Vol.2016-SLP-112 No.9
2016/7/29

音素エントロピーを利用した背景発話に頑健なDNNに基づく音声区間検出

藤田 悠哉[1,a)]　磯 健一[1]

概要：DNN に基づく音声区間検出に音素エントロピーによる棄却を追加することで背景発話の誤検出を抑制する方法を提案する．我々が運用している音声認識サービスでは, DNN に基づく音声区間検出を採用している．音声区間検出の誤りを観察したところ, そのほとんどが TV またはラジオや周囲の人の会話に由来する背景発話の誤検出だった．本稿ではそのような誤検出を抑制するために, DNN 音響モデルの音素事後確率のエントロピーに基づく信頼度スコアを導入する．背景発話はユーザーが音声認識サービスの利用を意図して行う発話よりもマイクロフォンとの距離が遠いことが多く, ノイズや残響の影響を受けやすい．従って背景発話音声は音素事後確率のエントロピーが大きな値を持つと考えられる．そこで, DNN に基づく音声区間検出により音声と判定されたフレームのうち, 音素事後確率のエントロピーが閾値以上のフレームを棄却し, 背景発話による誤検出を抑制する．実験により, 音声認識サービスの文誤り率が 10%以上削減できることを確認した．

Robust DNN-based VAD augmented with phone entropy based rejection of background speech

Yuya Fujita[1,a)]　Ken-ichi Iso[1]

Abstract: We propose a DNN-based voice activity detector augmented by entropy based frame rejection. DNN-based VAD classifies a frame into speech or non-speech and achieves significantly higher VAD performance compared to conventional statistical model-based VAD. We observed that many of the remaining errors are false alarms caused by background human speech, such as TV / radio or surrounding peoples' conversations. In order to reject such background speech frames, we introduce an entropy-based confidence measure using the phone posterior probability output by a DNN-based acoustic model. Compared to the target speaker's voice background speech tends to have relatively unclear pronunciation or is contaminated by other types of noises so its entropy becomes larger than audio signals with only the target speaker's voice. Combining DNN-based VAD and the entropy criterion, we reject speech frames classified by the DNN-based VAD as having an entropy larger than a threshold value. We have evaluated the proposed approach and confirmed greater than 10% reduction in Sentence Error Rate.

1. はじめに

音声区間検出 (Voice Activity Detection, VAD) は音声認識システムの重要な前段処理の 1 つである．システムに入力される音響信号をテキスト化不要な背景雑音のみの非音声区間とテキスト化したい音声を含む音声区間に分割することで, 認識精度の向上と計算量が削減できるからである．また, 品質の良いハンズフリーの無線通信や音声圧縮コーデックの実現にも欠かせない技術である．これまでに様々な手法が提案されているが, 大別すると次のように分類できる．まず最初は零交差頻度やエネルギーといった低レベルの音響特徴量を用いる手法である [1], [2]．次は確率モデルを用いる手法である．音声と非音声のフレームの特徴量がそれぞれ正規分布でモデル化され, それらの尤度比

[1] ヤフー株式会社
Yahoo Japan Corporation
Kioi Tower, Tokyo Garden Terrace Kioicho, 1-3 Kioicho, Chiyoda-ku, Tokyo, 102-8282
[a)] yuyfujit@yahoo-corp.jp

を用いてあるフレームが音声か非音声かを判定する．3 つ目は識別器を用いる手法である．サポートベクターマシン (Support Vector Machine, SVM) は機械学習において最もよく用いられる識別器の 1 つであり, VAD にも適用されている [3]．4 つ目はカルマンフィルタや隠れマルコフモデル (Hidden Markov Model, HMM) といった状態空間モデルを用いた手法である [4], [5]．最後に, 深層学習 (Deep Neural Network, DNN) を用いた手法が, 近年の音響モデルにおける成功を背景に数多く提案されている [6], [7], [8], [9], [10]．

本稿では, 我々が開発して運用している音声認識サービス (スマートフォン等から音声で web 等の検索を行うシステム) で採用している DNN に基づく VAD の改善について報告する．先述のような数多くの VAD 手法の中から DNN に基づく VAD を採用する理由の 1 つは, 音響モデル用の DNN のために作成した学習データやソースコードが再利用できるため, 実装や運用が容易になるからである．DNN に基づく VAD は確率モデルを用いる VAD に比べて大幅に性能が向上するが, まだ VAD の失敗により音声認識性能が低下する状況がある．誤認識された発話を分析したところ, そのような状況の 1 つとして周囲の人の会話やテレビ, ラジオのスピーカーに由来する背景発話による誤検出が多いことがわかった．我々のシステムに入力される発話は利用されるスマートフォン向けアプリケーションの種類によって大まかに 3 つのドメインに分類できる．1 つ目は典型的な音声検索アプリケーション (Search), そして 2 つ目はパーソナルアシスタントアプリケーション (Dialogue), 3 つ目は地図またはカーナビゲーションアプリケーション (Vehicle) である．分析の結果, 背景発話による誤検出が最も多いのは Vehicle ドメインだったことから, このドメインにおける背景発話を棄却する手法を検討する．

本稿で提案する手法は, 音響モデルの DNN が出力する事後確率のエントロピーを利用する．先述の通り, 誤検出する背景発話は周囲の人の会話やテレビ, ラジオのスピーカに由来する．そのような背景発話は音声認識サービスへの入力を意図してユーザーが発話する音声に比べてマイクロフォンから遠い位置にあることが多く, 周囲のノイズや残響の影響を受けやすい．ノイズや背景発話のないクリーンな状況においてユーザーが発話する音声を音響モデルに入力すると, どの音素に対応する状態が尤もらしいかを容易に決められるため, ある特定の状態の事後確率が大きな値を取ると考えられる．この場合, 事後確率のエントロピーは小さな値となる．一方, 背景発話が入力されると, ノイズや残響の影響によりどの音素に対応するか容易には決められず, いくつかの状態の事後確率が中間的な値を取ると考えられる．この場合の事後確率のエントロピーはクリーンな発話に比べて大きな値となる．そこで, 事後確率のエントロピーに基づく判断を追加することで背景発話が棄却できるのではないかと考えた．

我々が調べた限り, 背景発話を識別する研究は見つからない．一方, 音響信号のパワースペクトルのエントロピーを用いた VAD[11] や, 音声と音楽の識別に音響モデル事後確率のエントロピーを利用する方法などが提案されている [12]．しかし, 本稿で提案する手法はその目的とエントロピーの利用方法においてこれらの手法と異なるものである．

2. 提案手法

従来の DNN に基づく VAD では DNN が出力する音声状態 (例えば, 音素や性別など) の事後確率の値と非音声状態の事後確率の値を比べることであるフレームが音声か非音声かの識別が行われる．音響モデル用の学習データやソースコードを利用する前提だと, そのような DNN の構築方法として最も単純なものは音声と非音声の 2 状態を出力する DNN を学習する方法がまず挙げられる．他の方法としては, 音響モデルの DNN を直接利用して, 音声に対応するトライフォンに割り当てられた状態を音声状態とみなし, 無音に対応するトライフォンに割り当てられた状態を非音声状態とみなす方法がある．我々は, エントロピー計算に再利用できることから後者の方法を選択した．予備実験により, 前者の方法で学習した DNN と大きな性能差がないことを確認している．

ここから, 我々の VAD の詳細を述べる．ある時刻 t における音響特徴量ベクトルを $\boldsymbol{x}(t)$ とおき, L 層からなる音響モデル DNN の l 層目の重み行列とバイアスベクトルをそれぞれ $\boldsymbol{W}_l$, $\boldsymbol{b}_l$ とおくと, 事後確率は次のように計算される．まず, 1 層目の隠れ層の出力は次式で計算される．

$$\boldsymbol{h}_1(t) = \boldsymbol{W}_1\boldsymbol{x}(t) + \boldsymbol{b}_1 \tag{1}$$

$$\boldsymbol{o}_1(t) = g_1\left(\boldsymbol{h}_1(t)\right) \tag{2}$$

そして $l = \{2, \cdots, L\}$ 層目の出力は次式で計算される．

$$\boldsymbol{h}_l(t) = \boldsymbol{W}_l\boldsymbol{o}_{l-1}(t) + \boldsymbol{b}_l \tag{3}$$

$$\boldsymbol{o}_l(t) = g_l\left(\boldsymbol{h}_l(t)\right) \tag{4}$$

ここで, $g_l(\cdot)$ は l 層目の活性化関数である．$l = \{1, \cdots, L-1\}$ 層については次式で定義されるシグモイド関数を用いた．

$$g_l(y) = \frac{1}{1+\exp(-y)} \tag{5}$$

なお, L 層目については恒等関数 $g_L(y) = y$ を用いた．最後の L 層目の出力は次式のソフトマックス関数によって事後確率に変換される．

$$p(i|\boldsymbol{x}(t)) = \frac{\exp(o_L^i(t))}{\sum_{i'}\exp(o_L^{i'}(t))} \tag{6}$$

ここで, $o_L^i(t)$ はベクトル $\boldsymbol{o}_L(t)$ の i 番目の要素を表す．すると, 音声状態仮説 H_1 と非音声状態仮説 H_0 の事後確率は次式で計算される．

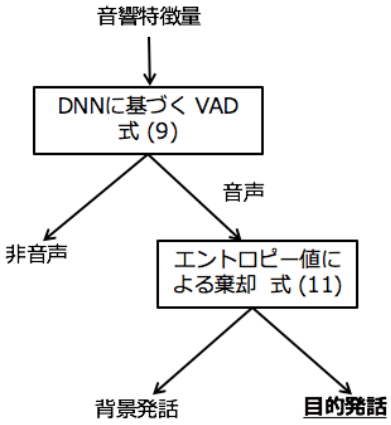

図 1 提案する VAD のブロックダイアグラム

$$p(H_1|\boldsymbol{x}(t)) = \sum_{i\in S} p(i|\boldsymbol{x}(t)) \tag{7}$$

$$p(H_0|\boldsymbol{x}(t)) = \sum_{i\in N} p(i|\boldsymbol{x}(t)) \tag{8}$$

ここで, S は音声状態を表すインデックスの集合, N は非音声状態を表すインデックスの集合である. 次式の条件が満たされた場合, 時刻 t のフレームは音声と識別される.

$$p(H_1|\boldsymbol{x}(t)) > p(H_0|\boldsymbol{x}(t)) \tag{9}$$

提案手法では, 音声と識別されたフレームに対してエントロピーに基づく背景発話の判定が追加される. エントロピーは次式で計算される.

$$e(t) = \sum_{i\in S\cup N} p(i|\boldsymbol{x}(t)) \log p(i|\boldsymbol{x}(t)) \tag{10}$$

そして, 次式の条件が満たされると, このフレームは目的とする発話つまり背景発話ではない音声だと識別されてデコーダに送られる.

$$e(t) < \tau \tag{11}$$

このアルゴリズムのブロックダイアグラムを図 1 に示す.

1 章で述べたように, 背景発話の事後確率はノイズや残響の影響によりいくつかの状態で中間的な値をとり, そういった影響のないクリーンな発話に比べると大きなエントロピーを取ると考えられる. これを確認するために, 図 2, 3 にある発話の波形, 人手でラベル付けされた音声区間, 音声状態の事後確率とエントロピー値をプロットしたものを示す. 図 2 はノイズ等のないクリーンな発話をプロットしたものであり, エントロピーは大きくないことがわかる. 一方, 図 3 は車内で発話されたものでラジオからの音声が背景発話として存在している. ラベル付けされた音声区間の前後でも音声状態の事後確率が大きくなっており, なおかつエントロピーが大きくなっていることがわかる.

また, 背景発話がエントロピーで識別できるか確認するために, 図 4 に開発セットのエントロピーのヒストグラムをプロットした. 開発セットの各フレームはアラインメント結果で生成されたラベルとフレームごとの VAD 結果によって true positive, true negative, false alarm, false rejection の 4 種に分類され, それぞれの分類ごとにヒストグラムを描いた. なお, 開発セットにおける false alarm の多くが背景発話に由来することを確認している. 図 4 から, false alarm のフレームは他のフレームに比べて大きなエントロピーを持つことがわかる. 加えて, 図 5, 6 に移動平均したエントロピーのヒストグラムもプロットした. これは, 文献 [13] で音声と音楽の識別にエントロピーを用いた場合, エントロピーを移動平均することで識別が容易になったとの記述があり, 同様の傾向が存在するかどうか確認するためである. しかしながら, 移動平均したエントロピーのヒストグラムでは, false alarm のフレームのエントロピー値が true positive のフレームのエントロピー値に近づいてしまい, 背景発話の識別において移動平均は逆効果であるこ

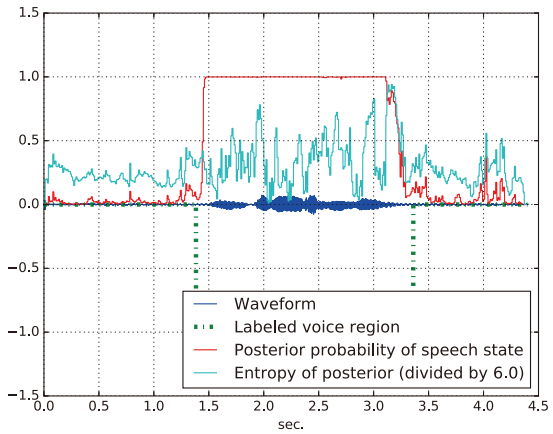

図 2 クリーンな発話の波形 (waveform), 人手でラベル付けされた音声区間 (labeled voice region), 音声状態の事後確率 (posterior probability of speech state) とエントロピー値 (entropy of posterior).

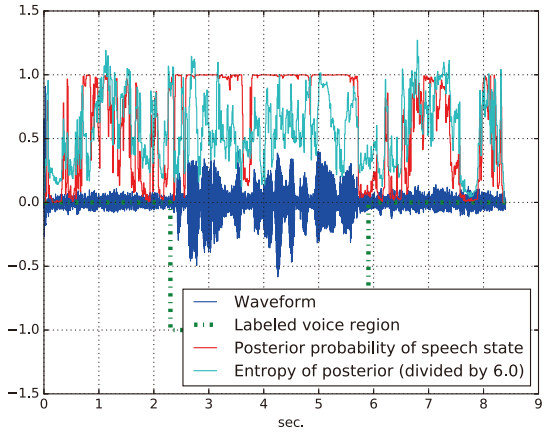

図 3 背景発話が存在する発話の波形 (waveform), 人手でラベル付けされた音声区間 (labeled voice region), 音声状態の事後確率 (posterior probability of speech state) とエントロピー値 (entropy of posterior).

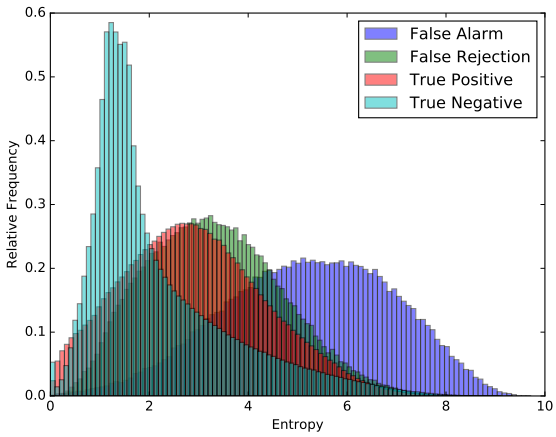

図 4 開発セットのエントロピー値のヒストグラム.

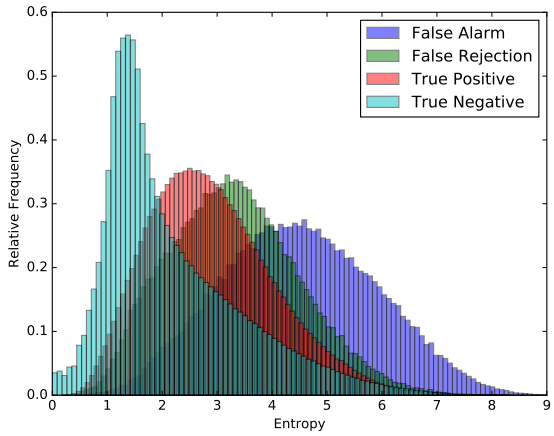

図 5 開発セットのエントロピー値を 10 フレームで移動平均したもののヒストグラム.

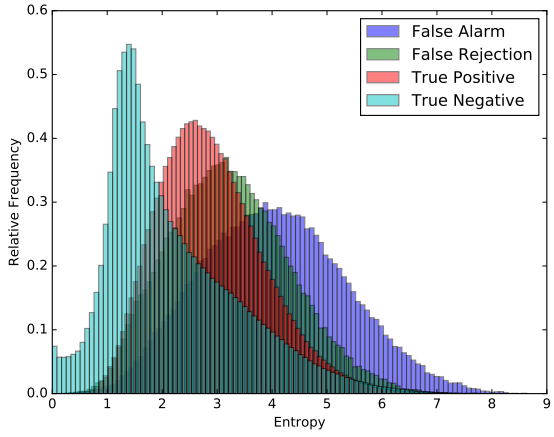

図 6 開発セットのエントロピー値を 20 フレームで移動平均したもののヒストグラム.

とがわかった. 従って, フレームごとにエントロピーに基づく判定を追加して背景発話を棄却することにした. つまり $e(t)$ がある閾値より大きい場合, そのフレームは背景発話と判定される.

3. 実験

3.1 実験条件

従来法 (baseline) と提案法 (proposed) の VAD について, 我々の音声認識サービスで収集された発話音声の書き起こしデータ約 1200 時間で学習した音響モデル DNN を用いて評価実験を行った. 従来法は式 (9) のみを用いて音声/非音声の判定を行い, 提案法は図 1 のように式 (9) と (11) の両方を用いて目的発話かどうかの識別を行う. 学習データとは別にカーナビゲーションアプリ (1 章で述べた Vehicle ドメイン) で収集された発話から 2 万発話をテストセットとして選び, 1 万発話ずつ開発セット (dev) と評価セット (eval) にした. なお, 開発セットと評価セットは異なる時期に異なるスマートフォンで発声された発話で構成されている. さらに, このテストセットのサブセットである縮小版テストセットも作成した. この縮小版テストセットは開発セット, 評価セットのうち人手でラベル付けされた音声区間でデコードした場合は正しく認識される発話のみで構成される. 一般に音声認識誤りの原因が VAD の失敗なのかデコード時のエラーなのかを特定するのは非常に困難であるため, この縮小版テストセットを用いることで VAD の性能が改善したことによる認識精度への貢献度合いを近似的に測ることにした.

性能評価には 2 つの指標を用いることにした. 1 つ目は VAD のフレームエラーレート (Frame Error Rate, FER) であり, 誤識別したフレーム数を総フレーム数で割った値である. 2 つ目は読みの文誤り率 (Sentence Error Rate, SER) である. 文誤り率を使う理由は, 我々のシステムは web などの検索クエリを音声で入力することを目的にしたものであり, 検索エンジンへの入力単語が 1 つ違うだけで結果が全く別なものになることもあることから, 広く用いられる単語誤り率 (Word Error Rate, WER) による評価結果が必ずしもユーザーが感じる主観的な性能評価の傾向を反映しないからである. また, 読みを用いる理由は, 漢字の送り仮名の違いなど検索結果に大きな影響を与えない表記の揺らぎを吸収し, VAD による認識精度への貢献をより正確に把握するためである.

テストセットの音響信号はまず VAD のプロセスに入力され, フレームごとに音声/非音声に識別される. 次に, その識別結果は人手でチューニングされた区間検出用のオートマトンに入力される. そして, オートマトンの出力によって分割された音声区間の音響信号がデコーダに渡される. デコーダは, 内製のシングルパス WFST デコーダ [14] を用いた. 言語モデルは Yahoo! JAPAN のテキストクエリと音声検索の書き起こしデータで学習した 3-gram モデルを用いた. その他のパラメータは表 1 に記載した.

表 1 音声認識システムのパラメータ.

name	value
音響特徴量	40 チャンネルのフィルタバンク
スプライシング	-5/+5
隠れ層のユニット数	1024
隠れ層の数	5
出力ユニットの数 (状態数)	4003
語彙数	1.3M

表 2 開発セットの FER.

Method	Entropy threshold	FER %
baseline	-	4.54
proposed	6.0	4.50
	7.0	4.29
	8.0	4.46
	9.0	4.54

表 3 評価セットの FER.

Method	Entropy threshold	FER %
baseline	-	4.60
proposed	7.0	4.49

表 4 縮小版テストセットの音声認識結果. 本表の SER 改善率は VAD 性能改善による音声認識精度への貢献度合いを近似的に測るものである.

	Condition	#Utts.	SER %	#Cor.	Red. %
dev.	baseline	8554	5.46	8087	
	proposed	8554	4.70	8152	13.9
eval.	baseline	8330	3.95	8001	
	proposed	8330	3.52	8037	10.8

3.2 実験結果

開発セットの FER を表 2 に示す. エントロピーの閾値を 7.0 に設定すると最良の FER となったので, これをシステムの動作点とした. この動作点における開発セットの FER の改善率は 5.5%だった. また, 評価セットの FER を表 3 に示す. 改善率は 2.4%だった.

次に, 縮小版テストセットの音声認識結果を表 4 に示す. 文誤り率で 10%以上の改善率が達成され, 提案法により従来法では VAD 誤りにより正しく認識できなかった発話を正しく認識できるようになった. また, テストセット全発話の音声認識結果を表 5 に示す. 改善率は開発セットで 4%, 評価セットで 2.2%だった. なお, テストセット全発話には VAD 誤り以外の原因により誤認識となっている発話も含まれており, 縮小版テストセットよりも改善率が小さくなっていると考えられる.

また, 今回は対象外としていた他 2 つのドメイン (Search, Dialogue) の評価も行った. 表 6 にそれぞれのドメインの音声認識結果を示す. 提案法は Vehicle ドメインに対して最適化されているにも関わらず, それ以外のドメインでも

表 5 テストセット全体の音声認識結果. このテストセットの中には VAD 誤り以外の要因で誤認識している発話も含まれている.

	Condition	#Utts	SER %
dev.	baseline	10000	16.78
	proposed	10000	16.10
eval.	baseline	10000	17.66
	proposed	10000	17.26

表 6 対象外ドメインの音声認識結果.

domain	system	#Utts	SER %
Search	baseline	10000	24.09
	proposed	10000	23.98
Dialogue	baseline	10000	23.79
	proposed	10000	23.71

従来法と性能差がほとんどなかった. 従って, 提案法は対象とした Vehicle ドメインの性能を改善しつつ, 他のドメインに悪影響を与えないことがわかった.

4. 結論

本稿では, 我々が運用している音声認識サービスの DNN に基づく VAD の主要なエラーである周囲の人の会話やテレビ, ラジオのスピーカに由来する背景発話の誤検出を抑制する手法を提案した. 背景発話はユーザーが音声認識サービスの利用を意図して行う発話よりもマイクロフォンとの距離が遠いことが多く, ノイズや残響の影響を受けやすい. そのような発話が音響モデルに入力されると, どの状態が尤もらしいかを容易には決められず, いくつかの状態の事後確率が中間的な値を取ることになり, クリーンな発話に比べてエントロピーが大きくなる. そこで, 音響モデルの DNN の事後確率のエントロピーを用いて背景発話を棄却する手法を提案した.

実験の結果, 提案法により開発セットの FER が 5.5%, 評価セットの FER が 2.4%削減された. また, 音声認識実験を行ったところ, 縮小版テストセット (VAD の性能改善による音声認識精度向上の貢献度合いを近似的に測るテストセット) の読みの文誤り率は開発セットで 13.9%, 評価セットで 10.8%削減された. テストセット全体の読みの文誤り率は開発セットで 4%, 評価セットで 2.2%削減された. また, 今回は対象外としたドメイン (Search, Dialogue) の評価により, 提案法は他のドメインに悪影響を与えず, 対象としたドメイン (Vehicle) の性能を改善できることがわかった.

参考文献

[1] ITU-T Recommendation G.729: *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)* (06/2012).
[2] ETSI ES 202 050: *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition;*

Advanced front-end feature extraction algorithm; Compression algorithms (2007).

[3] Wu, J. and Zhang, X. L.: Efficient Multiple Kernel Support Vector Machine Based Voice Activity Detection, *IEEE Signal Processing Letters*, Vol. 18, No. 8, pp. 466–469 (2011).

[4] Liang, Y., Liu, X., Lou, Y. and Shan, B.: An improved noise-robust voice activity detector based on hidden semi-Markov models, *Pattern Recognition Letters*, Vol. 32, No. 7, pp. 1044 – 1053 (2011).

[5] Fujimoto, M. and Ishizuka, K.: Noise Robust Voice Activity Detection Based on Switching Kalman Filter, *IEICE transactions on information and systems*, Vol. 91, No. 3, pp. 467–477 (2008).

[6] Zhang, X.-L. and Wu, J.: Deep Belief Networks Based Voice Activity Detection, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 4, pp. 697–710 (2013).

[7] Wang, Q., Du, J., Bao, X., Wang, Z., Dai, L. and Lee, C.: A universal VAD based on jointly trained deep neural networks, *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 2282–2286 (2015).

[8] Hwang, I., Sim, J., Kim, S., Song, K. and Chang, J.: A statistical model-based voice activity detection using multiple DNNs and noise awareness, *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 2277–2281 (2015).

[9] Ferrer, L., Graciarena, M. and Mitra, V.: A phonetically aware system for speech activity detection, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5710–5714 (2016).

[10] Obuchi, Y.: Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5715–5719 (2016).

[11] Shen, J.-L., Hung, J.-W. and Lee, L.-S.: Robust entropy-based endpoint detection for speech recognition in noisy environments., *ICSLP*, Vol. 98, pp. 232–235 (1998).

[12] Yang, C. and Hsieh, M.: Robust endpoint detection for in-car speech recognition, *Sixth International Conference on Spoken Language Processing, ICSLP 2000, Beijing, China, October 16-20, 2000*, pp. 1061–1064 (2000).

[13] Ajmera, J., McCowan, I. and Bourlard, H.: Speech/music segmentation using entropy and dynamism features in a HMM classification framework, *Speech Communication*, Vol. 40, No. 3, pp. 351 – 363 (2003).

[14] Iso, K., Whittaker, E., Emori, T. and Miyake, J.: Improvements in Japanese Voice Search, *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pp. 2109–2112 (2012).