音声変換を用いた方言音声認識のための音韻特徴分析
Analysis of phoneme features for dialect speech recognition by transforming

石垣将太
Shouta Ishigaki
法政大学情報科学部コンピュータ科学科
E-mail: shouta.ishigaki.dp@cis.hosei.ac.jp

Abstract

Today's speech recognition system is built on a database of standard Japanese, so its support is not enough to a dialect voice. In this study, we investigate the acoustic characteristics of dialect voices to solve this problem. The purpose of this study is to improve the accuracy of speech recognition by transforming dialect voice into standard Japanese based on this characteristic. However, the characteristics of the dialect voices are different in each region, so an object is only a Sendai dialect. The main characteristics of the Sendai dialect voices are to confusion vowels, between /i/ and /e/, between and /i/ and /u/, changing unvoiced phoneme to voiced phoneme, changing voiced phoneme to voiced nasal phoneme. In this study, we used two methods, linear predictive coding and line spectrum pair, to analyze formant frequency of the vowels. As a result, the characteristics of the vowel /i_e/ were a new second formant which was 500Hz to 1000Hz. And the formants of the vowel /i_u/ were near in frequency band of /u/. Therefore, the confusion of vowels was caused by middle vowel /i_e/ and changing vowels from /i/ to /u/. However, it is difficult to identify vowels including the middle vowel by only formant frequencies, so it is required to analyze other features.

1 まえがき

近年音声認識を用いた製品は、スマートフォン、カーナビゲーション、一部の自動販売機など私達の生活に近いところに存在している。音声認識はキーボードのような複雑なボタン操作を必要とせず、機械操作が苦手な高齢者にも有用で高齢化社会を迎える日本において発展が必要とされる技術であるが、各地方の高齢者には日常で方言を使用している人も多く存在する。製品化されている音声認識システムは標準語のデータベースを基に構築されているので方言への対応が十分ではなく、方言話者が音声認識システムを利用する際には無理に標準語を発話する必要があり、利用者側に負担が発生してしまう。

解決法として標準語を基に構築されたモデルを特定話者に適応させる手法 [1] も採られているが、音声認識システムの使用者が多数である場合はこの手法は不適切である。また話者適応の際にユーザーが訓練として文章を読み上げ、システムに学習させる必要があるが、方言話者には高齢者が多いため音素バランス文などを正しく読み上げることが難しく、方言音声学習作業においても大きな負担が発生してしまう。

本研究で提案するシステムの目的はこの問題を解決するために方言音声を分析された特徴量を基に合成し標準語へ変換を行った後に認識を行うことで音声認識システム利用時の方言話者の負担を軽減することである。しかし、方言は地方ごとに存在しそれぞれの特色も大きく異なる。そのため本研究では対象を東北地方宮城県に存在する仙台弁に限定した。

Supervisor: Prof. Katunobu Itou

2 音声合成変換を用いた音声認識システム

本研究で提案する音声認識システムは通常の認識を行う前に、分析された特徴量に基づき方言で標準語から変化する音韻特徴が含まれる音素を検出する。更に検出された音素に対して方言の特徴を除去する音声合成を行い標準語へ変換した後に音声認識を行う。

しかし方言は地方ごとに特色が異なるためサンプルデータ量の肥大化・分析の複雑化・検出時間の増大・方言の誤解釈などの問題があり、全ての方言に対応したシステム構築は不可能である。そのため本システムでは対象とする方言を仙台弁に限定し、音素検出・音声合成変換には仙台弁の特色として挙げられる音韻要素から母音/i/と母音/e/の混同（/i_e/）・母音/i/と母音/u/の混同（/i_u/）・特にカ行とタ行においての無声音の有声音化・濁音の鼻濁音化を用い、仙台弁母音/i_e//i_u/から通常の五母音への変換・有声音から有声成分除去・鼻濁音から鼻音成分除去の３ステップで変換を行う必要がある。また、本手法では標準語から方言へ変化する音響要素を別々に分析・検出・音声合成変換しているため、一つの方言で存在する音響要素に対応できれば、他の方言に存在する同様の音響要素にも対応することができると考えられる。

本研究では中音韻的特色の母音/i_e/、母音/i_u/について分析を行った。分析する特徴量としてフォルマント周波数を用いた、フォルマント周波数は母音により分布が異なるため通常の五母音識別に利用されていると共に、出雲弁 [2]、会津弁 [3]、名古屋弁 [4] で特色とされる特殊母音特徴分析においても用いられている。同様に五母音/a//i//u//e//o/、母音/i_e//i_u/の各フォルマント周波数を抽出、これを特徴量とし各母音と比較検証した。また、フォルマントを変換するフィルタを使用した音声合成により認識結果が改善されるか検証した。

しかし母音/i_e//i_u/を含めて母音識別を行うには各フォルマント周波数のみでは困難であったので、今後は新たな実験としてフォルマント周波数平均値からの距離や第 1・第 2・第 3 フォルマントの距離、帯域フィルタから出力される音声の各帯域パワー、といった特徴量を用いた実験を行い識別可能か検証する。

本研究では標準語から方言へ変化する要素である音韻変化・文法変化・単語変化の中から、音韻変化による音声認識精度低下について解決手法を提案した。文法変化・単語変化については音声認識システムの言語モデルに方言特有の文法・単語を新たに定義し追加することで対応できると考えられ、本研究手法と組み合わせることで方言に対応した音声認識システムを構築できると考えられる。

3 方言と仙台弁の音韻特徴

方言とは、地域社会ごとに見られる言語体系である。標準語から変化する要素として音韻・アクセント・文法・語彙などが挙げられ、日本国内においては一般的な標準語の他にも東日本方言・西日本方言・九州方言・琉球方言などが存在する。

実際は更に細かく分類され全ての方言を分析・特徴量抽出・音声変換するためには地域ごとの膨大な音声データと手法が必要になる。そのため本研究では東北地方宮城県に存在する仙台

弁に限って分析・特徴量抽出・評価を行う。

3.1 仙台弁の音韻特徴

仙台弁はいわゆる「ズーズー弁」と言われる東北地方方言の一つである。仙台弁の特色のうち音響的特色を以下にあげ、例として変化した単語と対応する標準語の単語をあげる [5][6][7]。

1. 無型アクセント（アクセントの決まりを持たず平坦な音調になる）を最大の特徴とする南奥方言に属し、アクセントで単語を区別しない
2. 母音/i/は共通語よりも広い母音で発音、/e/と似通った発音になる
3. 「シ・ス」「チ・ツ」「ジ・ズ」「ニ・ヌ」「ノ・ヌ」が混同
   例：ナス（茄子、梨）　クズ（口、靴）
4. 「シュ・チュ・ジュ」が「ス・ツ・ズ」になる
   例：テース（亭主）　スズツ（手術）
5. 「キ・ギ」が破擦音化し「キ・キャ・キュ・キョ・ギ・ギャ・ギュ・ギョ」が「チ・チャ・チュ・チョ・ジ・ジャ・ジュ・ジョ」になる
   例：チル（切る）　ヤチュー（野球）
6. カ・タ行の子音が語頭以外で有声音化
   例：サカ（坂）　ワフグ（和服）　ニワドリ（鶏）
7. カ・タ行の有声音化に並行してガ・ダ行の子音が鼻音を伴う
   例：アゲル（開ける）　アゲ’ル（上げる）　クギ（茎）　クギ’（釘）
8. 語頭以外のザ・バ行子音はほとんどの場合で鼻音を伴う
9. 鼻音を伴った有声子音が無声音化することがある、鼻音が独立して発音されることがある
   例：ワンツカ（わずか　フンチサン（富士山））
10. 長音、撥音、促音を十分に発音しない
    例：〜ナテ（〜なんて）　ワラタ（笑った）
11. 一音節語は長く伸ばして発音されることがある
    例：キー（木）　ハー（歯）

以上のことから仙台弁の音響特色は

1. 母音/i/と/e/の混同母音（/i_e/）
2. 母音/i/と/u/の混同母音（/i_u/）
3. 無声音の有声音化（特にカ・サ・タ行）
4. 濁音の鼻濁音化

と定義付けられる。

本研究ではこれらのうち仙台弁で現れる母音特徴である母音/i/と/e/の混同、母音/i/と/u/の混同（これらを母音/i_e/、/i_u/とする）の検出、合成のためフォルマント周波数を分析した。

4 音声データサンプル

仙台弁を分析するための音声データサンプルは日本のふるさとことば集成三巻に収録されている音声データを用いる。「仙台の昔の様子、神仏にまつわる話」をテーマとする 3 名の話者の談話音声データが収録されていて、実際に発話された文章と標準語訳が記載されている。

音声データサンプルの談話音声データは音素バランス文などとは異なり、話者が日常生活で使用している方言を自然に発声しているのでこれを分析・検出・音声合成変換に利用することで、実際に話者がシステムを利用する際にも負担の少ない自然な発声で利用することができると考えられる。なおひとりの話者が続けて話し、次の話者に交替するまでの連続した発言を 1 発話とする。

ただし、途中にあいづち・笑い声が入る場合や、複数の話者が同時に話している箇所が含まれている。話者情報と録音環境を表 1 に示す。

文章例（2 発話）

A：ヤムオエズ　ザオーサ　ニサンカイ　ノボッタ　ゴドアンノネー。

A：止むを得ず　蔵王に　2, 3 回　登った　こと [が]　あるのね。

C：コンデ　カミサンモ　ホドゲサンモー　ゴエンノ　アルー　ドゴド　ゴエンノ　ナイ　ドゴー（A　アー）　アラサルモンネ。

C：これで　神様も　仏様も　ご縁の　ある　ところと　ご縁の　ない　ところ [が]　（A　ああ）　ありなさるものね。

表 1　サンプル話者情報と録音環境

話者A	75 歳：男性
話者B	71 歳：男性
話者C	67 歳：女性
収録日時	1977 年 11 月 8 日
収録地点	宮城県仙台市八幡
総発話数	273
発話内訳	A：111　B：70　C：90
収録時間	22 分 04 秒
サンプリング周波数	22.050kHz
量子化ビット数	16bit
ファイル形式	RIFF
チャンネル	2（ステレオ）

カタカナの文章は実際に発話された仙台弁音声を文章に書き起こしたものであり、併せて標準語に直した文章を記載している。() 内は発話中のあいづちを表し、[] 内は仙台弁では発声されず標準語で補われる部分を表す。

4.1 サンプル中に含まれる音素のカウント

サンプル中に含まれる分析に用いる音素数を調査する。母音の混同（/i_e//i_u/）が発生している箇所と母音/a//i//u//e//o/が含まれる箇所をカウントした。ただし五母音は母音が主観により判別できる箇所のみをカウントした。

また「でげもの（出来物）」のように母音/i_e/と無声音の有声音化が同一音素に現れている場合があるがこれも含めてカウントを行った。カウント結果を表 2 に示す。

カウントの結果、同じ単語が繰り返し発話される場合が多数見られた。これはサンプル音声が談話を録音したデータであり、話の流れに沿った単語が繰り返し発話されたためだと考えられる。混同母音/i_e/は 46 箇所中で 41 箇所が本来標準語では母音/i/で発声されるべき音素が/e/に近付いた音素（「フルイ」から「フルエ」に変化など）で発声され、反対に母音/e/から母音/i/に近付いた音素（「エキ」から「イギ」など）は 5 箇所のみであった。混同母音/i_u/は 79 箇所中全てが本来標準語では母音/i/で発声されるべき音素が母音/u/に近付いた音素（「コクブンジ」から「コクブンズ」など）で発声されていた。

表 2　サンプル中に含まれる音素数

混同母音/i_e/	46
混同母音/i_u/	79
母音/a/	45
母音/i/	27
母音/u/	29
母音/e/	32
母音/o/	26

4.2 音素の切り出し

分析のためにサンプル全体から各音素（/i_e//i_u//a//i//u//e//o/）を切り出す。まず切り出す音素が現れている発話から一文程度の長さで切り出し、その後で Julian を用いて音素アライメントを出力する。出力された音素アライメントとフレームにより音素の位置を確認した後に音素の切り出しを行う。

4.2.1 Julian による音素アライメント

Julian はフリーで提供されている大語彙連続音声認識システムである。これは通常の音声認識結果だけではなく音素アラ

イメントの認識結果も出力できる。これを使用して切り出す音素の位置を特定する。提供されている音響モデルと言語モデルと発話に合わせた文法を用いて音素アライメントの認識を行った。

Julian の文法ファイルを次の表 3 のように変更し、Voca ファイル中の HATSUWA の内容を各発話文章に変更することで各発話内容がより高い精度で認識される。なお、表中の NS_B、NS_E は文頭・文末の無音区間に対応している。

表 3　発話文章と文法ファイル例

発話文章	ソエツ　ノボッテ　イグ　ドネ
Grammer	S : NS_B HATSUWA NS_E
Voca	HATSUWA：s o e ts u n o b o q t e i g u d o n e

認識結果から音素の位置を特定する。標準設定ではフレーム長 400 点・フレームシフト長 160 点となっていて音素に対して窓長が長いため、認識された音素位置がずれてしまった。そのためフレームシフト長を 80 点に変更し認識を行った。以上のように Julian で認識された結果を基に最終的には人手で切り出す範囲を調整して切り出しを行う。

5　母音/i_e/・/i_u/フォルマント分析

サンプル音声データより切り出された各母音を分析した。母音の認識を行う場合、日本語の五母音/a//i//u//e//o/は第 1・第 2 フォルマントによって識別される。各方言の母音を扱った先行研究 [3] でも同様にフォルマント周波数を用いているため、本研究でも日本語五母音と仙台弁の特殊な混同母音/i_e//i_u/の特徴量としてフォルマント周波数を用いた。

5.1　フォルマント周波数抽出

人間の声は声帯によって発声した音波が声道の形状によって共振し、調音変化したものである。共振で特定帯域ごとに増幅された音がスペクトルのピークとなり、これをフォルマントと呼ぶ [8]。フォルマントは低周波から第 1、第 2、第 3 フォルマントとよばれる。

スペクトル包絡を計算するために次数 14 の線形予測（LPC）分析を用いた。導かれたスペクトル包絡からピークを検出し、フォルマント周波数を抽出した。

5.2　フォルマント周波数抽出結果例

混同母音/i_e/のフォルマント周波数を抽出した結果を図 1、母音/i/のフォルマント周波数を抽出した結果を図 2 に示す。図中の赤丸が検出抽出されたフォルマント周波数である。

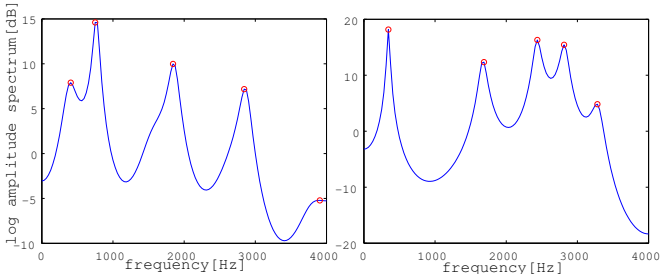

図 1　フォルマント周波数 (/i_e/)　図 2　フォルマント周波数 (/i/)

5.3　フォルマントの時間変化

音声は時間により変化する信号であり、フォルマントも時間により変化するため音声波形をフレームごとに分析する必要がある。フレーム長 256 点・フレームシフト長 128 点としてフォルマント周波数抽出を行い、線スペクトル対係数（LSP）によるフォルマント推移分析を行った。

線スペクトル対係数はフォルマント近傍に位置し、線形予測係数に比べて時間方向の変化が滑らかで補間特性に優れるためこの変化を追従することで、あるフォルマントがその前後でどのフォルマントに対応するのかを明確にすることができる。

サンプルとして「毎年（ma-/i_e/-ne-n)」と発声した音声で分析を行った例を図 3 に示す。赤破線が線スペクトル対係数で、黒線で囲われている範囲が母音/i_e/に対応している。

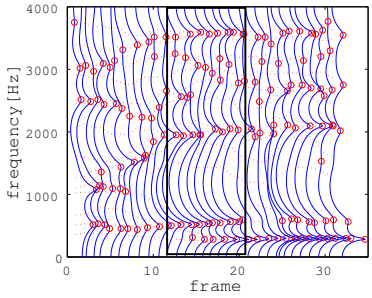

図 3　「ma-/i_e/-ne-n」フォルマント時間変化分析結果

5.4　母音フォルマント周波数分析結果

各母音の第 1 フォルマント・第 2 フォルマント・第 3 フォルマント周波数を分析・抽出し、算出された平均値を表 4 に示す。

表 4　各母音のフォルマント周波数平均値

母音	第 1(Hz)	第 2(Hz)	第 3(Hz)	第 1'(Hz)
/i_e/	427.30	1858.01	2535.32	669.27
/i_u/	416.53	1138.84	1765.82	-
/a/	580.46	1043.34	2423.44	-
/i/	367.96	1653.90	2312.50	-
/u/	340.62	1346.87	1767.96	-
/e/	521.09	1477.34	2347.65	-
/o/	469.53	919.53	2208.59	-

5.5　フォルマント周波数分布

フォルマント周波数の分布を作成する。混同母音/i_e/の第 2 フォルマントを除外し抽出された本来の第 3 フォルマントを第 2 フォルマント、第 4 フォルマントを第 3 フォルマントとして作成した。

各フォルマント周波数を軸としたフォルマントの 2 次元分布図を図 4、図 5 で示し第 1・第 2・第 3 フォルマントの 3 次元分布図を図 6 で示す。

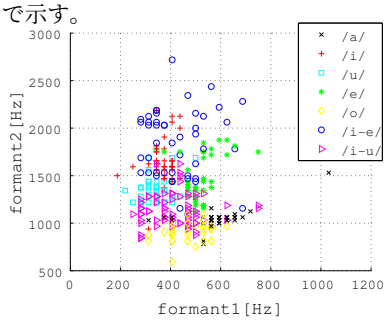

図 4　/i_e//i_u//a//i//u//e//o/分布（第 1・第 2 フォルマント）

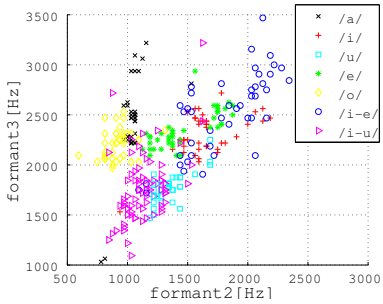

図 5　/i_e//i_u//a//i//u//e//o/分布（第 2・第 3 フォルマント）

5.6　混同母音/i_e/のフォルマント周波数考察

46 個の母音/i_e/のフォルマントを分析した結果、そのうち 37 個に図 1 の 750Hz 付近、図 3 の黒枠内 500Hz 付近で現れているような第 2 フォルマント（これを第 1' フォルマントとする）が存在した、出現する帯域は 500Hz から 1000Hz で平均値は 669.27Hz となった。これは図 2 の 750Hz 付近においては現れない新たなフォルマントであり、同様に母音/e/でも現れないため母音/i_e/と母音/i//e/を分ける独自の特徴と言える。

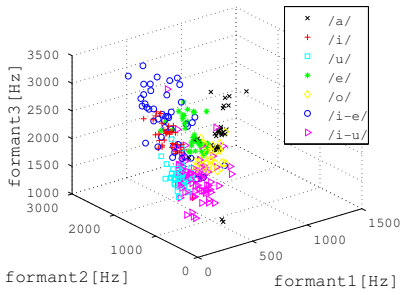

図 6　/i_e//i_u//a//i//u//e//o/分布（第 1・第 2・第 3 フォルマント）

この結果から、仙台弁における母音/i/と/e/の混同とは第 1・第 2・第 3 フォルマントがこれら 2 母音の帯域に広く分布し、更に第 1' フォルマントが 670Hz 程度の帯域に生成された新しい中間的母音の存在が原因であると言える。

5.7　混同母音/i_u/のフォルマント周波数考察

79 個の母音/i_u/のフォルマントを分析した結果、図 4（F2:1500-2000Hz）、図 5（F2:1500-2000Hz、F3:2000-2500Hz）のように母音/i/の周波数帯域では分布数が少なくなっていて、母音/u/の周波数帯域に分布している。特に図 5 ではそれがより顕著に表れている。

また、実際に母音/i_u/が現れている単語は「イシムロ（石室）：イスムロ」「コッチカラ：コッツカラ」のように母音/i/から/u/へと変化しているものが全てであり、母音/u/から/i/へ変化したものは無かった。

以上から仙台弁における母音/i/と/u/の混同とは母音/i_e/のように母音/i/と/u/の中間音によるものではなく、母音/i/が母音/u/へ変化したことが原因であると言える。

5.8　混同母音の前後音節・単語中の音節位置

混同母音は単独で発声されるのではなく、他の音素と連続して発声される。そのため混同母音の発声は前後の母音に影響される可能性がある。これを検証するため混同母音/i_e//i_u/の前後の母音、単語中の音節位置を調査した。

その結果混同母音の前音節は/a//i/が多く、特に/i_e/は半分以上 (26/46) が前に/a/を持ち、/i_u/は前音節に/a/(25/79)、/i/（25/79）を多く持つことが分かった。また、単語中の混同母音は後音節に位置していることが分かった。

以上から混同母音は前の音素による影響が強く特に/i_e/の前に母音/a/が存在する場合、/i_u/の前に母音/a//i/が存在する場合についてフォルマント時間変化を分析する必要があると考えられる。

6　/i_e/音声合成変換

/i_e/の分析結果として 650Hz 周辺に新たなフォルマント（第 1' フォルマント）の存在が分かった。これを除去し/i_e/を/i//e/へ合成変換することで認識精度向上を図った。

合成のためにノッチフィルタを作成した、ノッチフィルタとは特定帯域を減衰させるフィルタの一種である。ノッチフィルタは減衰帯域が狭くその他の帯域の成分に与える影響が少ないという利点があり、この音声合成においては第 1' フォルマントのみを減衰させることが目的のため使用した。ノッチフィルタの振幅応答を図 7 に示す。

サンプルとして「マエ（i_e）ネン」と発声された文章を合成した例を図 8 に示す。青線が加工前、赤線が加工後のスペクトル包絡である。650Hz 付近の第 1' フォルマントが除去されていることが確認できる。

Julian でマ [イ] ネン、マ [エ] ネン、マ [ア] ネンのように加工箇所で 5 母音を定義した文法を用いて認識結果を比較した。その結果加工前音声では「マエネン」と認識されていたものが、加工後は「マイネン」と正しく認識されていた。

このように/i_e/において現れた第 1' フォルマントをノッチフィルタにより除去し音声変換を行うことで音声認識結果が改善されたため、本手法により仙台弁から標準語への変換が可能であると言える。

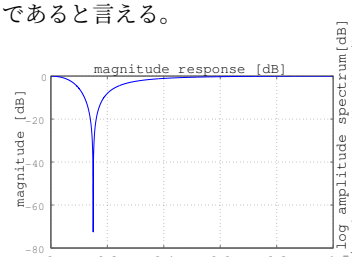

図 7　ノッチフィルタ

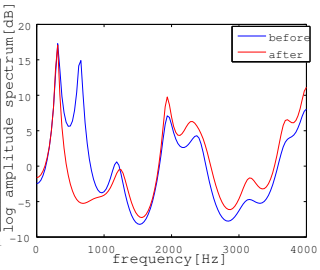

図 8　音声合成（マエネン：マイネン）

7　あとがき

本研究では仙台弁の特色が現れている音素を検出し仙台弁音声を合成し標準語へ変換する機能を持つ音声認識システムを提案し、その前段階として必要な仙台弁の音韻特徴分析を行った。仙台弁の音韻特徴から母音の混同（/i_e/・/i_u/）についてフォルマント周波数を用いた分析を行い、混同母音/i_e/では結果として通常の母音/i//e/では現れない新たな第 1' フォルマントが 500Hz から 1000Hz の周波数帯域に存在し、第 1' フォルマントを除外した分布では母音/i//e/の帯域に広く分布していることが確認された。また、第 1' フォルマントの出現確率は 80.43 %となった。

また混同母音/i_u/は各フォルマントが母音/u/に近い帯域で分布していると共に主観認識では母音/u/と判断できることから、混同母音/i_e/のように母音/i/と/u/の「中間的母音」ではなく母音/i/から/u/へ変化している音素であると言える。

以上の結果から母音/i/と/u/の混同には単語辞書データに「石（イス）」のような単語を追加することで対処できるが、中間的母音/i_e/には対処できない。そのため/i_e/への対処としてノッチフィルタを使用し音声合成変換（第 1' フォルマント除去）を行った。例として「マエネン（毎年）」のエ（i_e）箇所を変換し、音声認識結果を「マエネン」から「マイネン」に改善することができた。以上から本研究の手法により方言による音韻変化に対応できると考えられる。

しかし、システム自動化に必要である中間的母音/i_e/検出のためにはフォルマント周波数だけではなく更なる特徴量が必要である。よって今後は周波数帯域ごとのパワー [8]、フォルマント同士の距離などを用いた分析を行い、仙台弁音声から中間的母音/i_e/を検出することが可能であるか検討する。

また仙台弁で変化する音韻特徴には母音変化の他にも、無声音の有声音化、濁音の鼻濁音化が存在する。これらにも対応した音声認識システム構築のため、今後は無声音・有声音特徴分析、濁音・鼻濁音特徴分析が必要である。

参考文献

[1] Dong YANG Koji IWAO Sadaoki FURUI, "Accent Analysis for Mandarin Large Vocabulary Continuous Speech Recognition", 社団法人電子情報通信学会, 2008-03

[2] 吉廣綾子 岸江信介 大山玄, "出雲方言における中舌母音の音響的特性について", 徳島大学, 言語文化研究 14, 203-218, 2006-12

[3] 坂本通治, "会津地方における各方言のフォルマント分析", 日本方言研究会発表原稿集 90 回, 2010-09-16

[4] 今西由華 林陽子 金森康和 犬養隆, "名古屋弁音声の特徴抽出と分析", 社団法人電子情報通信学会, 研究報告.SP, 音声 106(263), 31-36, 2006-09-19

[5] "全国方言談話データベース　日本のふるさとことば集成　第 3 巻　宮城・山形・福島",　国立国語研究所資料集 13-3, 国書刊行会, 2006-05-31

[6] 大橋純一, "東北方言音声の研究", 國語學 54(3), 145-150, 日本語学会, 2003-07-01

[7] 三輪 譲二, "現代日本語方言音声の音響分析", 日本音響学会誌 51(11), 893-898, 1995-11-01 社団法人日本音響学会

[8] 齋藤寿樹, "周波数帯域パワーとフォルマント周波数を用いた母音認識の初歩的研究", 島根大学総合理工学部数理・情報システム学科田中研究室 2004 年度卒業論文