多重解像度解析を用いた音素分類木と音声超解像への適用

理工学研究科 数理情報学専攻
T14M003 杉井 勇貴
指導教員 馬 青

概要

音声認識の一般的な手法には，HMM や統計的手法がある．これらの手法は，雑音が少ない環境下やよく学習された話者に対しては有効であるが，雑音下や異なる話者に対しての認識精度は高くない．また，HMM や統計的手法は学習データ数が少ないと認識率が著しく落ちてしまう．本研究では，独立した単一日本語音素を入力としたときに，学習データが少なく話者が異なる場合でもその音素，もしくは複数の音素候補を同定する音素分類木を提案する．

音素データには様々な周波数情報が含まれている．ある音素データに含まれる特定の周波数情報のみを用いて，音素を 50 音のいずれかに分類することは極めて困難である．そこで本研究では，Wavelet の多重解像度解析から得られる複数のレベルの周波数情報と，周波数情報に対する主成分分析を行うことで，音素分類木を作成する．1 つの音素データは周波数レベル L と圧縮次元 D によって表現される複数の空間上に表すことができ，これらを L-D 空間と呼ぶ．全ての L-D 空間からある音素が最も分離しやすい L-D 空間を見つけ，そこでその音素を分離していき音素分類木を作成する．

実験では，音素分類木を用いた音素を同定する手法と，SVM や多重解像度解析のテンプレートマッチングを用いた音素の同定手法との比較を行った．評価は，音素を正しく同定した音素の数で行った．音素分類木から単一音素を同定した場合，識別率 7%程度で比較対象とした手法と大きな違いは見られなかった．しかし，音素分類木の深さ 1 で複数の音素候補を同定するための識別をを行った結果，40%程度の候補識別率を得ることができた．次に，多重解像度解析から得られるレベルを制限した際，識別率にどのような変化が生じるかを実験した．多重解像度解析から得られるレベルを制限することは，入力音素のサンプリング周波数が変更されることと等しい．人が音素識別不可能なサンプリング周波数 1000Hz の音素が入力された場合の識別率は，16000Hz の音素が入力された場合と同等かそれ以上の識別性能が確認できた．

提案された音素分類木は，ダウンサンプリングされた音声データの音素識別に対しても有効であることが示された．そこで，ダウンサプンリングされた音素データに対して，音素分類木を用いることで音素を同定し，その音素が本来持っていた高周波数成分を補う音声超解像を提案する．

音声超解像の実験では，話者の異なる同一音素のデータを用いる．ダウンサンプリングされたある話者の音声に対して，異なる話者の音声の高周波数成分を多重解像度解析の再構成の性質を用いて補間する．評価は，フレーム補間 (NN 法，Linear 法) を行った音素のスペクトログラムと主観的な評価を用いて比較を行った．フレーム補間された音素のスペクトログラムは，本来持っていなかった高周波数成分も補間してしまい，主観的な評価でも音素を聞き取ることができなかった．一方，音声超解像を行った音素のスペクトログラムでは高周波数成分がもっともらしく補間されており，主観的な評価でも音素を聞き取ることができた．しかし，ダウンサンプリングを施した音声の話者らしさ，あるいは高周波数成分を補った音声の話者らしさなどの話者固有の特徴は復元されていなかった．