験的モード分解と変調スペクトルを用いた ロバストな音声区間検出法に関する研究

金井　康昭 (0910013)

北陸先端科学技術大学院大学 情報科学研究科

2012 年 2 月 6 日

キーワード: 音声区間検出，経験的モード分解，変調スペクトル分析，正答率.

音声区間検出（Voice Activity Detection, VAD）とは，解析信号から音声の存在する区間（音声区間）と音声の存在しない区間（非音声区間）を自動的に検出する技術である．VAD 自体は要素技術であるが，その応用範囲は，音声符号化技術や雑音抑圧技術，音声認識技術への応用など多種多様である．そのため，実用的な VAD を考えると，対象音声以外の音の混入や，背景雑音がある状況でも，正確に解析信号から音声区間を検出できる必要がある．古典的な VAD 法は，信号のパワーの閾値処理やゼロ交差数を利用して行われる．この手法は，観測信号が対象音声のみをであれば，ほぼ完全に音声区間を検出できるが，対象音声以外の成分（雑音成分）が混入すると検出精度の著しい低下を招くという問題がある．実際問題として，VAD を活用する様々なアプリケーションは雑音のある環境や対象音声以外の非音声信号が混在する環境で利用されるため，実環境で VAD を活用するためには利用環境や対象信号に依存せずに正確に音声区間を検出できなければならない．近年，これらの問題を解決するために，閾値の決定に統計的手法を用いる Otsu の閾値法や複数のパラメータ（スペクトル歪み，全帯域のエネルギー，低帯域のエネルギー，ゼロ交差数の４種類）を音声／非音声の判別に利用する G.729 Annex B の VAD 法 や音声の非線形振動特性に着目した Exponentioal 自己回帰モデルを用いた手法など，実環境を想定した VAD 法が提案されている．これらの方法は，雑音環境でも高精度な音声区間検出が行えるが，音声と非音声を区別して検出する能力は弱い．そのため，非音声信号が存在する環境では，高精度な VAD を行うことはできない．本研究では，実環境で正確な音声区間検出を行うために，背景雑音の混入，および，目的音声以外の非音声信号の混入に対して，ロバストな VAD 法の実現を目的とする．そのために，経験的モード分解（Empirical Mode Decomposition, EMD）を利用して定常雑音の除去を行い，変調スペクトル分析（Modulation Spectrum Analysis, MSA）を用いて得られた信号区間の音声／非音声判別を行う，二段階処理からなるロバスト VAD 法を提案する．

EMD は，解析信号を定常性の高い成分から順に固有モード関数（Intrinsic Mode Function, IMF）に分解する特徴をもつ．そのため，解析信号の成分を定常雑音の成分と非定

常な音声信号の成分に容易に分解できるが，その境界を IMF 上で正確に区別することが難しい．一方，MSA は解析信号の変調スペクトルを分析するものである．変調スペクトルは音声や環境雑音など，信号の種類によって異なった特徴を示すことが知られており，この特徴を比較することで解析信号が音声か非音声かを区別することができる．しかし，解析信号に背景雑音が含まれる場合，それぞれの変調スペクトルの特徴が混ざってしまい，目的音声の特徴を正確に判別することが難くなってしまう．しかし，EMD により背景雑音を除去し，MSA により音声特有の特徴を正確に抽出することで，互いの欠点を補いつつ，背景雑音と非音声信号の混入に頑健な VAD を実現できるものと考えられる．

5 種類の環境を想定した実験刺激を作成し，その実験刺激に対して提案法 VAD を行うことによって，各環境における提案法の性能評価を行った．また，従来法として Otsu の閾値法や G.729 Annex B の方法，パワーエンベロープの閾値処理による VAD の正答率も調べ，比較を行った． 想定した環境は，音声信号のみのクリーンな環境，背景に定常雑音が存在する環境，音声以外に非音声信号も存在する環境，非音声信号と背景雑音（定常）が同時に存在する環境，そして，現実的な環境（音声以外にも他の非音声信号を含み，背景に非定常雑音が存在する環境）である．これらの環境に対する評価を，それぞれ実験 1 から実験 5 として行った．評価には正答率（%）と誤受理率（False acceptance rate, FAR），誤棄却率（False rejection rate, FRR）を用いた．正答率は，全データ数のうち，音声／非音声の判別が正しく行われたデータ数の割合から求めた．FAR は非音声区間を音声区間として誤検出した割合であり，FRR は音声区間を非音声区間として誤棄却した割合である．この二つの値はトレードオフの関係にある．実験に用いたデータは，ATR データベース A-set から音声信号と定常雑音，Noisex-92 から環境雑音，RWC 研究用音楽データベースから楽器・演奏音，そして，Avian Vocalizations Center のデータベースから鳥の鳴声である．これらのデータから，実験刺激を作成し，解析信号として用いた．実験刺激は全て，サンプリング周波数 20 kHz で調整した．実験 1 の結果，どの手法でも，正答率，FAR，FRR ともに高い数値を示していた．このことから，提案法，従来法ともに，クリーンな環境では高い精度で音声区間の検出を行えることが確認できた．実験 2 の結果，従来法では雑音に紛れた音声区間を正確に検出できなかったり，SNR の低下に伴い，雑音に紛れた音声区間を検出しそこなうことで，FRR が増加してしまっていることが確認できた．しかし，提案法は SNR が低くなっても正答率の減少や FAR，FRR の増加はほとんどみられなかった．この結果から，EMD による定常雑音除去の効果を確認できた．実験 3 から，従来法は，非音声の区間も音声区間として検出してしまっていた．これに対し提案法では，検出された信号が存在する区間に対し，それぞれの区間の変調スペクトルから，その区間が音声の区間かどうかの判別を行い，音声区間のみを検出することで平均的に高い正答率と FAR を維持できていた．この結果から，MSA による音声／非音声区間判別の効果を確認できた．実験 4 から，従来法は，背景雑音の影響で非音声区間の判別ができていなかったり，非音声区間を音声区間と検出するエラーがみられた．提案法は，音声区間のみをうまく検出できていた．また， SNR の増加に伴う正答率の減少

や FAR，FRR の増加は見られなかった．この結果から，提案法は，非音声信号を含み雑音のある環境下でも，高い精度で音声区間検出を行えることが確認できた．実験 5 から，提案法が，音声区間のみを検出できていることに対して，従来法は非音声区間を含む，ほとんどの区間を音声区間として検出してしまっていることが分かった．しかし，提案法は SNR が増加しても提案法に比べ高い正答率を維持できていた．この結果から，実環境をそいていした環境でも，提案法は効果があることが確認できた．

以上の結果から，EMD と MSA を組み合わせて用いることで，背景雑音と非音声信号の混入にロバストな VAD の開発という目的を達成することができた．