論文の内容の要旨

論文題目　　背景雑音と話者の違いに頑健な音声認識

氏　名　　鈴木　雅之

音声認識は様々なシステムの要素技術として利用されている．例えば，カーナビシステム，スマートフォンの音声対話システム，企業のコールセンタにおける電話自動応対システムなど，その応用範囲は多岐に渡る．音声認識の精度を高めることは，これらのシステムのユーザ満足度を向上させることに直結する．そのため，音声認識の精度を向上させるために着実に研究を進めていくことが重要である．

音声認識の精度は，様々な要因によって低下してしまうことが知られている．例えば背景雑音が音声に重畳してしまった場合，何も対処を行わないと音声認識精度は大幅に低下してしまう．他にも，話者の違い，マイクとの距離，部屋の残響，話している内容と，様々な要因によって音声認識精度が低下してしまう．

本論文では，背景雑音と話者の違いに対して頑健なシステムを構築することで，より精度の高い音声認識の実現を目指す．単純に目指すといっても，既に音声認識に関するこれまでの研究の歴史の中で，背景雑音や話者の違いに頑健にするための手法が数多く提案されている．そのため従来手法をよくサーベイし，それでもなお精度向上が見込める分野を重点的に研究していくのが望ましい．

現時点で考えうる，計算コストを無視して話者・雑音に頑健な手法の一つは，まず VAD で音声区間を求め，その区間から求めた MFCC や PLP などの音響特徴量を，特徴量正規化し，VTLN し，それを前後数フレーム連結して LDA し，STC し，fMLLR し，特徴量強調したものを特徴量にして，音響モデルとして HMM/DNN，言語モデルとして modified Kneser-Ney

smoothing をかけた N-gram を用いて WFST デコーダで音声認識し，それを様々な特徴量を利用した識別モデルでリランキングし，このようなシステムを複数集めてシステムコンビネーションしたもの，となる．

サーベイの結果，本論文で特に注目したのは，背景雑音に頑健にするための技術である，特徴量ドメインでの雑音抑圧と，話者の違いに頑健になることが予想される，識別的リランキングにおける音声の構造的表象の利用である．

まず，近年の音声認識では音響モデルに HMM/DNN を使うケースが増えたため，話者や雑音のミスマッチ問題は，モデル適応ではなく，特徴量側で解決していくことが必要になると考えられる．そこで本論文では，雑音のミスマッチを特徴量側で解決する，特徴量強調法に注目する．特徴量強調では，VTS 強調や SPLICE が精度が高い手法として知られているが，それぞれに関して解決すべき問題点が残されている．

VTS 系の特徴量強調アルゴリズムは，クリーン音声 GMM のインデックスの事後確率を求める際に，分散共分散行列の逆行列を求める必要があるが，FBANK を利用する場合は対角になるため計算量が問題にならないが，MFCC を利用する場合には全角になるため，計算量がかかる．この処理は，雑音モデルが変化する度に必要になるため，非定常雑音環境下で雑音モデルが時間と共にすばやく変動する場合には現実的でなくなってしまう．結局，MFCC より精度の低い FBANK 領域を用いるか，精度の高い MFCC を使う代わりに雑音モデルが数秒の間固定したままにするか，のどちらかが必要になる．また VTS 系の特徴量強調では，特徴量として PLP や，前後数フレームの特徴量に LDA をかけた特徴量空間では利用できないことも問題点の一つである．

SPLICE 系のアルゴリズムは，任意の特徴量空間で利用することができて，しかも非常に高速に動作する．しかし，ステレオデータを用いる手法であるため，突発的な非定常雑音など，学習用ステレオデータの雑音環境に含まれていない雑音が重畳してしまった場合には，正しく特

徴量強調を行うことができない．その一つの解決策として NMN-SPLICE があるが，NMN-SPLICE 対数をとった後の特徴量空間において引き算を行うというヒューリスティックな手法であり，なぜそれでうまく動作するのかには疑問が残る．

本論文では，高速に動作するという点で，SPLICE などのステレオデータを用いる特徴量強調に注目する．そして，ステレオベースの特徴量強調を非定常雑音にも頑健になるように改良する手法を提案する．具体的には，区分的線形変換において各部分空間の事後確率を求める部分の計算を，クリーン音声状態の識別と捉える考え方を導入し，その入力特徴量として，観測したノイジー音声の特徴量に加え，推定した雑音特徴量や，前後数フレームの特徴量を入力として利用することを提案する．加えて，線形変換の次元数が高くなった場合に L2 正則化を導入する手法も提案する．AURORA2 データベースを用いた実験の結果，クリーン音声状態の識別，結合特徴量を線形変換に用いること，正則化にそれぞれ効果があり，SPLICE や NMN-SPLICE を越える精度が実現できることが分かった．

次に注目すべき点として，音響モデルの研究からでてきた識別モデルを用いた音声認識と，識別的言語モデルの研究は，ほぼ同じような手法と目的を持ちつつ，ここまで互いに独立に発展してきていることがある．特に，音響モデル側の研究では識別モデルを使うことそのものに注目した研究が多く，どのような特徴量を用いるかについてないがしろにされていた点である．逆に識別的言語モデルの研究では，NNLM の尤度など，文全体にまたがる特徴量を積極的に利用しようとする研究が行われている．

そこで本論文では，識別的言語モデルで広く用いられている N-best リストの識別的リランキング手法において，長時間にわたって定義される音響的特徴量を利用する手法を提案する．この情報は，これまで利用されていなかった側面の情報であるため，認識精度をさらに向上させられる可能性がある．具体的には，この長時間にわたる音響特徴量として，音声の構造的表象を利用する．

音声の構造的表象とは，話者の違いに非常に高い頑健性を持つ特徴で，これまで孤立単語音声認識や外国語自動発音評価に利用され，効果が示されている．本論文の提案手法は，音声の構造的表象を初めて大語彙音声認識に適用する手法となる．提案手法により，日本語の大語彙音声認識実験の結果，HMM/GMM ベースのシステムから 6.69% の文字誤り率削減を実現することができた．