音声研究会に参加して

杉 井 勇 貴
Yuki SUGII
数理情報学専攻修士課程　2 年

1. はじめに

2015 年 3 月 1 日～3 月 2 日に石垣島の南の美ら花ホテルミヤヒラにおいて開催された音声研究会に参加し，「音声超解像のための Haar-Wavelet を用いた劣化音素分類の検討」という題目でポスター発表を行った．

2. 背景

超解像とは，主に画像処理で用いられる補間技術の一つである．われわれは，この超解像処理を音声データに適用し，人が言語音素を識別できないような劣化音声データを，音素識別可能なレベルまで超解像することを目指している．劣化音素を適切に超解像するためには，その音素が劣化以前にはどの音素として発話されたものであるかを音声特定し，その各音素に応じた超解像処理を施す必要がある．

そこで本報告では，Haar-Wavelet による多重解像度解析で得られる複数の解像度レベルと，それを主成分分析によって様々な次元に圧縮した情報を用いることで，劣化音素を分類するための決定木を生成する方法を提案する．

3. Haar-Wavelet の多重解像度解析

代表的な時間周波数解析の手法に，短時間フーリエ変換があるが，不確定性原理より時間分解能と周波数分解能はトレードオフとなる．そこで検出したい周波数によって窓の大きさを変更する Wavelet 変換が用いられる．本報告では，Haar-Wavelet の Scaling 関数 $\phi(t)$ と Wavelet 基底 $\psi(t)$ を用いた多重解像度解析を用いる．

Haar-Wavelet による任意の関数 $f(t)=f_0(t)$ の Wavelet 変換は，$\phi(t)$ で級数展開を行い，任意の関数に対しての近似する．このとき，L をレベルと呼び，レベル L が大きいほど近似が細かく，小さいほど近似が粗い．

$$f_L(t)=\sum_k c_k^{(L)}\phi(2^{-L}t-k) \tag{1}$$

このとき，$c_k^{(L)}$ がレベル L での近似の結果となる．

レベル L の $f_L(t)$ とレベル $L-1$ の $f_{L+1}(t)$ との差を，レベル $L+1$ の $g_{L+1}(t)$ とおくと，$g_{L+1}(t)$ は Wavelet 基底 $\psi(t)$ の級数展開で表される．このとき，係数 $d_k^{(L-1)}$ がレベル $L-1$ での時間周波数解析の結果となり，レベルが小さいほど高周波数成分を，レベルが大きいほど低周波数成分を表す．

$$\begin{aligned} g_{L+1}(t) &= f_L(t)-f_{L+1}(t) \\ &= \sum_k d_k^{(L+1)}\psi(2^{-L-1}t-k) \end{aligned} \tag{2}$$

このとき，$f_L(t)$ から $f_{L+1}(t)$ と $g_{L+1}(t)$ を導くことを分解，$f_{L+1}(t)$ と $g_{L+1}(t)$ から $f_L(t)$ を導くことを再構成と呼ぶ．

$f_L(t)$ の分解を繰り返すと，最終的に直流成分 $c_0^{(L-M)}\phi(2^{-L+M}t)$ と $\psi(t)$ の級数展開との和で表すことができる．

$$\begin{aligned} f_L(t) &= f_{L+M}(t)+g_{L+1}(t)+\cdots+g_{L+M}(t) \\ &= f_{L+M}(t)+\sum_{i=1}^{M} g_{L+i}(t) \\ &= \sum_k c_k^{(L+M)}\phi(2^{-L-M}t-k) \end{aligned} \tag{3}$$

$$\begin{aligned} &+\sum_{i=1}^{M}\left\{\sum_k d_k^{(L+i)}\psi(2^{-L-i}t-k)\right\} \\ &= c_0^{(L+M)}\phi(2^{-L-M}t) \\ &+\sum_{i=1}^{M}\left\{\sum_k d_k^{(L+i)}\psi(2^{-L-i}t-k)\right\} \end{aligned} \tag{4}$$

このように様々なレベルで任意の関数 $f(t)$ を表すことを多重解像度解析と呼ぶ．ただし，M は Wavelet 基底 $\psi(t)$ が直流成分を表すまで分解したレベルを表すための変数とする．

4. 劣化音素分類のための決定木

人が認識不可能な劣化音素データを超解像するには，劣化音素デーが本来どの音素であったかを同定する必要がある．本報告では，音声超解像を行うために学習データから決定木を生成する手法を提案する．

4.1 決定木の作成手順

以下に提案する決定木作成の手順を示す．

1．あらかじめ音素をラベル付けした学習データに対して Haar-Wavelet の多重解像度解析を行う．
2．各レベルに対しての $\psi(t)$ の係数ベクトルを求める．このとき，レベル L の係数ベクトルを $d^{(L)}$ とする．
3．$d^{(L)}$ を主成分分析し，これを D 次元で表したベクトルを $d^{(L,D)}$ とする．
4．全ての $L-D$ 空間で全音素の分離度を求める．
5．もっとも高い分離度をもつ音素から順に，パラメータ $L-D$ で分離しその音素を枝とする．

4.2 分類木による分類手順

1．入力データに対して Haar-Wavelet 変換の多重解像度解析を行う．
2．分類木の各枝の分離パラメータ $L-D$ にしたがって，$d^{(L,D)}$ を求める．
3．$L-D$ 空間上で音素を分類することを順に繰り返し，葉に辿り着くまで繰り返す．

5. 実験と結果

本報告で提案した決定木かあら日本語劣化音素の分類を行い，SVM での分類との比較を行う．音声データは，男性話者 8 名の 45 音の日本語音素 16000 Hz データを用いた．前処理として，音声データは単一音素に区切り，値が最大になる時間を t_{max}，最小となる時間を t_{min} とし，$\frac{1}{2}(t_{max}+t_{min})$ が一致するように時間シフトし正規化を行った．さらに，全ての音素のピッチを一定に揃えた．

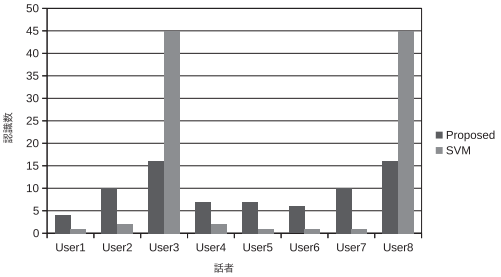

図 1　User 3 と User 8 を学習データとした時の識別数

劣化音素は 16000 Hz のデータを音素識別不可能な 1000 Hz まで間引いたデータとする．テストデータの分類はノード内の $L-D$ 空間での重心の距離でノードを決定し分類を行う．SVM での学習データの選択は，1000 Hz の学習データを超平面で分類し，テストデータが属する音素を劣化音素の分類とした．

5.1 実験

8 人の音声データから 2 人選び，2 人 1 セットの学習データとし，学習データに対して 8 人の 45 音をテストデータとし分類した．ただし，提案手法での評価は，木の深さ 1（クラス数 8）での正答率とする（図 1）．

図 1 から，提案手法と SVM を比較すると，認識数は学習データがテストデータと同じ場合を除くと，提案手法の方が高いことが示された．提案手法において，学習データがテストデータと同じ場合にも関わらず，認識率が 100% となっていない．これは，決定木の生成過程でノード内の音素の重心が移動しているためだと考えられる．

6. おわりに

実験を行うにあたり，音声録音に協力していただいたみなさま，研究に関してご指導いただいた佐野彰先生に深く御礼申し上げます．