\mathcal{H}_N$ と音声区間 $\mathcal{H}_S$ での音声と騒音の STFT 係数をそれぞれ $\boldsymbol{s}_t = [s_1, \ldots, s_f, \ldots, s_F]^\top$，$\boldsymbol{n}_t = [n_1, \ldots, n_f, \ldots, n_F]^\top$ とすると，観測音はそれぞれの区間で，

$$\mathcal{H}_N:\ \boldsymbol{x}_t = \boldsymbol{n}_t,\ \mathcal{H}_S:\ \boldsymbol{x}_t = \boldsymbol{n}_t + \boldsymbol{s}_t \tag{1}$$

のように表される．ここで $\mathcal{H}_N$，$\mathcal{H}_S$ において，それぞれの $\boldsymbol{x}_t$ の確率密度関数が，次式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．

$$\begin{aligned} p(\boldsymbol{x}_t|\mathcal{H}_N) &= \prod_{f=1}^{F} \frac{1}{\pi\lambda_f^N} e^{-\frac{|x_f|^2}{\lambda_f^N}} \\ p(\boldsymbol{x}_t|\mathcal{H}_S) &= \prod_{f=1}^{F} \frac{1}{\pi[\lambda_f^N + \lambda_f^S]} e^{-\frac{|x_f|^2}{[\lambda_f^N + \lambda_f^S]}} \end{aligned} \tag{2}$$

ここで λ_f^N，λ_f^S は n_f，s_f の分散を表す．すると f 次元目の音声・非音声の尤度比は，式 (3) で表される．

$$\Lambda_f(x_f) = \frac{p(x_f|\mathcal{H}_S)}{p(x_f|\mathcal{H}_N)} = \frac{1}{1+\xi_f} e^{\frac{\gamma_f \xi_f}{1+\xi_f}} \tag{3}$$

$$\xi_f = \lambda_f^S/\lambda_f^N,\ \gamma_f(x_f) = |x_f|^2/\lambda_f^N$$

ここで ξ_f，γ_f はそれぞれ事前，事後 SN（signal-to-noise）比と呼ばれる．それぞれの次元の尤度比の幾何平均により，音声・非音声を判断できる．

$$\log \Lambda(\boldsymbol{x}_t) = \frac{1}{F}\sum_{f=1}^{F} \log\left(\Lambda_f(x_f)\right) \underset{\mathcal{H}_N}{\overset{\mathcal{H}_S}{\gtrless}} \eta \tag{4}$$

$\log \Lambda(\boldsymbol{x}_t)$ が閾値 η よりも大きければ時刻 t は $\mathcal{H}_S$，小さければ $\mathcal{H}_N$ となる．ここで λ_f^N は観測された騒音の分散であり，事前に推定しておく．音声の分散 λ_f^S を最尤基準により推定すると，最終的に音声・非音声の判別式は式 (5) のようになる．

$$\log \Lambda^{(ML)}(\boldsymbol{x}_t) = \frac{1}{F}\sum_{f=1}^{F} \left(\gamma_f(x_f) - \log \gamma_f(x_f) - 1\right) \underset{\mathcal{H}_N}{\overset{\mathcal{H}_S}{\gtrless}} \eta \tag{5}$$

これに対して hidden Marikov model（HMM）hangover という手法により，判定を安定化させる手法が提案されている．HMM hangover では得られた Λ を直接使わず，前フレームの結果も用いて再帰的に計算した

$$\Gamma(t) = \frac{a_{01} + a_{11}\Gamma(t-1)}{a_{00} + a_{10}\Gamma(t-1)} \log \Lambda(\boldsymbol{x}_t) \tag{6}$$

を用いて，$\Gamma(t) \underset{\mathcal{H}_N}{\overset{\mathcal{H}_S}{\gtrless}} \eta$ により判定する．

3. DNN に基づく音声区間検出法

音声認識で DNN の有効性が示されるのとほぼ同時に，音声区間検出においても DNN の有効性が確認された [6], [7]．$\boldsymbol{X}$ より得られるパワースペクトル，フィルタバンク特徴量，Mel-frequency cepstral coefficient（MFCC）等の F' 次元のスペクトル特徴量 $\boldsymbol{X}' \in \mathbb{R}^{F'\times T}$ を入力として，たとえば 2 ノードの出力を設けて置き，学習データの音声 (0)/非音声 (1) の状態に対応して，一方のノードの出力が 1 になるように DNN を学習する．式 (7) のように，スペクトル特徴量 $\boldsymbol{X}'$ を DNN に入力し（変換を φ で表す），出力値 $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_T] \in \mathbb{R}^{2\times T}$ を得る．

$$\boldsymbol{Y} = \varphi(\boldsymbol{X}') \tag{7}$$

運用時にはそれらの出力のソフトマックスをとることで，音声の事後確率を算出し，それが閾値 η' を超えていれば音声区間，それ以外は非音声区間と判定できる．

$$\sigma(\boldsymbol{y}_t) = \frac{e^{y_t(0)}}{e^{y_t(0)} + e^{y_t(1)}} \underset{\mathcal{H}_N}{\overset{\mathcal{H}_S}{\gtrless}} \eta'. \tag{8}$$

4. 補助音声特徴量の利用

「はじめに」にも述べたとおり，音声区間検出においても，補助音声特徴量の利用が有効であると考えられる．音声区間検出を行う問題は，音声と騒音を区別する問題であるが，音声は多様性が大きい．そこで，補助特徴量として，音声のパターンを限定するような特徴量を用いれば，音声の多様性を縮小できる．たとえば，音素を表す特徴量を用いれば，先の音声と騒音を区別する問題が，ある特定の音素と騒音を区別する問題に単純化でき，音声区間検出の性能が向上することが期待される．また，話者性を表す特徴量を用いれば，特定の話者の音声を騒音の中から探す問題に単純化できる．**図 1** に提案のシステムを示す．音素や話者性の情報を表していると考えられる NMF によるアクティベーション，もしくは不特定話者の音響モデル（GMM/DNN）のスコアを補助特徴量として，DNN により音声区間検出を行う．

4.1 NMF アクティベーション

騒音が混ざった音声 $\boldsymbol{X}$ のパワースペクトルを NMF によって，騒音と音声に分離する．文献 [4] 同様に，クリーン音声から事前に音声の基底を構築しておくことで，音声

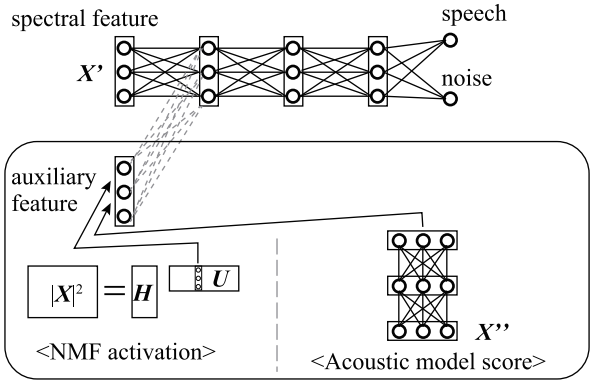

図 1 提案の補助音声モデルを併用した DNN に基づく音声区間検出システム

Fig. 1 The proposed DNN-based voice activity detection (VAD) system using auxiliary speech models.

と雑音を分離することができる．

$$|\boldsymbol{X}|^2 \simeq \hat{\boldsymbol{X}} = \boldsymbol{H}\boldsymbol{U} = [\boldsymbol{H}_s\boldsymbol{H}_n][\boldsymbol{U}_s;\boldsymbol{U}_n] = \boldsymbol{H}_s\boldsymbol{U}_s + \boldsymbol{H}_n\boldsymbol{U}_n \tag{9}$$

ここで $|\cdot|^2$ は行列の各要素の絶対値の 2 乗をとる関数，$\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{F\times K}$ は K 個の基底 $H_{f,k}$ からなる基底行列，$\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{K\times T}$ は，基底 k の時刻 t における活性化度 $U_{k,t}$ を表すアクティベーション行列である．基底 $\boldsymbol{H}$ をクリーン音声の基底 $\boldsymbol{H}_s$ と騒音の基底 $\boldsymbol{H}_n$ に分けると，アクティベーション $\boldsymbol{U}$ もそれぞれに対応して $\boldsymbol{U}_s$ と $\boldsymbol{U}_n$ に分けられる．ここでは，この $\boldsymbol{U}$ もしくは $\boldsymbol{U}_s$ に着目する．板倉斎藤擬距離に基づく乗法更新則 (10) により，$\boldsymbol{H}_n$ と $\boldsymbol{U}$ を更新する [18]．更新式は再構成された行列 $\hat{\boldsymbol{X}} = \boldsymbol{H}\boldsymbol{U}$ の各成分を $\hat{X}_{ft}$ とすると，

$$\begin{aligned} H_{f,k} &\leftarrow H_{f,k}\sqrt{\frac{\sum_t |X_{f,t}|^2 U_{k,t}/\hat{X}_{f,t}^2}{\sum_t U_{k,t}\hat{X}_{f,t}}} \\ U_{k,t} &\leftarrow U_{k,t}\sqrt{\frac{\sum_f |X_{f,t}|^2 H_{f,k}/\hat{X}_{f,t}^2}{\sum_f H_{f,k}\hat{X}_{f,t}}} \end{aligned} \tag{10}$$

のようになる．ただし $\boldsymbol{H}$ は $\boldsymbol{H}_n$ に対応する k のみ更新する．手順をまとめると $\boldsymbol{H}_s$ をクリーン音声からランダムに選択，$\boldsymbol{H}_n$ はランダム初期化し，$\boldsymbol{U}_s$，$\boldsymbol{U}_n$ と $\boldsymbol{H}_n$ を式 (10) により更新する．

NMF のアクティベーションは，音声の基底 $\boldsymbol{H}_s$ を音声の特徴を持つよう構築しておけば，発話に含まれる音声のアクティベーションを表すと考えられる．実際文献 [19] では，音声の基底に対応するアクティベーションを利用して，条件付き確率場で音声区間検出を行っている．また 5 章の実験でも，

$$\boldsymbol{Y} = \varphi([\boldsymbol{U}_s]),\ \boldsymbol{Y} = \varphi([\boldsymbol{U}]) \tag{11}$$

のように，DNN に基づく音声区間検出の特徴量としても有効であることを追試している．また 3 章のシステムに加えて，$\boldsymbol{U}$

$$\boldsymbol{Y} = \varphi([\boldsymbol{X}';\boldsymbol{U}]) \tag{12}$$

もしくは $\boldsymbol{U}_s$ を

$$\boldsymbol{Y} = \varphi([\boldsymbol{X}';\boldsymbol{U}_s]) \tag{13}$$

のように，補助特徴量として用いる．

4.2 音声認識の音響モデルの音響スコア

文献 [12], [13] では，音声認識の結果をフィードバックすることで，音声強調の性能を向上させる方法が提案されている．ただし「はじめに」に記したように，音声区間検出はデコードよりも負荷が小さい必要があり，デコードしたのでは負荷が大きすぎる．それを簡易的に考慮するために，小さい規模の音響モデルにより，音素ごとの音響スコアを算出し，それを特徴量/補助特徴量として利用することが考えられる．音響モデルによるスペクトル特徴量 $\boldsymbol{X}''$ の音素ごとのスコアへの変換を ψ とする．GMM 音響モデル，DNN 音響モデルのいずれも使えるため，両者で実験を行っている．音響スコアを正規化して[*3] DNN の入力とすると，式 (14) のようになる．

$$\boldsymbol{Y} = \varphi\left(\left[\boldsymbol{X}';\frac{\psi(\boldsymbol{X}'')}{|\psi(\boldsymbol{X}'')|}\right]\right) \tag{14}$$

4.3 結合型

結合型では上記の補助特徴量を結合した特徴量を使う．ただしそのままでは次元が高くなりすぎる懸念があるので，必要に応じて主成分分析（Principal Component Analysis; PCA）により次元削減行列 $\boldsymbol{P}$ をかける．$\boldsymbol{P}$ は学習データの特徴量から事前に求めておく．

$$\boldsymbol{Y} = \varphi\left(\left[\boldsymbol{X}';\boldsymbol{P}\left[\boldsymbol{U};\frac{\psi(\boldsymbol{X}'')}{|\psi(\boldsymbol{X}'')|}\right]\right]\right) \tag{15}$$

5. 騒音環境における音声区間検出実験

5.1 実験条件

車内で実収録された音声データセットである CENSREC-2 [16] を用いて[*4]，音声区間検出のためのデータセットを構築した．CENSREC-2 では発話ごとにファイルが切り出されているが，これを連結して 1 人あたり 1 分程度の音声データを作成し，音声区間検出の実験を行った．1 つの走行速度につき，話者数は学習セット 58 人，評価セット 15 人である[*5]．音声区間検出用 DNN の学習は，3 種の走行速

[*3] 必ずしも正規化する必要はないが，音響スコアは値のレンジが大きいため，何らかの配慮が必要となる．

[*4] 音声区間の実験を行うためのデータセットとして，CENSREC-1C があるが，サンプリング周波数が 8 kHz で実情と合わないため，データセットを構築した．

[*5] CENSREC-2 の評価セットには接話マイクの音声がなく音声区間のラベリングが困難であったので，CENSREC-2 の学習セットを分割して，新たに学習セットと評価セットを構築した．

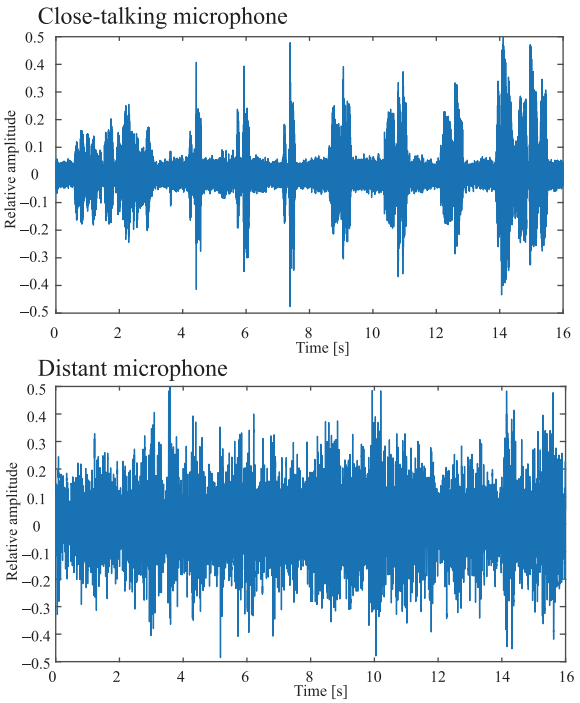

図 2 高騒音環境における近接マイクと遠隔マイクで収録された音声波形

Fig. 2 Speech waveforms recorded by close-talking and distant microphones in a highly noisy environment.

度（アイドリング（i_a），低速（市街地）走行（c_a），高速走行（e_a））すべての音声を用いて行った．各走行速度において，4 種類の車内環境（通常走行，エアコン On，オーディオ On，窓開）をほぼ同じ割合で組み合わせた 12 種類の環境が存在するが，集計は走行速度別に行った．発話は数字 11 種類（1〜9，0（まる），Z（ぜろ））から構成される．CENSREC-2 には，音声区間の時間ラベルが含まれていないため，ラベル付けは接話マイク収録された音声を自動音声認識して行った[*6]．**図 2** に，高速走行時の近接マイクと遠隔マイクにより収録された音声の比較を示す．近接マイクにより収録された音声は音声区間を視察で与えることもできそうだが，遠隔マイクの方は全体が騒音に埋もれており，視察では音声区間を特定することは難しい．

表 1 に実験の設定を示す．音響特徴量は，音声区間検出用の DNN と音響スコア計算用の DNN には，0 次から 22 次のフィルタバンク（filter bank; fb）特徴量を，前後 4 フレームコンテキスト拡張したものを用いた．NMF のアクティベーションはすべての基底に対する $\boldsymbol{U}$（式 (12)）と音声のみの基底に対応する $\boldsymbol{U}_s$（式 (13)）の 2 通りで実験した．NMF の基底 $\boldsymbol{H}_s$ は，15 時間の英語のクリーン音声（Wall street journal 0（WSJ0））からランダム選択により構築し，基底数は VAD の実験結果より決定した．$\boldsymbol{H}_n$ はランダム初期化し，$\boldsymbol{U}$ とともに乗法更新則により更新した．また GMM による音響スコアの計算には，0 次から 12 次までの MFCC 特徴量とその動的特徴量を用いた．音響スコア計算用の音響モデルは，基底 $\boldsymbol{H}_s$ を選択したのと同じ 15 時間の英語クリーン音声から学習した．音響モデルの状態数は英語のモノフォンに対応する 84 とした．DNN の隠れ層数やノード数は経験的に決定しており，最適化は行っていない．

[*6] 付属のスクリプトで「Condition 3」で音響モデルを適合学習し，そのモデルにより正解の文を音声認識した際のアライメントをとった結果の時間情報から，10 ms 周期のフレーム別で音声/非音声 2 値のラベル付けを行った．

表 1 音声区間検出システムの設定

Table 1 Setup for VAD system.

Acoustic features	
Sampling frequency	16 kHz
Window length/shift	25 ms/10 ms
Features	0–22th fbanks
Splice	9 frames
Model for VAD	
# DNN output nodes	2
# DNN nodes per layer	1,000 nodes
# DNN layers	3 layers
Model for auxiliary features	
# NMF bases	50 for each
# NMF iterations	100
# GMM/DNN output states	84 monophones
# mixtures of GMM	12
# DNN nodes per layer	200
# DNN layers	3

Sohn の方法では，λ_f^N は発話冒頭の騒音部分 10 フレームの分散とした．Sohn の方法のバリエーションとして，信号の振幅を最小 2 乗誤差（minimum mean square error; MMSE）基準で推定した decision directed（DD）手法もあり，主に騒音の部分で判定結果を安定させる効果がある [2]．DD 法では事前 SN 比 ξ の推定を MMSE-STSA（short-time spectral amplitude estimator）法 [20] により行う．よって MMSE-STSA 法により騒音抑圧した音声に対して音声区間検出を行うことで同様の効果が得られるため，これとの比較も行った．Sohn の方法はすべて HMM hangover 処理を行い，式 (6) のパラメータ $a_{00} = 1 - a_{01} = 0.8$，$a_{11} = 1 - a_{10} = 0.9$ とした．

音声区間検出の閾値は，Sohn の方法では，平均的に最も良い音声区間検出精度が得られるときの閾値 η を，環境に共通で与えた．DNN では，2 ノードの出力のソフトマックス値をとり，音声の事後確率が $\eta' = 0.5$ を超えた場合に音声，それ以外は騒音として判定した．

結合型の場合は，結合特徴量が高次元になりすぎる可能性があるため[*7]，両補助特徴量を単純に結合したものと，NMF アクティベーションを PCA により 400 次元に圧縮

[*7] fb 特徴量が 23 次元，NMF アクティベーションが 100 次元，GMM/DNN 音響モデルの事後確率が 83 次元であり，それぞれコンテキスト拡張するので，207，900，747 次元となる．

表 2 フレーム別の音声・非音声の判定精度 [%]．DNN の性能を Sohn の方法と比較した．DNN はフィルタバンク特徴量（fb）と NMF アクティベーションを使用

Table 2 Average frame-level VAD accuracy [%]. The performance of DNN was compared with that of Sohn's method. DNN used filterbank (fb) features with NMF activations.

	e_a	c_a	i_a
Sohn	52.08	63.34	63.23
Sohn（w MMSE-STSA）	62.25	60.81	65.06
DNN（fb）	77.76	86.03	91.62
DNN（speech activation $\boldsymbol{U}_s$）	76.00	84.36	90.00
DNN（all activation $\boldsymbol{U}$）	77.48	85.75	90.89
DNN（fb + speech activation $\boldsymbol{U}_s$）	79.37	**87.92**	**92.70**
DNN（fb + all activation $\boldsymbol{U}$）	**79.38**	87.79	92.64

したのちに特徴量結合したもの[*8]を比較した．

短すぎる発話を棄却し，発話中の短い非音声区間（無音や促音等）を音声区間として連結させるために，スムージング処理を行った．これは発話の最小継続長を 10 フレーム，発話中の 3 フレーム以下の無音区間は音声区間として連結させる処理で，音声区間検出ではたいていこのようなスムージング処理が行われる．

得られた音声区間に対して，CENSREC-2 付属のスクリプトにより学習した音響モデルにより，音声認識実験を行った．認識は数字単位の単語モデルで，GMM の混合数は各 20 である．

5.2 ベースライン

表 2 には，各手法により，フレームごとに音声・非音声を判定した結果の判定精度を示す．Sohn の方法のベースライン，MMSE-STSA 法により騒音抑圧した後に Sohn の方法を用いたものである．MMSE-STSA 法により，高騒音下（e_a）において Sohn の方法の性能が向上している．また，DNN（fb）が DNN に基づく手法の fb 特徴量でのベースラインであるが，Sohn の方法に比べてすべての条件で非常に高い性能を示している．

5.3 NMF アクティベーション

表 2 には，NMF の音声の基底に対応するアクティベーション $\boldsymbol{U}_s$ のみ（speech activation）と全基底に対するアクティベーション $\boldsymbol{U}$（all activation）を，入力特徴量として DNN により音声区間検出を行った結果（DNN（* activation））も示している．DNN（fb）よりは性能が若干低いものの，音声区間検出はできており，アクティベーションに音声区間検出に有用な情報が含まれていることが確認できた．

[*8] 式 (15) の PCA 行列 $\boldsymbol{P}$ の音響モデルスコアに対応する部分を対角行列としたことに相当する．音響スコアに対しては PCA をかけると性能が低下したため，NMF アクティベーションのみに PCA をかけた．

表 3 フレーム別の音声・非音声の判定精度 [%]．DNN により騒音音声を 11 種の数字と騒音に分類する

Table 3 Average frame-level VAD accuracy [%]. DNN classifies noisy speech into eleven digits and noise.

	e_a	c_a	i_a
DNN（fb）	78.00	86.53	91.79
DNN（fb + speech activation）	78.72	87.98	92.39

表 4 フレーム別の音声・非音声の判定精度 [%]．DNN は GMM/DNN クリーン音響モデルの音響スコアを併用

Table 4 Average frame-level VAD accuracy [%]. DNN additionally used GMM/DNN clean speech acoustic model scores.

	e_a	c_a	i_a
DNN（fb）	77.76	86.03	91.62
DNN（speech GMM）	78.41	87.30	92.15
DNN（speech DNN）	80.96	89.32	93.72
DNN（fb + speech GMM）	80.23	88.33	92.94
DNN（fb + speech DNN）	**81.70**	**90.14**	**94.28**

これに対して，補助特徴量として fb 特徴量に加えてアクティベーション（$\boldsymbol{U}$，$\boldsymbol{U}_s$）を用いることで，補助特徴量を用いないものに比べて，どちらの場合も性能が向上した．すべてのアクティベーションを用いたもの $\boldsymbol{U}$ と音声のアクティベーションだけを用いたもの $\boldsymbol{U}_s$ の性能の差異が小さかったことから，騒音の基底に対するアクティベーション $\boldsymbol{U}_n$ を用いても，精度は向上しないことが分かった．このことから，音声の基底に対するアクティベーション $\boldsymbol{U}_s$ を用いる有効性が示された．

表 3 は，DNN の出力を音声・非音声の 2 分類ではなく，11 種類の数字と騒音の 12 分類に細かくモデリングした場合である．若干の推定精度の向上が見られるものもあったが，認識対象が変わると使えないという欠点があるため，以後は上記同様音声・非音声 2 分類の場合で評価する．

5.4 音響スコア

表 4 には，音響モデル（GMM/DNN）の音響スコアを特徴量/補助特徴量とした場合の結果を示す．5.3 節に示した NMF アクティベーションを用いた結果に比べ全体的に精度が高く，GMM よりも DNN 音響モデルを用いた場合の方が有効性が高いことが分かった．また音響スコアを補助特徴量として用いた方が，単独で用いるよりも性能が高いことが分かった．

5.5 尤度比，事後確率の比較

図 3 には，図 2 の音声を与えたときの，式 (5) で計算される Sohn の方法による対数尤度比 $\log\Lambda$ と，提案の DNN の音声区間検出モデルに DNN の音響モデルの音響スコア

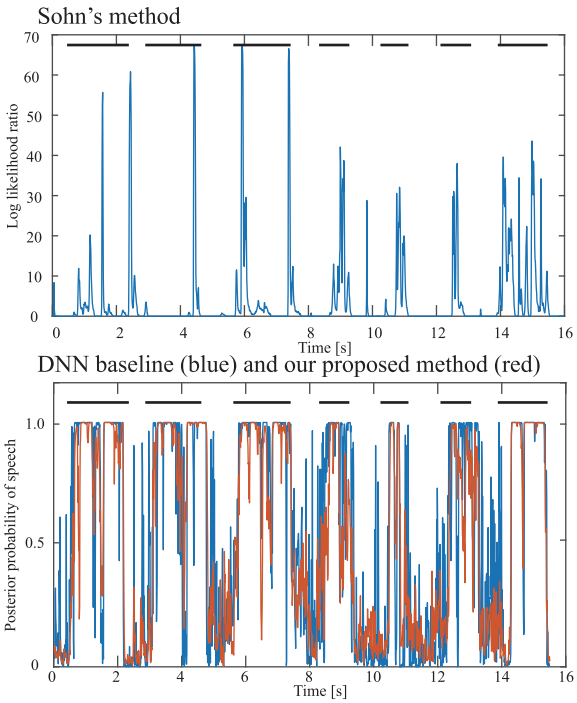

図 3 Sohn の方法での対数尤度比 $\log\Lambda$ と，DNN ベースライン（青線）と提案法（赤線）での音声の事後確率 σ．上の線が正解の音声区間

Fig. 3 Log likelihood ratio of Sohn's method, $\log\Lambda$, and the speech posterior probability σ of DNN baseline and the proposed method. Upper line indicates the reference speech area.

を補助特徴量として与えた際に算出された音声の事後確率（式 (8)）を示す．横線は正解の発話区間を表す．どちらの方法も発話を取り逃してはいないものの，Sohn の方法の結果が非常に変動が大きく，始終端検出の性能は，閾値 η による影響を大きく受けることが分かる．これに対して DNN の結果は変動も少なく，閾値処理も容易である．青線の音響特徴量のみを用いた DNN よりも，提案の補助特徴量を併用した DNN の方がより，発話区間においては値が安定し，無音区間においては誤検出を抑制できている．これにより，音声と騒音を識別する性能が向上しているといえる．

5.6 結合型

表 5 には結合型の音声区間検出性能を示す．全体的に音声の基底に対応するアクティベーションのみを使ったものの方がよく，DNN のスコアを用いた方が，GMM のスコアと組み合わせたものよりも判定精度が高いことが分かる．ただし表 4 の結果と比べて性能の向上はほとんど見られなかった．PCA の有無はあまり性能に影響しておらず，PCA を行わなくても高次元の特徴量を DNN が問題なく扱うことができていることが分かる．

表 5 フレーム別の音声・非音声の判定精度 [%]．DNN は両補助特徴量を使用

Table 5 Average frame-level VAD accuracy [%]. DNN used both auxiliary features.

activation	GMM/DNN	PCA	e_a	c_a	i_a
speech	GMM	-	80.18	88.87	93.42
all	GMM	-	79.97	88.24	93.19
speech	DNN	-	81.51	**90.08**	94.11
all	DNN	-	81.49	89.91	94.00
speech	GMM	✓	80.01	88.47	93.30
all	GMM	✓	79.75	87.64	92.53
speech	DNN	✓	**81.66**	89.88	**94.38**
all	DNN	✓	81.44	89.51	94.10

表 6 スムージングしたフレーム別の音声・非音声の判定精度 [%]

Table 6 Smoothed average frame-level VAD accuracy [%].

	e_a	c_a	i_a
Sohn	52.14	63.66	63.46
Sohn（w MMSE-STSA）	62.41	60.81	65.22
DNN（fb）	80.91	89.02	94.03
DNN（speech activation）	78.62	87.21	91.68
DNN（all activation）	80.14	88.47	92.76
DNN（speech GMM）	79.86	89.69	93.55
DNN（speech DNN）	82.62	90.65	94.58
DNN（fb + auxiliary features）			
activation GMM/DNN			
speech -	81.90	89.93	94.02
all -	82.13	90.25	94.39
- GMM	82.25	88.33	92.94
- DNN	**82.68**	**91.19**	**94.91**

5.7 スムージングの必要性

表 6 には，フレーム別の音声区間検出結果を，スムージング処理した場合を示す．DNN に基づく手法の場合に，顕著に性能が向上している．Sohn の方法では HMM hangover によるスムージング効果がすでに入っているが，DNN は入力特徴量の隣接コンテキストを利用することで暗に与えているだけなので，性能が向上した．

結合型に関しては，スムージングなしではあまり性能が向上しなかったものの，**表 7** に示すとおり，スムージングをすることで，特に騒音の大きな環境（e_a）で性能が大幅に改善した．2 つの異なる特徴量を使うことで，それらが相補的に働き，より安定的に音声区間が検出できるようになったと考えられる．

5.8 提案法の効果に関する考察

図 4 に，話者別の性能改善量［pt（ポイント）］のヒストグラムを示す．speech + DNN の場合の判定精度から，ベースライン（fb）の判定精度を引いたものである．このように話者・環境を問わず 39/45 条件で性能が向上しており，提案法により，話者・環境に適応できていることが分

表 7 スムージングしたフレーム別の音声・非音声の判定精度 [%]. DNN は両補助特徴量を使用

Table 7 Smoothed average frame-level VAD accuracy [%]. DNN used both auxiliary features.

activation	GMM/DNN	PCA	e_a	c_a	i_a
speech	GMM	-	81.87	90.77	94.26
all	GMM	-	82.12	90.16	94.28
speech	DNN	-	82.80	**91.14**	94.72
all	DNN	-	82.86	91.05	94.74
speech	GMM	✓	82.08	90.81	94.54
all	GMM	✓	82.32	89.96	94.26
speech	DNN	✓	**83.40**	**91.14**	**95.22**
all	DNN	✓	83.16	91.01	94.91

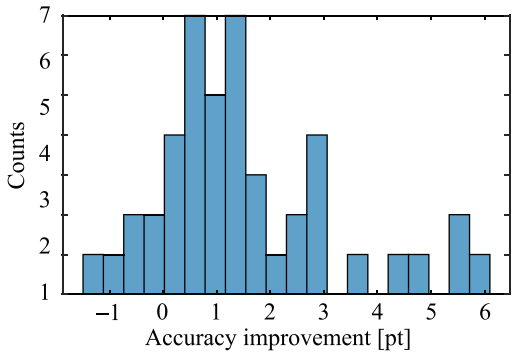

図 4 条件別の性能改善量 [pt] のヒストグラム（15 speakers and 4 environments）

Fig. 4 Histogram of accuracy improvement [pt] per condition (15 speakers and 4 environments).

かる．

5.9 音声認識での評価

表 8 に単語正解精度（word accuracy）を示す．5.7 節のスムージング処理を行ったのちに発話単位で認識を行い，その認識結果を結合したうえで評価した．音声区間検出なしですべてデコードした場合の結果（VAD none）もあわせて示す．これは計算量がかなり多くなるのにもかかわらず，e_a で音声強調のない Sohn の方法を上回るが，それ以外はほぼ変わらないか低い性能となり，音声強調を入れた Sohn の方法，もしくは DNN に基づく方法のすべての場合を下回った．こちらも DNN に基づく音声区間検出を行った場合の性能が高く，補助特徴量の利用によりさらに性能が向上している．音声区間検出でのとりこぼしは音声認識性能の低下に直結するため，高精度に音声区間検出を行う重要性が示された．また結合型により全体的に性能が向上し，特に高騒音下で大幅に改善された．

6. まとめ

DNN による音声区間検出の性能を向上させるために，補助特徴量を併用する方法を提案した．CENSREC-2 による音声区間検出実験を行ったところ，従来のパワーに基づく方法よりも，DNN に基づく方法の性能が顕著に高いことが分かった．また，NMF のアクティベーションや GMM/DNN 音響モデルのスコアを補助特徴量とした実験により，補助特徴量の利用が有効であることを確認した．加えて音声認識の性能も向上させることを確認した．また 2 つの補助特徴量を併用することでさらに性能が改善し，特に音声認識性能の大幅な改善が見られた．

表 8 検出された音声の音声認識での単語正解精度 [%]

Table 8 Word accuracy [%] of automatic speech recognition for the detected speech.

			e_a	c_a	i_a
Baseline					
VAD none			41.21	39.83	41.81
Sohn			18.37	40.79	44.70
Sohn（w MMSE-STSA）			50.93	59.38	68.67
DNN（fb）			69.33	78.30	87.25
DNN（all activation）			67.76	73.99	82.71
DNN（speech DNN）			73.82	79.09	88.66
DNN（fb + auxiliary features）					
activation	GMM/DNN	PCA			
speech	-	-	72.70	78.87	87.37
all	-	-	73.00	79.15	87.06
-	GMM	-	73.75	80.57	88.35
-	DNN	-	72.70	80.28	89.03
speech	GMM	✓	73.15	80.11	88.72
all	GMM	✓	74.79	81.19	89.21
speech	DNN	✓	**76.51**	80.85	**89.88**
all	DNN	✓	74.57	**81.98**	89.27

参考文献

[1] Rabiner, L. and Sambur, M.: An Algorithm for Determining the Endpoints of Isolated Utterances, *The Bell System Technical Journal*, Vol.54, pp.297–315 (1975).
[2] Sohn, J., Kim, N.S. and Sung, W.: A Statistical Model-based Voice Activity Detection, *IEEE Signal Processing Letters*, Vol.6, pp.1–3 (1999).
[3] 太刀岡勇気，花沢利行，成田知宏，石井 純：音声と騒音の密度比推定を用いた音声区間検出法，電気学会論文誌 C，Vol.133, pp.1549–1555 (2013).
[4] Fujimoto, M. and Ishizuka, K.: Noise Robust Voice Activity Detection Based on Switching Kalman Filter, *IEICE Trans. Information and Systems*, Vol.E91-D, pp.467–477 (2008).
[5] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Processing Magazine*, Vol.28, pp.82–97 (2012).
[6] Zhang, X.-L. and Wu, J.: Deep Belief Networks Based Voice Activity Detection, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.21, No.4, pp.1–14 (2013).
[7] Hughes, T. and Mierle, K.: Recurrent Neural Networks for Voice Activity Detection, *Proc. ICASSP*, pp.7378–7382 (2013).
[8] Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-end Factor Analysis for Speaker Verifi-

cation, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.19, No.4, pp.788–798 (2011).
[9] Delcroix, M., Kinoshita, K., Hori, T. and Nakatani, T.: Context Adaptive Deep Neural Networks for Fast Acoustic Model Adaptation, *Proc. ICASSP*, pp.4535–4539 (2015).
[10] Tran, D., Delcroix, M., Ogawa, A., Hümmer, C. and Nakatani, T.: Feedback Connection for Deep Neural Network-Based Acoustic Modeling, *Proc. ICASSP* (2017).
[11] Hojo, N., Ijima, Y. and Mizuno, H.: An Investigation of DNN-Based Speech Synthesis Using Speaker Codes, *Proc. INTERSPEECH*, pp.2278–2282 (2016).
[12] Sohrab, F. and Erdogan, H.: Recognize and Separate Approach for Speech Denoising Using Nonnegative Matrix Factorization, *Proc. EUSIPCO* (2015).
[13] Kinoshita, K., Delcroix, M., Ogawa, A. and Nakatani, T.: Text-informed Speech Enhancement with Deep Neural Networks, *Proc. INTERSPEECH*, pp.1760–1764 (2015).
[14] Lee, D.D. and Seung, S.: Learning the Parts of Objects by Non-negative Matrix Factorization, *Nature*, Vol.401, No.6755, pp.788–791 (1999).
[15] Smaragdis, P.: Convolutive Speech Bases and Their Application to Supervised Speech Separation, *IEEE Trans. Audio, Speech and Language Processing*, Vol.15, No.1, pp.1–12 (2007).
[16] Takeda, K., Fujimura, H., Itou, K., Kawaguchi, N., Matsubara, S. and Itakura, F.: Construction and Evaluation a Large In-Car Speech Corpus, *IEICE Trans. Information and Systems*, Vol.E88-D, No.3, pp.553–561 (2005).
[17] Ohnish, T., Dixon, P., Iwano, K. and Furui, S.: Robust Speech Recognition Using VAD-measure Embedded Decoder, *Proc. INTERSPEECH*, pp.2239–2242 (2009).
[18] Nakano, M., Kameoka, H., Le Roux, J., Kitano, Y., Ono, N. and Sagayama, S.: Convergence-guaranteed Multiplicative Algorithms for Non-negative Matrix Factorization with Beta-divergence, *Proc. MLSP*, pp.283–288 (2010).
[19] Teng, P. and Jia, Y.: Voice Activity Detection via Noise Reducing Using Non-Negative Sparse Coding, *IEEE Signal Processing Letters*, Vol.20, No.5, pp.475–478 (2013).
[20] Ephraim, Y. and Malah, D.: 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).

太刀岡 勇気 （正会員）

1983 年生．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．同年三菱電機株式会社入社．2017 年退社．同年株式会社デンソーアイティーラボラトリ入社．音声認識の研究に従事．2018 年東京工業大学工学院（情報通信系）博士課程修了．博士（工学）．日本音響学会，ISCA，計量国語学会各会員．