雑音環境下で頑健な音声区間検出技術の開発

片桐　一浩

音声区間検出(Voice Activity Detection、以下VAD)とは、音声と雑音が含まれる信号から音声が存在する区間とそれ以外の区間を判別する技術である。VAD自体は単独で使用されることはほとんどなく、他の音響信号処理技術と組み合わせて使われることが多い[1)]。事前に音声区間と雑音区間を判別することで、その後の音響信号処理の性能を最大限引き出すことができる。例えば雑音抑圧では、雑音区間でフィルタの学習を行うことで雑音の特性を精度良く推定し、雑音を抑圧できる。また音声符号化では、音声のみ符号化することにより、効率的に帯域を利用できるようになる。さらに音声認識では、音声区間で処理を行うことが認識率の向上や演算量の削減につながっている。このようにVADは多くの音響信号処理技術において欠かせないものとなっている。通常VADは前処理として使用するため、VADの検出精度が後段の処理性能を大きく左右することになる。そのためVADは、あらゆる雑音環境下で高精度であることが強く求められる。

本稿では、雑音環境下で頑健なVADの技術を紹介し、実験による評価結果を報告する。

スペクトルエントロピー法

VADでは音声区間と雑音区間を判定するために、様々な音響特徴量を利用する。最も基本的なVADは、信号のパワーを利用するものである。この手法は演算量が少ないため、一般的に使用されている。しかしパワーの情報だけでは、入力信号のレベルが急に大きくなると雑音区間を音声区間と誤検出してしまう欠点がある。

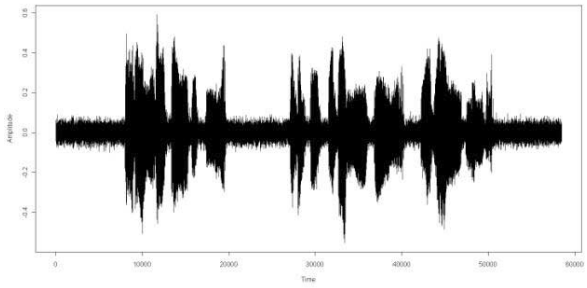
入力信号レベルの変動の影響を受け難いVAD手法として、スペクトルエントロピーを利用したものがある[2)]。スペクトルエントロピーとは、信号のスペクトルを確率分布とみなし情報エントロピーを計算したものであり、信号の白色性を示した特徴量である。スペクトルエントロピーは、ホワイトノイズのようなスペクトルが均一である信号では高い値となり、音声信号のようなスペクトルが不均一な信号では低い値となる。**図1**にSN比10dBでホワイトノイズを含む信号に対するスペクトルエントロピー示す。**図1**(b)よりスペクトルエントロピーは雑音区間では高く、音声区間では低くなっていることが分かる。この違いを利用し、音声区間の検出を行うことができる。

スペクトルエントロピーは以下の式から算出される。

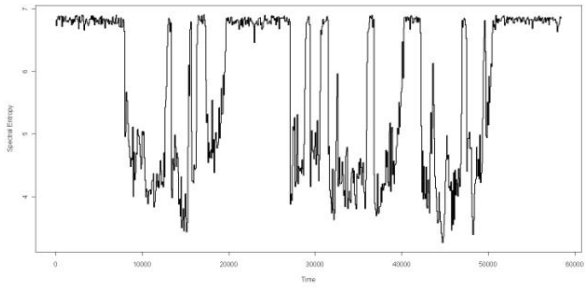
$$H = -\sum_{k=M}^{N} p_k \log_2 p_k \qquad (1)$$

$$p_k = s_k / \sum_{i=M}^{N} s_i \qquad (2)$$

ここでp_kはk番目の周波数のパワーの存在確率、s_kはk番目の周波数のパワーである。またMとNは処理対象の周波数帯域の下限値と上限値である。スペクトルエントロピーは、(2)式により正規化を行なっているので、入力信号のレベルが変化したとしても、スペクトルの形状が変わらなければスペクトルエントロピーは変わらない。この特性により、信号のレベルが変動しても安定して音声区間を検出することができる。また雑音のスペクトルが



(a) 入力信号

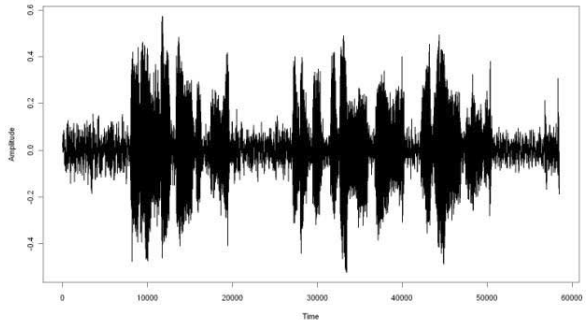


(b) スペクトルエントロピー

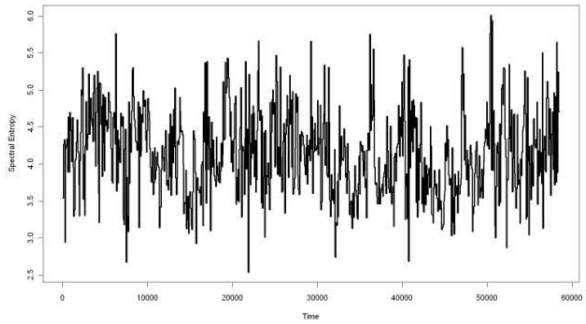
図1　SN比10dBで定常雑音（ホワイトノイズ）を含む音声信号のスペクトルエントロピー

均一でなくても、時間によりスペクトルが変化しない定常雑音であれば、入力信号のスペクトルを推定した雑音のスペクトルで除算し、雑音成分のスペクトルを均一化することにより対処できる[3)]。

しかし、雑音のスペクトルが均一でなくかつ時間的に変化する非定常雑音では、雑音のスペクトルが推定できず対処できないという問題が残っている。**図2**にSN比10dBで非定常雑音のオフィスノイズを含む信号に対するスペクトルエントロピー示す。**図2(b)**から音声と非定常雑音のスペクトルエントロピーの差がほとんどなく、音声区間の検出が困難であることがわかる。



(a) 入力信号



(b) スペクトルエントロピー

図2　SN比10dBで非定常雑音（オフィスノイズ）を含む信号のスペクトルエントロピー

提　案　手　法

本稿では、スペクトルエントロピーの収束性と変化率の2つの特性を利用することで、非定常雑音に対処することができる新たなスペクトルエントロピー法を提案する。

先行研究[1)]で示されているように、雑音のスペクトルがホワイトノイズのように均一であれば、スペクトルエントロピー法は高い検出精度を誇る。つまり非定常雑音のスペクトルを均一に近づけることができれば、検出精度を改善できると考えられる。本提案手法では、入力信号の全周波数帯域でパワーを加算することにより、非定常雑音のスペクトルを均一に近づけ、検出精度を改善する。パワーの加算は、次式により行う。

$$S'_k = S_k + \alpha_i \tag{3}$$

ここで、α_iはパワーの加算量を示す変数である。α_iの値は、以下のように決定する。

まず最大スペクトルエントロピーを算出する。最大スペクトルエントロピーは、周波数分解能によって一意に決まる値であり、次の式から算出される。

$$H_{max} = \log_2(N+1-M) \tag{4}$$

次に雑音のスペクトルエントロピーを算出する。この雑音のスペクトルエントロピーと最大スペクトルエントロピーとの比を加算量の指標とし、次のように定義する。

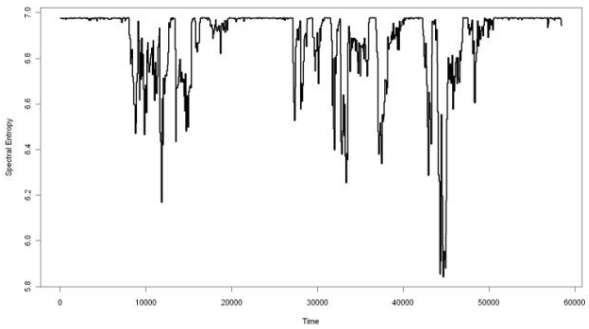
$$\beta \leq \frac{\overline{H'}_{noise}}{H_{max}} \tag{5}$$

$\overline{H'}_{noise}$は、パワー加算後の雑音のスペクトルエントロピーである。βで設定した値を上回るように雑音のパワーを加算していき、βを上回る最小の加算量をα_iと設定する。α_iは、音声が検出されている間同じ値が適用され、次に雑音を検出した際更新される。この加算指標を使うことで、様々な雑音の特徴に応じてパワーを適切に加算することがでる。

提案手法の効果を**図3**に示す。**図2(b)**の既存手法と比べ、雑音区間のスペクトルエントロピーが高い値を保っており、音声区間が明確に分かるまでに改善されていることがわかる。

提案手法が定常雑音だけでなく非定常雑音にも対処できる理由は、スペクトルエントロピーの持つ2つの重要な特性によるものである。

まず一つ目の特性は、信号のパワーを全周波数帯域で



提案手法によるエントロピー

図3　SN比10dBで非定常雑音（オフィスノイズ）を含む信号（図2（a））に対する提案手法のスペクトルエントロピー

同量加算したときのスペクトルエントロピーの収束性である。パワーを加算するとスペクトルはもとの形状に比べ均一になり、スペクトルエントロピーが大きくなる。また(1)式から加算量を増やすとスペクトルエントロピーは対数的に変化し、最大エントロピーの値に向かって収束していく。パワーの加算量を増やしていったときの非定常雑音(オフィスノイズ)のスペクトルの変化を**図4**に示す。図中の曲線は、オフィスノイズのスペクトルエントロピーを16ミリ秒ごとに100回算出し、さらにそれぞれパワースペクトルを徐々に加算していったときのスペクトルエントロピーを表している。また図中の一番上の直線は最大スペクトルエントロピーである。**図4**から、非定常雑音は時間によりスペクトルが異なるためスペクトルエントロピーの初期値は異なるが、パワーを一定量加算してしまえばスペクトルエントロピーはほぼ同じ値に収束することがわかる。

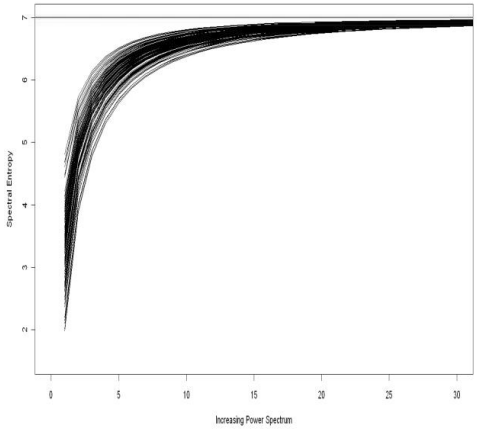


図4　オフィスノイズのスペクトルエントロピーの収束性

二つ目の特性は、パワーを加算したときのスペクトルエントロピーの変化の度合いは、もとの信号のパワーに依存するというものである。**図5**は、スペクトルの形状が同じでパワーが違う信号に対して、それぞれパワーを同量加算したときのスペクトルエントロピーの変化を示している。スペクトルの形状は同じなのでスペクトルエントロピーの初期値は同じであるが、もとのパワーが大きいほど影響を受け難く、また(1)式からスペクトルエントロピーは対数的に変化するので少しの加算で大きな差が生まれることになる。

これらの特性を利用することで、非定常雑音への対処が可能となる。信号の全周波数帯域で(3)式に従いパワーを加算すると、定常雑音、非定常雑音にかかわらず、雑音のスペクトルエントロピーは最大値に向かって対数的

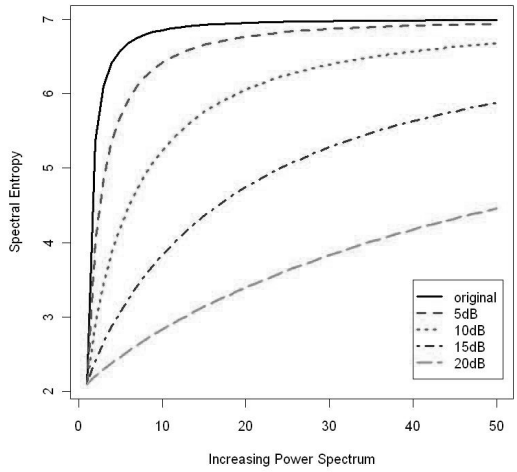


図5　パワーの違いによるスペクトルエントロピーの変化。原信号をそれぞれ5dB、10dB、15dB、20dB大きくした信号に対し、パワーを加算したときのスペクトルエントロピーの変化を示している。横軸は加算量であり、例えば数値が10のときは、原信号の平均パワースペクトルを10倍した値をそれぞれの信号に加えている。

に増加していく。同様に音声のスペクトルエントロピーも増加していくが、通常音声は雑音よりパワーが大きいため雑音のスペクトルエントロピーに比べ変化が小さい。このため雑音が非定常雑音であっても、音声区間と雑音区間のスペクトルエントロピーに大きな差が生まれ、検出精度を向上させることができる。

性能評価実験

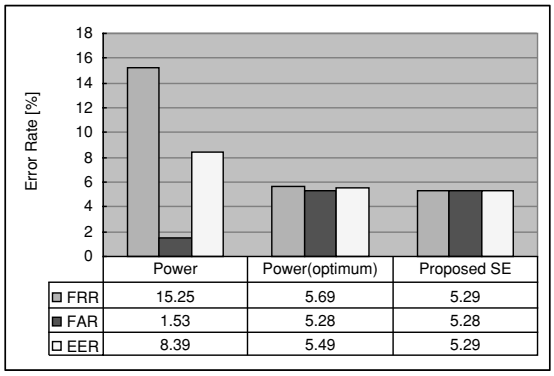
音声信号に定常雑音(ホワイトノイズ)と2種類の性質の違う非定常雑音(オフィスノイズ、パーティノイズ)を10dBで重畳した信号に対して、提案手法の性能評価を行った。オフィスノイズは突発的に音声や電話の呼出音などが発生し、スペクトルが急激に変動するノイズである。それに対し、パーティノイズは常に多数の音声が混ざっている雑音であり、スペクトルの時間変化はオフィスノイズと比べると小さい。比較対象として、信号のパワーを利用したVAD(以後パワーVAD)を用い、同じ条件で性能評価を行った。使用した音声データは女性の音声で、サンプリングレートは8kHzである。

性能評価の指標として、誤棄却率(False Rejection Rate；FRR)：音声区間を誤って雑音区間として検出したフレームの割合、誤受理率(False Acceptance Rate；FAR)：雑音区間を誤って音声区間として検出したフレームの割合、等価エラー率(Equal Error Rate；EER)：FRR＝FARとなる点での誤検出率を使用する。しかし実際FRR＝FARとなる点を見つけることは難しいため、本稿ではFRRとFARがトレードオフの関係にある点を考慮し、推定EERを次のように定義する。

$$EER_{est} = \frac{FRR+FAR}{2} \quad (6)$$

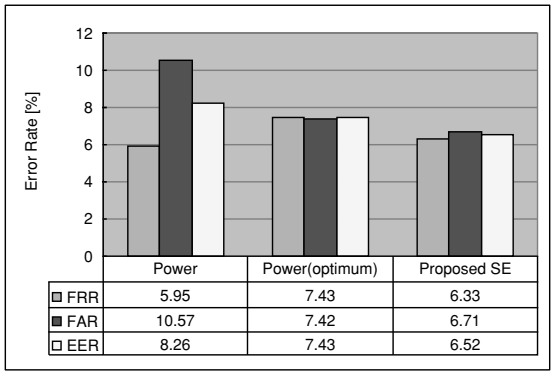
これらの指標を用いてVADの安定性と検出精度を評価する。含まれる雑音の種類にかかわらずFRRとFARの差が少ないほど安定性が高く、またEERが低いほど検出精度が良いことを示している。

図6は入力信号に対して、提案手法(Proposed SE)とパワーVADを用いてFRR、FAR及びEERを求めたものである。パワーVADに関しては、判定閾値などの各種パラメータを毎回同じ値に設定したもの(Power)と、入力信号ごとにEERが最も低くなるように(FRRとFARの値がほぼ同じになるように)パラメータを調節したもの(Power(optimum)2つを調べた。Power(optimum)は本評価実験で使用したパワーVADの性能限界を表している。



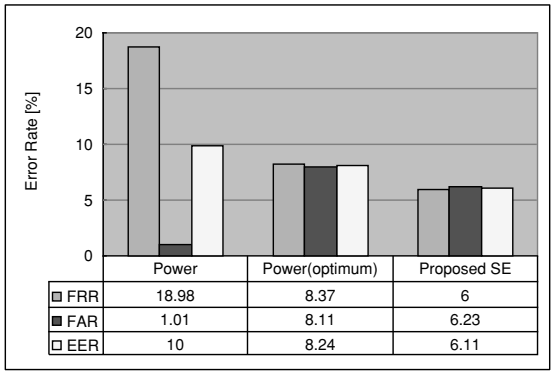
	Power	Power(optimum)	Proposed SE
FRR	15.25	5.69	5.29
FAR	1.53	5.28	5.28
EER	8.39	5.49	5.29

(a) ホワイトノイズ



	Power	Power(optimum)	Proposed SE
FRR	5.95	7.43	6.33
FAR	10.57	7.42	6.71
EER	8.26	7.43	6.52

(b) オフィスノイズ



	Power	Power(optimum)	Proposed SE
FRR	18.98	8.37	6
FAR	1.01	8.11	6.23
EER	10	8.24	6.11

(c) パーティノイズ

図6　雑音に対する各方式の検出精度

まず安定性について評価する。パワーVAD(Power)のFRRとFARの差が平均で12.1ポイントであるのに対し、提案手法では平均で0.2ポイントと差がほとんどなく安定していることが分かる。

次に検出精度について評価する。提案手法は、定常雑音(a)と非定常雑音(b)、(c)対してEERが7%以下と高い検出精度を誇っていることがわかる。特に非定常雑音のパーティノイズ(c)では、Powerよりも検出精度が39%改善する結果となった。またPower(optimum)と比較しても、EERは提案手法の方が総じて低く、パワーVADの性能限界よりも検出精度が優れていることがわかる。

ま と め

本稿では、定常雑音だけでなく非定常雑音にも対応できる新たなVADの技術を紹介した。従来のスペクトルエントロピー法を改良することにより、雑音の種類にかかわらず音声区間と雑音区間のスペクトルエントロピーの差を明確にできることを示した。また性能評価実験により、提案手法の検出精度が非定常雑音でも安定して高く、一般的なパワーVADよりも優れていることを示した。

今後の課題としては、演算量と検出精度のコストパフォーマンスを解析、評価することが挙げられる。◆◆

■参考文献

1) 石塚健太郎，藤本雅清，中谷智広：音声区間検出技術の最近の研究動向，日本音響学会誌，65巻，10号，pp.537–543，2009年

2) J. Shen, J. Hung and L. Lee,: "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments," ICSLP-98, 1998.

3) P. Renevey and A. Drygajlo,: "Entropy Based Voice Activity Detection in Very Noisy Conditions," Eurospeech 2001, 2001.

●筆者紹介

片桐一浩：Kazuhiro Katagiri. 研究開発センタ メディア処理技術研究開発部