審 査 の 結 果 の 要 旨

氏名　エスピ　マルケス　ミケル

本論文は、「Characterization and classification of speech and non-speech acoustic events (音声・非音声の特徴づけと自動分類に関する研究)」と題し，英文で記されており，7 章から成る.

第1章は序論で，本論文が扱う分野の説明と応用領域などについて説明されている。人間は聴覚を頼りに、周囲の環境把握、音声言語理解、危険の判断などをすることができる。この機能の計算機による実現はマルチメディア処理、ロボティクスなどにおいて有用である。特に音声区間検出は音声認識のフロントエンド、音響ドキュメント（ミーティングの録音など）のdiarization, サーベイランスに必要な機能である。本論文は音声イベントの特徴づけ（characterization）、音響イベントの認識（recognition）、実環境理解による音響イベント理解（understanding）の3つの観点から、音響的側面から言語的側面に渡ってモデルを構築している。

第2章は関連研究であり，音声音響イベント検出には主に２つのアプローチが考えられる：音源分離に基づくアプローチ（Computational Auditory Scene Analysis）と、知識に基づくアプローチ（音声認識）である。しかし人間が雑音環境下で、知らない言語であっても、その音声区間を検出できることは、先の２アプローチとは別のアプローチを示唆している。

第3章は「characterization」で、音声の階層的な定義とその特徴付けが論じられている。音声を定義するには言語から音響までのレベルの考慮が必要である：発話（単語列）、単語（音素列）、音素、それぞれの音響的な特徴（エネルギ、ピッチ）などがある。本論文ではintra-phonemeレベル（音素と瞬時的な音響特徴の間）にまだ考慮されていない特徴量があり、声は音素中でも定常でなく変動することが議論されている。そして、あらゆる目的に有効な頑健な特徴量はなく、これら特徴量をあわせて（ANDでもORでもなく）音声を特徴づけることが必要であると主張されている。本研究では従来特徴量にさらに加える形で、intra-phonemeのスペクトル変動に関する特徴量を論じた。従来法である2段HPSS(Harmonic-Percussive Sound Separation)を音声に用いることで、音声らしいスペクトル変動がある部分を抽出することができる。CENSREC-1-Cデータベースの様々な雑音下での音声強調実験、音声区間検出の実験により、従来特徴量と今回のスペクトル変動特徴量の複合アプローチの有効性が示された。

次に第4章は「recognition」、単一イベントモデルと先見的知識に基づいた情報統合による分類が論じられている。まず、非音声音響イベント認識には、イベントがとてもheterogeneousであるという問題がある、すなわちそれぞれの音響イベントごとに異なった妥当な特徴量があると考えられる。心理学に拠ると、人間は音響イベントの認識の際に、categorical perceptionをしているとされ、生理学では音響処理のためにそれぞれの脳領域が別々なスペクトロ変動スケールから情報を抽出している可能性が指摘されており、その情報を組み合わせることで音響イベントを認識しているとされる。そこで本論文では、そのようなアプローチをできるように、別々なスペクトル変動スケールから同時に動く情報抽出部、そしてその情報結合して認識するために、音声処理のようにシーケンスを扱えるモデルを構築する必要がある。従来のtandem connectionist モデルでは、そのような情報抽出（多層パーセプトロン）とシーケンスモデ

ルを実現している。しかしながらtandem connectionistは初期統合を行う方法であった。本論文では知識に基づいてスペクトル変動スケールごとに情報抽出するための結果統合モデルを提案し、シーケンスモデルは隠れマルコフモデルでモデル化するとよいことを論じた。CHIL2007データベース（ミーティング中での12個の音響イベント）の非音声音響イベント認識タスクの実験により、全体的な精度およびイベントごとのパフォーマンスも向上が見られ、更に近似精度改善に伴い認識率が向上することが示された。

次に第5章は「understanding」、実環境シーンやイベントの発生についての言語モデルが提案されている。実環境での処理では、音響イベントが雑音や他の音響イベントとオーバーラップするという問題がある。音声認識や音楽認識の分野で言語モデルを利用することでパフォーマンスが向上しているが、音声や音楽の場合は文法や語彙があるため言語モデルが比較的議論されやすかったと考えられる。人間の知覚では、場所や状況（situations）を頼りに音響イベントを認識をしている。これは事前確率として扱うことができる。例えばsituationによってイベント発生の尤度が変わると仮定することができる（学会の休憩ではspoon clingingがtriangleの演奏より起こりやすい）。よって、situations・events・acoustic signalの３つのエンティティの階層モデルが考えられ、これら３階層には関係がある。まず、situationsとeventsの関係では、音響シーンにはeventsがあり、それぞれのイベント毎回全く同じではなく分散があり、situationが分かれば認識パフォーマンスが上がる。これと同形に、手書き単語認識にも「単語」があり、分散もあり、「トピック」が分かれば認識パフォーマンスが向上する。それに、situationは重畳することがあるため、situationsは組み合わせとして扱うことが必要である。これは、ドキュメント分類のようにProbabilistic Latent Semantic Analysis (pLSA)でモデルできると考えられる。次にeventsとacoustic signalの関係では、イベントはスペクトログラムのテンプレートと考えられ、音響信号がそのテンプレートとイベントactivationsに分解ができ、重畳や雑音にロバストな2次元 Nonnegative Matrix Factorization (2DNMF) を利用できる。このような考察によって、この章ではトピックモデルのpLSA(言語)と信号処理の2DNMF（音響）を組み合わせたsituationモデルが提案される。CHIL2007データベースの非音声音響イベント認識タスクの実験により、いくつかのsituation数の設定のもとでも全体的に正解率が向上することが示された。

第6章では、音声・非音声の分類について議論されている。聴覚知覚の種類を音響的な複雑性と言語的な複雑性の2次元平面で考えると、第3章のcharacterizationは音響的な複雑性が高い一方言語的な複雑性が低く、第4章のrecognitionは音響・言語ともに中程度の複雑性があり、第5章のunderstandingは音響的な複雑性が低く言語的な複雑性が高いと考えられる。人間の聴覚知覚では、音響的な複雑性と言語的な複雑性がともに高い場合も扱えると考えられ、本論文の手法はそれには及ばないが、人間のような聴覚知覚を実現するための一里塚となったと考えられる。

最後に第 7 章「Conclusion（結論）」では，本論文の研究成果をまとめ，今後の可能性について言及している。

本論文の大きな貢献は、観測音響信号の分類において、characterization（音響）・recognition（イベント毎）・understanding（シーン）の階層的なアプローチを初めて論じたことにあり、永年を費やして音声分野において築かれた音響・音素・単語・文・対話の階層モデルに相当する意義がある。この論文により、言語的側面から音響的側面までの音声特徴の階層的な定義と、新たな特徴のこれら階層内の位置づけ、単一イベントモデルの構築と知識の情報統合による音響イベントの分類、実環境理解のための音響イベント発生の言語モデル構築の３つの課題に解決を与えた。すなわち本研究は情報理工学に関する研究的意義と共に、情報理工学における創造的実践の観点でも価値が認められる。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。