雑音にロバストな音声と非音声の判別技術

Voice Activity Detection Technology with Robust Performance in Noisy Environments

山本　幸一　　赤嶺　政巳
■ YAMAMOTO Koichi　■ AKAMINE Masami

入力信号がユーザーの音声かどうかを判別するためにVAD（Voice Activity Detection）技術が利用されているが，自動車内のような雑音が大きな環境下では，雑音に対する十分なロバスト（頑健）性が必要である。

東芝は，このような環境下でもロバストに音声と非音声を判別できるVADを開発した。この新手法では，特徴抽出部に新しい特徴量を，また判別部に決定木（DT：Decision Tree）を用いた識別器を導入して雑音環境下でのロバスト性を向上させている。雑音が重畳した音声を用いて判別実験を行い，新しい特徴量が従来の特徴量に比べて高い雑音ロバスト性を示すこと，また，識別器としてDTが性能と演算量の両面で優れていることを確認した。

Voice activity detection (VAD) technology for preprocessing of speech recognition is used in many applications to judge whether or not input signals are the user's voice. The enhancement of robustness against background noises is required for VAD under noisy circumstances such as inside an automobile.

Toshiba has developed a VAD technology that can reliably detect the user's voice in noisy environments. In our newly proposed method, robustness in noisy circumstances is improved by introducing new features for the feature extraction portion and decision trees (DTs) for the distinction portion. Experiments using noisy speech data confirmed that the new features achieve better performance than the existing features and that the DTs provide advantages in terms of both VAD performance and computational costs.

1　まえがき

VADは，マイクロホンから入力された信号がユーザーの音声か否かを判別する技術であり，音声認識の前処理や携帯電話の音声圧縮など様々なアプリケーションに利用されている。VADは音声信号処理の前段に位置し，その性能がアプリケーション全体に与える影響は大きい。実環境で使用されるVADには，雑音環境下でも十分な性能を発揮する雑音ロバスト性と，実時間で動作するリアルタイム性が要求される。特に，自動車内のように雑音が大きな環境でのVADは難しく，雑音ロバスト性を向上させる研究が多数報告されている。

一般に，VADは二つの要素で構成される（**図1**）。一つは入力信号から音響的な特徴量を抽出する特徴抽出部，もう一つは抽出された特徴量を用いて音声か否かを判断する判別部である。特徴量としてもっとも一般的なものは信号のエネルギーであり，判別部が備える識別器がこのエネルギーを評価し，あるしきい値を超えた区間を音声と判別する。しかしこの手法では，自動車走行雑音のようにエネルギーが大きな雑音を音声と誤って判別してしまうことがあり，雑音に対するロバスト性は不十分である。

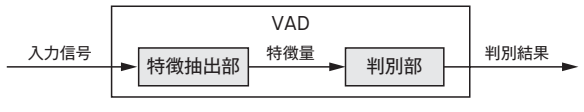


図1. VADの構成 ― 入力信号から音響的な特徴量を抽出し，その特徴量を用いて判別部の識別器で音声と非音声を判別する。
Configuration of VAD

東芝は，雑音が大きな環境下でもロバストに音声と非音声を判別できるVADを開発した[1]。ここでは，提案手法のアルゴリズムと，雑音が重畳した音声を用いて実施した判別実験の結果について述べる。

2　提案手法のアルゴリズム

2.1　特徴抽出部

当社は，雑音に対して十分なロバスト性を持つ新しい特徴量（以下，新特徴量と言う）を開発した。新特徴量は，従来から広く利用されているスペクトルエントロピー[2]に短時間SNR（Signal-to-Noise Ratio）とスペクトル間余弦値を組み合わせている。更に，判別対象となるフレーム，つまり20～30 msの長さで切り出した入力信号だけでなく，その近傍のフレームから抽出した特徴量も利用している。

2.1.1　スペクトルエントロピー　スペクトルエントロピーHは，振幅スペクトルの白色性を表した特徴量であり，(1)式で計算する。

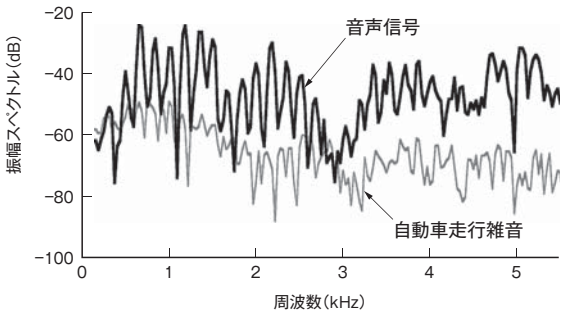


図2. 音声信号と自動車走行雑音 — 音声信号は自動車走行雑音に比べてスペクトルが不均一で，スペクトルエントロピーが低くなる。

Speech signal and noise in running car

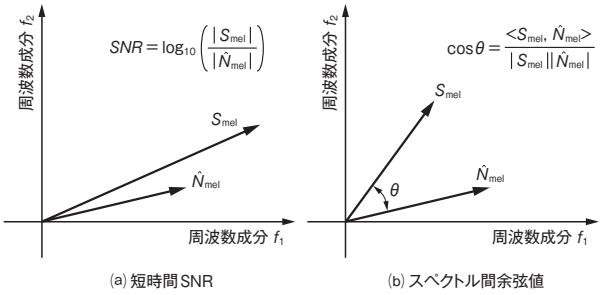


図3. 短時間SNRとスペクトル間余弦値 — 短時間SNRが大きい場合，入力信号に音声が含まれる可能性が高い(a)。スペクトル形状が違う場合，スペクトル間の余弦値を用いて表現できる(b)。

Short-time signal-to-noise ratio (SNR) and cosine value between spectra

$$H = -\sum_f P_f \cdot \log P_f$$
$$P_f = \frac{S_f}{\sum_f S_f} \qquad (1)$$

ここで，S_fは入力信号を離散フーリエ変換して得られる周波数成分fの振幅スペクトルである。Hはスペクトルが均一な白色信号では高い値となる。一般に，音声信号は自動車走行雑音などと比較するとスペクトルが不均一な有色信号であり(**図2**)，Hは低い値を示す。従来手法[2]では，Hがあるしきい値以下となった信号を音声と判別している。

2.1.2 短時間SNR　短時間SNRは，入力信号と雑音信号の相対的な大きさを表した特徴量であり(**図3**(a))，各フレームから(2)式により計算する。

$$SNR = \log_{10}\left(\frac{|S_{mel}|}{|\hat{N}_{mel}|}\right) \qquad (2)$$

S_{mel}及び$\hat{N}_{mel}$は，メルフィルタバンク処理(注1)を施した入力スペクトル及び雑音スペクトルを表している。なお，$\hat{N}_{mel}$は，音声が入力されていない区間の平均スペクトルから推定する。一般に，ユーザーが発声した区間での入力信号の大きさは背景雑音と比較して大きくなる。したがって，短時間SNRが大きい場合，入力信号に音声が含まれる可能性が高いと言える。Hは入力信号と雑音信号の相対的な大きさを抽出できない。短時間SNRを組み合わせることにより，Hの性能を補完できると考えられる。

2.1.3 スペクトル間余弦値　スペクトル間余弦値$\cos\theta$は，S_{mel}及び$\hat{N}_{mel}$の間の余弦値を表しており，(3)式で計算される(図3(b))。

$$\cos\theta = \frac{<S_{mel}, \hat{N}_{mel}>}{|S_{mel}||\hat{N}_{mel}|} \qquad (3)$$

(注1)　聴覚特性を考慮して行う帯域分割処理。

ここで，(3)式の分子で$<\cdot,\cdot>$はベクトルの内積を示している。入力信号に音声が含まれると，S_{mel}は$\hat{N}_{mel}$とは異なった形状を示す。この特徴量は，スペクトルの相対的な大きさではなくスペクトル間の余弦値を用いて，スペクトル形状の違いを表現している。

2.1.4 フレーム結合　判別対象となるフレームから抽出したH，短時間SNR，$\cos\theta$の3次元の特徴量と，その前後10フレーム目から抽出した3次元の特徴量を組み合わせることで計9次元の特徴量(新特徴量)を得る。これにより，単一フレームの特徴量では表現できないスペクトルの時間変化情報を抽出できる。

2.2 判別部

VADの判別部で一般的に利用されるのはGMM (Gaussian Mixture Model) やSVM (Support Vector Machine) と呼ばれる識別器であるが，提案手法では識別器としてDTを用いて音声と非音声を判別する。

2.2.1 DT　DTは，複数のノードと複数のリーフから成る木構造を持つ(**図4**)。木構造の最上位にあるノードはルートノードと呼ばれる。各ノードには，特徴量に関する質問

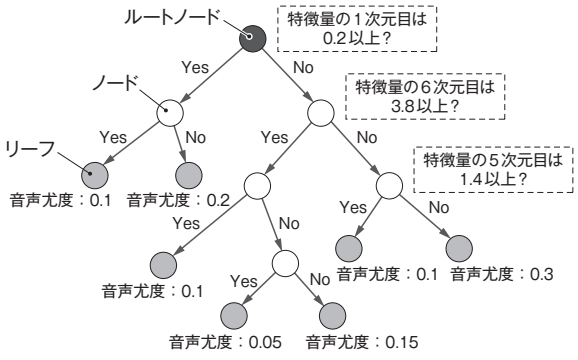


図4. DTの木構造 — ルートノードに9次元の特徴量を入力し，各ノードの質問に答えることで音声尤度を得る。

Structure of DT

があらかじめ設定されている。例えば質問は，「特徴量の1次元目が0.2以上?」といったものである。各ノードは，質問に対する答えに応じて“Yes”又は“No”の子ノードに分岐する。分岐のないノードはリーフと呼ばれ，入力された特徴量に対する音声らしさを表す音声尤度（ゆうど）を出力する。

学習時には，学習データに対するトータルの音声尤度が最大となるように，各ノードの質問やリーフ数が最適化される。各リーフでの音声尤度L_{leaf}は，(4)式で計算される。

$$L_{\mathrm{leaf}} = \frac{1}{P_{\mathrm{r}}} \cdot \frac{N_{\mathrm{speech}}}{(N_{\mathrm{speech}} + N_{\mathrm{nonspeech}})} \tag{4}$$

ここで，N_{speech}及び$N_{\mathrm{nonspeech}}$は各リーフに到達した音声フレーム数と非音声フレーム数であり，P_{r}は学習データに含まれる音声フレームの確率で，あらかじめ学習で求めておく。

判別時には各フレームの特徴量がルートノードに入力され，リーフに到達するまで各ノードの質問に答えていく。そして，到達したリーフが出力する音声尤度とあるしきい値を比較することで，当該フレームの音声と非音声を判別する。

DTにはこのほか，入力される特徴量の種類や分布に制約がない，判別に有益な特徴量を選択できるなどの特長がある。

2.2.2 DT-forest　DTから出力される音声尤度は，リーフ数の制限で離散的になる。これにより，特徴量のわずかな変化で音声尤度が大きく変動し，動作が不安定になることがある。この問題を解決するために当社は，複数の木構造を用いて音声尤度を計算するDT-forestを開発した[3]。DT-forestでは，(5)式で計算される，複数の木構造から出力される音声尤度の重み付き加算を最終的な音声尤度として出力する（**図5**）。

$$\begin{aligned} \text{音声尤度} = {} & W_1 \times L\,(X\mathrm{t}/DT_1) + W_2 \times \\ & L\,(X\mathrm{t}/DT_2) + W_3 \times L\,(X\mathrm{t}/DT_3) \end{aligned} \tag{5}$$

ここで，Xtは特徴量，W_1はDT_1での重み，$L(X\mathrm{t}\times DT_1)$は$DT_1$での音声尤度である。複数の木構造の出力を加算することにより，特徴量のわずかな変化で音声尤度が大きく変動する事態は起きにくくなる。DT-forestで音声尤度がスムージングされ，安定した判別が可能になる。なお，これ以降，一つの木構造を用いて音声尤度を計算するときはDT-singleと言う。

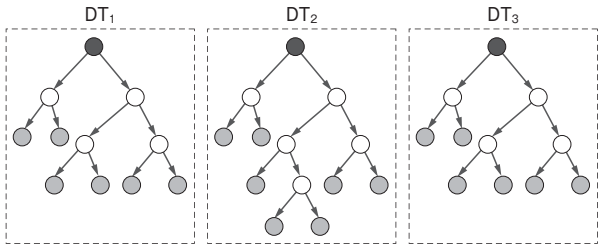


図5．DT-forestの木構造 ― 複数の木構造から出力される音声尤度に重み付き加算した最終的な音声尤度は，特徴量のわずかな変化で大きく変動しにくくなり，スムージングされて安定した判別が可能になる。

Structure of DT-forest

3　実験

3.1　実験条件

雑音が重畳した日本語の都市名発声の音声を用いて，提案手法の性能を評価した。総発声数は1,600であり，静かな環境で音声を収録した。これに，学習時に使用していない白色雑音，自動車走行雑音，バブルノイズなどの雑音を重畳した。雑音を重畳させる際のSNRは，0，5，10，及び20 dBとした。入力信号のサンプリング周波数は11,025 Hz，フレーム長は23 ms，フレームの切出し間隔は8 msに設定した。また，DT-forestにおける木構造の数は5とした。

今回の実験では，DTの比較対象としてGMM及びSVMの性能も評価した。GMMは複数の正規分布によるモデル化に，SVMはカーネルトリック(注2)を利用した非線形な識別面の構築に特徴がある。GMMの混合数は，音声と非音声モデルともに32とした。また，SVMのサポートベクター数は約3,500となった。予備実験での性能を基にこれらのパラメータを最適化した。

3.2　音声と非音声の判別性能

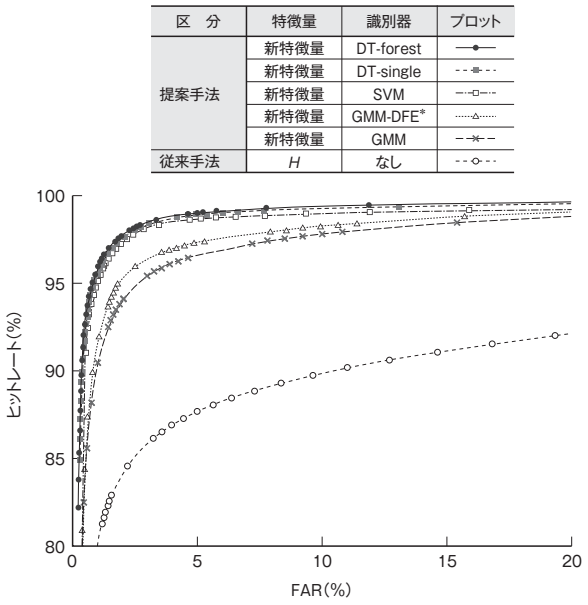
SNR 5 dBの自動車走行雑音に対するROC（Receiver Operating Characteristic）曲線を**図6**に示す。図中の横軸は非音声フレームを音声と誤って判別した割合であるFAR（False Acceptance Rate）を，縦軸は音声フレームを正しく音声と判別した割合であるヒットレートを表している。曲線上の各点は，判別部の識別器から出力される音声尤度に対してしきい値を変化させたときの結果である。特徴量Hを用いた従来手法[2]の結果と，新特徴量に対して識別器GMM，SVM，DT-single，及びDT-forestそれぞれを用いた提案手法の結果を示している。

各雑音条件での平均EER（Equal Error Rate）を**表1**に示す。EERは，FARとFRR（False Rejection Rate）が等しくなるようにしきい値を調整した場合の誤り率で，この値が小さいほど雑音に対するロバスト性は高い。

図6と表1で示すように，すべての雑音条件で新特徴量を用いた提案手法が従来手法の性能を上回っている。特にDT-forestは，従来手法と比較して判別性能が高く，全雑音における平均EERを72.4 %改善している。以上から，新特徴量で導入した短時間SNR，$\cos\theta$，及びフレーム結合が雑音ロバスト性の向上に寄与することがわかる。

識別器の性能では，SVMとDTがGMMと比較して高い性能を示している。GMMは正規分布を用いて特徴量の分布をモデル化する。新特徴量の分布が正規分布に適合していな

(注2)　特徴ベクトルを非線形変換して，その空間で線形の識別を行う方法。



*GMMの学習に組織的な枠組み[4]を導入した手法

図6. 自動車走行雑音に対するROC曲線 ― ROC曲線が左上に近づくほど，雑音に対して高いロバスト性を示す。

Receiver operating characteristics (ROC) curve for running car noise

表1. 平均EER

Average equal error rate (EER)

手　法		EER (%)			
特徴量	識別器	白色雑音	自動車走行雑音	バブルノイズ	全雑音
H	なし	14.21	8.95	16.24	12.13
新特徴量	GMM	2.58	3.95	7.99	4.78
新特徴量	GMM-DFE	2.39	3.50	6.88	4.27
新特徴量	SVM	2.12	2.60	5.40	3.65
新特徴量	DT-single	2.10	2.34	5.86	3.59
新特徴量	DT-forest	2.00	2.28	5.46	3.35

かったためGMMの性能が劣化した可能性がある。SVMとDTの比較では，全雑音の平均EERにおいてDTの性能がわずかに高くなっている。また，DT-singleとDT-forestの比較では，DT-forestがすべての条件でDT-singleを上回る性能を示しており，DTがほかの識別器と比較して高い雑音ロバスト性を示すことがわかる。

3.3　識別器の演算量

次に，各識別器の演算量について検討する。GMMの演算量は$(D \times M)$のオーダ，SVMの演算量は$(D \times V)$のオーダになる。ここで，Dは特徴量の次元数，MはGMMの総混合数，Vはサポートベクターの数である。今回の実験において，Dは9，Mは64，Vは3,500であり，演算量は非常に大きくなる。

一方，この実験で用いたDTの木構造の高さは平均で10であり，10回の比較演算だけでDT-singleの音声尤度を計算できる。DT-forestではDT-singleの約5倍の演算量を必要とするが，GMMやSVMと比較すると演算量は圧倒的に少なく，DTが演算量の面でも優れていることがわかる。

4　あとがき

雑音が大きな環境下でもロバストに音声と非音声を判別できるVAD技術を開発した。提案手法のポイントは，新特徴量の導入とDTを用いた識別器である。新特徴量では，従来の特徴量のスペクトルエントロピーに短時間SNRとスペクトル間余弦値を組み合わせた後，判別対象となるフレームとその近傍のフレームから抽出した特徴量を結合している。雑音が重畳した音声を用いて実施した音声と非音声を判別する実験の結果，提案手法では従来手法に比べて平均の判別誤りが72.4％改善されることを確認した。また，判別部に導入した識別器のDT，GMM，及びSVMを比較し，性能面と演算量の両面でDTがもっとも優れていることを確認した。

今後は，提案手法を音声認識の前処理として利用し，雑音に対するロバスト性を更に向上させる音声認識エンジンの開発を進めていく。

文　献

(1) Yamamoto, K., et al. "Comparative Evaluation of Different Methods for Voice Activity Detection". 9th Annual Conference of the International Speech Communication Association (INTERSPEECH). Brisbane, Australia, 2008-09, INTERSPEECH. (CD-ROM).

(2) Renevey, P.; Drygajlo, A. "Entropy Based Voice Activity Detection in Very Noisy Conditions". 7th European Conference on Speech Communication and Technology (EUROSPEECH). Aalborg, Denmark, 2001-09, International Speech Communication Association. (CD-ROM).

(3) Teunen, R.; Akamine, M. "HMM-based Speech Recognition Using Decision Trees Instead of GMMs". 8th Annual Conference of the International Speech Communication Association (INTERSPEECH). Antwerp, Belgium, 2007-08, INTERSPEECH. (CD-ROM).

(4) Yamamoto, K., et al. "Robust Endpoint Detection for Speech Recognition based on Discriminative Feature Extraction". IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Toulouse, France, 2006-05, IEEE Signal Processing Society. (CD-ROM).

山本　幸一　YAMAMOTO Koichi
研究開発センター　知識メディアラボラトリー研究主務。
音声認識技術の研究・開発に従事。
Knowledge Media Lab.

赤嶺　政巳　AKAMINE Masami, D.Eng.
研究開発センター技監，工博。
音声符号化・音声合成・音声認識技術の研究・開発に従事。
電子情報通信学会，日本音響学会会員。IEEEシニア会員。
Corporate Research and Development Center