ロバスト音響モデルと連続音声認識システムに関する研究

大貫和永 *

Study on the Design of Automatic Continuous Speech Recognition System with Robust Acoustic Models

Kazunaga OHNUKI

1 はじめに

平成 21 年 9 月 25 日，北海道大学より博士 (情報科学) の学位を授与された。学位論文は連続音声認識で利用する音素単位のデータベースである音響モデルのロバスト性を向上させる方法について報告している。

本研究により，雑音が含まれた音声に対する認識率が 3%程度向上したロバスト音響モデルの構築が可能となった。

2 連続音声認識システム

連続音声認識システムは図 1 に示すように音声の最小単位である音素の特徴を記憶した音響モデルと，音素の並びと認識対象単語の対応表である単語辞書，日本語の言語的特性として単語のつながり規則が記録されている言語モデルの情報を基に，入力される発話音声を文字情報に変換するシステムである。

音声が発声される際，周囲で発生する環境音は雑音となり，認識結果に著しい悪影響をあたえるため，雑音に強い認識システムを構築することが，音声認識を実用化するために求められている。

日本語を表記する際に必要となる音素は 43 個であるが,「赤」を発声する場合の aka と「秋」を発声する場合の aki では中心音素 k の音が変化する。このように前後の音素により音が変化する特徴があるため，音響モデルには 43 の 3 乗個の音素の特徴が必要である。

日本音響学会では 39 の研究機関の協力により，男女各 153 名が各 200 文を読み上げた「新聞記事読み上げ音声コーパス」(JNAS)[1] が提供されている。これを学習データとすることにより，雑音の無い環境で 90%以上の精度で認識が可能な音響モデルを構築できる。

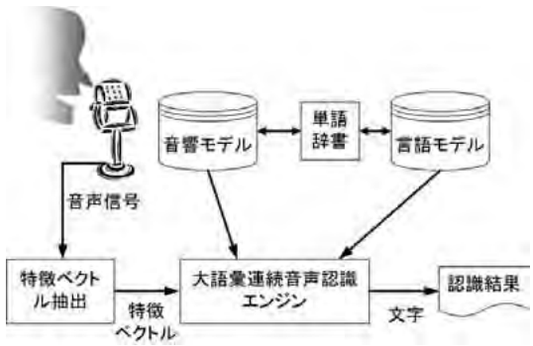

図 1: 大語彙連続音声認識

本研究では音響モデルを学習する際に，学習データである JNAS に対して次節で説明する雑音低減技術 RSA を適用することによりロバスト性をもった，ロバスト音響モデルの構築が可能となった。

3 雑音低減技術：RSA

発話データを 25[msec] の短区間毎に周波数成分を分析して，MFCC 特徴ベクトルと呼ばれる特徴量を抽出する。この短区間を 10[msec] 毎にシフトして発話全体を分析することにより発話内容を認識する。

各短区間毎に得られた MFCC 特徴ベクトルの 5 番目の値の時系列変化を図 2 に示す。そのスペクトルは図 3 である。図 4 は，図 3 の高域を削減したものである。これを逆フーリエ変換して図 5 が得られる。得られた時系列変化は元の図 2 に比較して滑らかになっており雑音成分が低減されている。これを Running Spectrum Analysis(RSA) と呼ぶ。

JNAS には新聞記事から選ばれた発話時間 4 秒程度の文を発声した男女各 23,651 発話のデータ含まれているが，これら発話データから RSA を掛けた MFCC

*釧路高専情報工学科

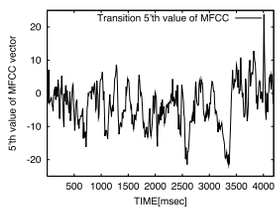

図 2: 時間変化

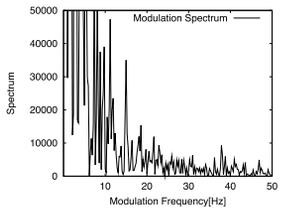

図 3: スペクトル

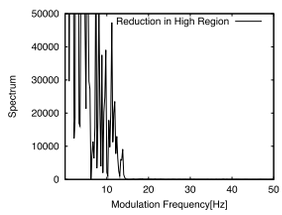

図 4: 高域低減

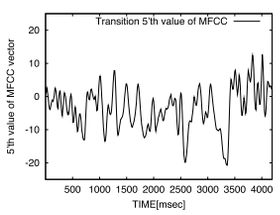

図 5: 滑らかになった図 2

特徴ベクトル，RSAMFCC 特徴ベクトルを抽出し学習することにより，本研究で提案するロバスト音響モデルを構築できる。

4　性能評価

音響モデルと認識する特徴ベクトルの組み合わせを，表 1 として認識実験を行った。

雑音データベース NOISEX-92[2] の雑音を 3 種の SNR で付加した音声を認識した。表 2 より組み合わせ B が最も認識率が高く，ロバスト音響モデルで標準 MFCC 特徴ベクトルを認識すると雑音に強い認識ができ，提案の音響モデルはロバストである。

表 1: 音響モデルと認識データの組み合わせ

	RSA	標準
ロバスト音響モデル	A	B
標準音響モデル	D	C

表 2: 雑音を含んだ音声の認識結果

組み合わせ	A	B	C	D
男特 SNR=20dB	76.33	79.19	76.71	69.65
男特 SNR=15dB	57.02	61.64	57.91	50.43
男特 SNR=10dB	35.89	40.15	37.07	31.39

5　ロバスト音響モデルの音素間距離

ロバスト音響モデルの各音素間のマハラノビス距離を計算し，標準の音響モデルと比較した。

図 6 に相対累積頻度を示す。RSA 処理を行ったロバスト音響モデルの音素間距離が大きい。これにより雑音に強い認識ができる。

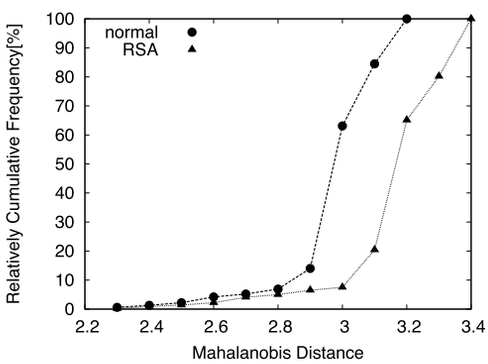

図 6: 音素間マハラノビス距離の相対累積度数

6　むすび

提案した音響モデルがロバスト性を持つことを示した。認識する信号に対しては RSA 処理は不要であり，認識速度には影響を与えず高速な認識を行える。

7　謝辞

北海道大学大学院情報科学研究科メディアネットワーク専攻情報通信システム学講座通信ネットワーク研究室の宮永喜一教授にご指導をいただいた。また，岸浪建史校長，情報工学科の同僚をはじめ，本校の教職員には，一方ならぬご配慮・ご援助をいただいた。併せて感謝します。

参考文献

[1] http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/

[2] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html