補助音声モデルを用いたDNNによる音声区間検出法*

○太刀岡 勇気 (デンソーアイティーラボラトリ)

1 はじめに

遠隔マイクでの音声操作や発話スイッチレス音声認識など実環境で音声インターフェースを利用する機会の増加に伴い、騒音環境下で対象話者が発話した区間を検出する音声区間検出技術が、重要性を増している。以前はパワーに基づく方法が使われていたが、次第に高騒音下で検出力を向上させるため、尤度比検定に基づく方法 [1, 2] が主流となった。この他に、音声モデルを用いる方法 [3] の有効性も知られている。近年では、Deep Neural Network (DNN) を使うことで、さらに性能が向上することが示されている [4]。

一方、DNN に基づく音声強調の際に、音声認識の情報などを補助的な特徴量として使う方法が提案されている [5, 6]。本報では、DNN 音声区間検出において、スペクトル特徴量に加えて、補助音声モデルより出力される補助特徴量を併用することで、音声区間検出の精度を向上させる方法を提案する。補助音声モデルとしては、非負値行列因子分解 (NMF) と音声認識用の音響モデルを利用する。前者では、NMF のアクティベーションを、後者では、音響モデルの音響スコアを補助特徴量として用いる。CENSREC-2 を用いた実験により、提案法の有効性を示す。

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 [1] を概観する。短時間フーリエ変換により、観測音の FFT 係数 $\boldsymbol{X} \in \mathbb{C}^{F\times T}$ の時刻 t における特徴量 $\boldsymbol{X}_t = \{X_{f=1,\ldots,F}\} \in \mathbb{C}^F$ 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 FFT 係数をそれぞれ $\boldsymbol{S}_t = \{S_{f=1,\ldots,F}\}$, $\boldsymbol{N}_t = \{N_{f=1,\ldots,F}\}$ とすると、観測音はそれぞれの区間で、$H_N:\ \boldsymbol{X}_t = \boldsymbol{N}_t$、$H_S:\ \boldsymbol{X}_t = \boldsymbol{N}_t + \boldsymbol{S}_t$ のように表される。ここで H_N、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。

$$
\begin{aligned}
p(\boldsymbol{X}_t|H_N) &= \prod_{f=1}^{F} \frac{1}{\pi\lambda_f^N} e^{-\frac{|X_f|^2}{\lambda_f^N}} \\
p(\boldsymbol{X}_t|H_S) &= \prod_{f=1}^{F} \frac{1}{\pi[\lambda_f^N + \lambda_f^S]} e^{-\frac{|X_f|^2}{[\lambda_f^N + \lambda_f^S]}}
\end{aligned}
\tag{1}
$$

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 (2) で表される。

$$
\Lambda_f(X_f) = \frac{p(X_f|H_S)}{p(X_f|H_N)} = \frac{1}{1+\xi_f} e^{\frac{\gamma_f \xi_f}{1+\xi_f}} \tag{2}
$$

$$
\xi_f = \lambda_f^S/\lambda_f^N,\ \gamma_f = |X_f|^2/\lambda_f^N
$$

ここで ξ_f, γ_f はそれぞれ事前、事後 SN 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。

$$
\log\Lambda(\boldsymbol{X}_t) = \frac{1}{F}\sum_{f=1}^{F} \log\left(\Lambda_f(X_f)\right) \underset{H_N}{\overset{H_S}{\gtrless}} \eta \tag{3}
$$

$\log\Lambda(\boldsymbol{X}_t)$ が閾値 η よりも大きければ時刻 t は H_S、小さければ H_N となる。ここで λ_f^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。音声モデル 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 (4) のようになる。

$$
\log\Lambda^{(ML)}(\boldsymbol{X}_t) = \frac{1}{F}\sum_{f=1}^{F} (\gamma_f - \log\gamma_f - 1) \tag{4}
$$

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 [4]。$\boldsymbol{X}$ より得られるスペクトル特徴量を入力として、例えば 2 ノードの出力を設けて置き、学習データの音声/非音声の状態に対応して、一方のノードの出力が 1 になるように DNN を学習する。テスト時にはそれらの出力の softmax をとることで、音声の事後確率を算出することができる。式 (5) のように、スペクトル特徴量 $\boldsymbol{X'} \in \mathbb{R}^{F'\times T}$ を DNN に入力し (変換を f で表す)、出力値 $\boldsymbol{y} \in \mathbb{R}^{2\times T}$ を得る。

$$
\boldsymbol{y} = f(\boldsymbol{X'}) \tag{5}
$$

4 補助特徴量の利用

音声強調の分野において、スペクトル特徴に加えて補助的な特徴量を用いることで、音声強調の性能が向上することが示されている [5, 6]。また音声認識の分野においても、DNN の音響モデルに対して話者性を表す特徴量を補助的に入力することで、音声認識の

*DNN-based voice activity detection using auxiliary speech models. by TACHIOKA, Yuuki (Denso IT Laboratory)

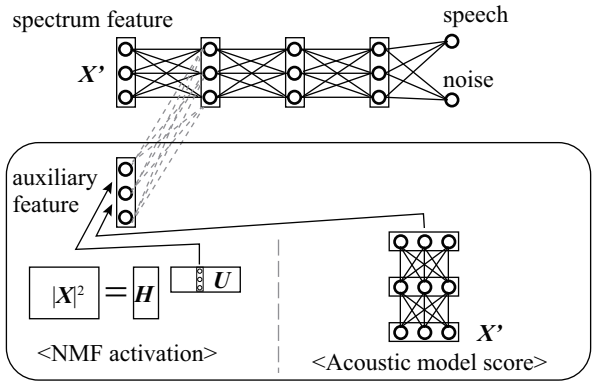

Fig. 1 The proposed DNN-based voice activity detection (VAD) system using auxiliary speech models.

性能が向上することが知られている [7]。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ることができる。音声区間検出においても、このような手法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くこの問題を直接解くことは難しい。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。例えば、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。図 1 に提案のシステムを示す。NMF によるアクティベーション、もしくは音響モデルのスコアを補助特徴量として用いて、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。

$$|\boldsymbol{X}|^2 \simeq \boldsymbol{H}\boldsymbol{U} = \boldsymbol{H}_s\boldsymbol{U}_s + \boldsymbol{H}_n\boldsymbol{U}_n \qquad (6)$$

ここで $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{F\times K}$ は K 個の基底からなる基底行列、$\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{K\times T}$ 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。基底を音声の基底 $\boldsymbol{H}_s$ と騒音の基底 $\boldsymbol{H}_n$ に分けると、それぞれのアクティベーションも $\boldsymbol{U}_s$ 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。NMF のアクティベーションは基底 $\boldsymbol{H}$ が適切なものであるならば、発話に含まれる音声の特徴をよく表していると考えられる。実際 [8] 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。そこで $\boldsymbol{U}$ もしくは $\boldsymbol{U}_s$ を

$$\boldsymbol{y} = f([\boldsymbol{X}';\boldsymbol{U}]) \text{ or } \boldsymbol{y} = f([\boldsymbol{X}';\boldsymbol{U}_s]) \qquad (7)$$

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 [5, 6] 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。ここにおいても、音声認識に用いる音響モデルにより、音素毎に属する確率を算出し、それを補助特徴量として用いる方法も考えられる。音響モデルによるスペクトル特徴量 $\boldsymbol{X}'$[1]。の音素毎の事後確率への変換を g とする[2]。音響モデルのスコアを正規化して[3]DNN の入力とすると、式 (8) のようになる

$$\boldsymbol{y} = f\left(\left[\boldsymbol{X}'; \frac{g(\boldsymbol{X}')}{|g(\boldsymbol{X}')|}\right]\right) \qquad (8)$$

5 CENSREC-2 による実験

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 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 [9]。CENSREC-2 では発話毎にファイルが切り出されているが、これを連結して一人当たり 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。一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]。音声区間検出用の DNN の学習は、3 種の走行速度 (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。付属のス

Table 1 Setup for the VAD system.

Sampling frequency	16 kHz
Window length	25 ms
Window shift	10 ms
Features	0–22th fbanks
Splice	9 frames
# NMF bases	50
# DNN output nodes	2
# DNN nodes per layer	1,000 nodes
DNN layer size	3 layers

[1]式 (5) でのスペクトル特徴量と同じである必要はない。
[2]GMM 音響モデル、DNN 音響モデルのいずれも使える。
[3]必ずしも正規化する必要はないが、値のレンジの大きなスコアを扱う場合は何らかの配慮が必要となる。
[4]音声区間の実験をおこな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 8kHz で実情と合わないため、独自の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。

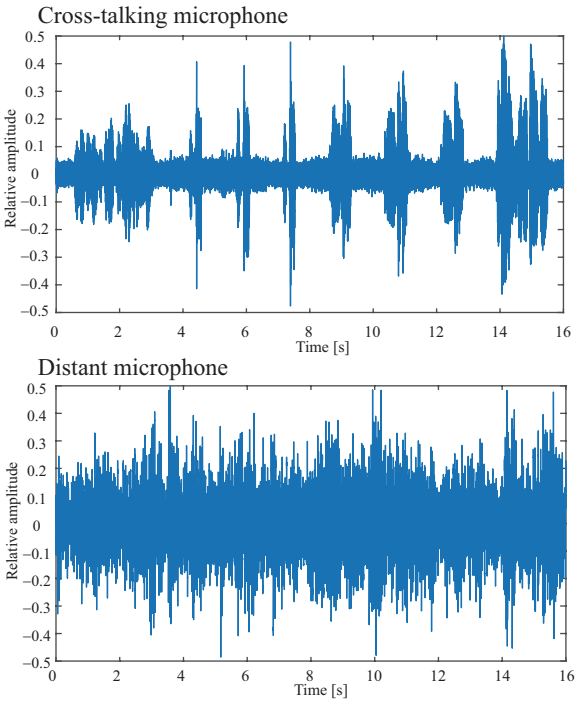

Fig. 2 Waveforms recorded by cross-talking and distant microphones in a highly noisy environment.

クリプトで「Condition 3」で音響モデルを適合学習し、そのモデルにより音声認識した際のアライメント結果の時間情報から、10ms 周期のフレーム単位で音声/非音声 2 値のラベル付けを行った。

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 次のフィルターバンク (fbank) 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。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。NMF のアクティベーションは音声の基底に対応する $\boldsymbol{U}_s$ と、すべての基底に対する $\boldsymbol{U}$ の両方で実験した。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を、環境に共通で与えた。DNN の方法では、2 ノードの出力の softmax 値を取り、音声の事後確率が 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。

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。

5.2 ベースライン

表 2 には、フレーム単位での平均音声区間検出精度を示す。Sohn の方法のベースライン、MMSE-STSA により騒音抑圧した後に Sohn の方法を用いたもの、多層パーセプトロン (MLP) によるベースラインである。MLP が DNN に基づく手法のベースラインであるが、Sohn の方法に比べて非常に高い性能を示している。

Table 2 Average frame-level VAD accuracy [%]. The performance of MLP was compared with that of the conventional Sohn's method. MLP used filterbank 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
MLP baseline	77.76	86.03	91.62
+ speech activation	79.37	**87.92**	**92.70**
+ speech & noise activation	**79.38**	87.79	92.64

Table 3 Average frame-level VAD accuracy [%]. MLP used filterbank features with clean speech acoustic model (GMM/DNN) outputs.

	e_a	c_a	i_a
MLP baseline	77.76	86.03	91.62
+ speech GMM	80.23	88.33	92.94
+ speech DNN	**81.70**	**90.14**	**94.28**

5.3 NMF アクティベーション

表 2 に、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 $\boldsymbol{U}$(speech & noise activation) を、補助特徴量として加えた結果を合わせて示している。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が、騒音の基底に対するアクティベーションを用いても、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の有効性が示された。

5.4 音響スコア

表 3 には、音響モデル (GMM/DNN) の音響スコアを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。

図 3 には、図 2 の音声を与えたときの、式 (4) で計算される Sohn の方法の対数尤度比 $\log\Lambda$ 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を示す。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わかる。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。

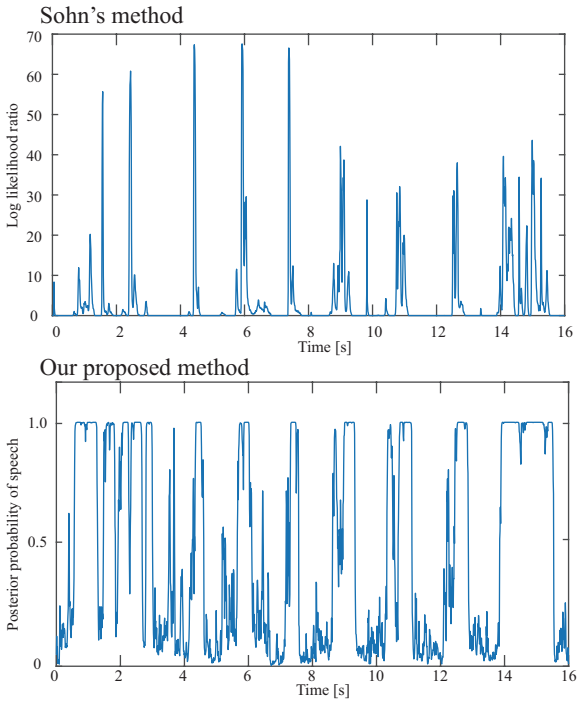

Fig. 3　Log likelihood ratio of Sohn's method, $\log \Lambda$, and the speech posterior probability of the proposed method.

Table 4　Average frame-level VAD accuracy [%] with smoothing.

	e_a	c_a	i_a
Sohn	52.14	63.66	63.46
MLP baseline	80.91	89.02	94.03
+ speech activation	81.90	89.93	94.02
+ speech & noise activation	82.13	90.25	94.39
+ speech GMM	82.25	88.33	92.94
+ speech DNN	**82.68**	**91.19**	**94.91**

5.5　スムージングの必要性

表 4 には、フレーム単位の音声区間検出結果に対して、隣接数フレームをまたいでスムージングした場合の性能を示す。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 といったスムージング手法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。

5.6　音声認識での評価

CENSREC-2 付属のスクリプトにより学習した GMM 音響モデルにより、音声認識実験を行った。表 5 に単語正解率を示す。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。音声区間検出での取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。

Table 5　Word accuracy [%] of automatic speech recognition for the detected speech.

	e_a	c_a	i_a
Sohn	18.37	40.79	44.70
MLP baseline	69.33	78.30	87.25
+ speech activation	72.70	78.87	87.37
+ speech & noise activation	73.00	79.15	87.06
+ speech GMM	**73.75**	**80.57**	88.35
+ speech NN	72.70	80.28	**89.03**

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、補助音声モデルによる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。加えて音声認識実験においても、提案法の有効性を確認した。

参考文献

[1] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, **6**, 1–3 (1999).

[2] 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，**133**，1549–1555 (2013).

[3] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," IEICE Transactions on Information and Systems, **E91-D**, 467–477 (2008).

[4] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," IEEE Transactions on Audio, Speech, and Language Processing, **21**, 1–14 (2013).

[5] F. Sohrab and H. Erdogan, "Recognize and separate approach for speech denoising using nonnegative matrix factorization," Proceedings of EUSIPCO (2015).

[6] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," Proceedings of INTERSPEECH, pp.1760–1764 (2015).

[7] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," Proceedings of ICASSP, pp.4535–4539 (2015).

[8] P. Teng and Y. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," IEEE Signal Processing Letters, **20**, 475–478 (2013).

[9] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and F. Itakura, "Construction and evaluation a large in-car speech corpus," IEICE Transactions on Information and Systems, **E88-D**, 553–561 (2005).