Vol.2016-HCI-167 No.10
2016/3/9

方言の音声認識モデル構築に向けた音声データ収集の効率化検討

吉田裕範[†1] 松本和芳[†1] 関義則[†1] 榑松理樹[†2]

概要：本稿では，方言音声認識システムの学習用音声データのラベル付け作業の効率化について述べる．ラベル付け作業とは，現地で収録した方言音声データへカナ文字を関連付け機械学習のための識別情報を作成することである．従来のラベル付け作業は，音声解析ソフトで音声波形を図示した後は，手作業が大部分を占め，作業効率が悪い．今回，ラベル付け作業負荷の軽減，時間短縮を図る手法として，従来手法のように事前に音響モデルを事前準備することなく，音素開始時間に関連する特徴量を算出し，音素の区切り位置となる時間をラベル付け作業者に対して情報提示する手法を検討した．これにより，方言音声に関する知識や経験の無い作業者でもラベル付け作業を効率よく進められる仕組みを整備する．

キーワード：音声情報処理，音声音響データベース，方言音声，ラベル付け

A Study of an Improvement Approach to Collect Speech Data for a Dialect Speech Recognition Model

HIRONORI YOSHIDA[†1] KAZUYOSHI MATSUMOTO[†1] YOSHINORI SEKI[†1]
MASAKI KUREMATSU[†2]

Abstract: In this paper, we describe the efficiency of labeling work of training speech data of dialect speech recognition system. The labeling work of training speech data, is to create a information associated with the Kana character to the dialect voice data was recorded in the field. The conventional labeling work, after illustrating the speech waveform in voice analysis software, manually accounted for a large part, it is poor work efficiency. This time, reduce labeling workloads, as a method for shortening time without Prerequisite an acoustic model in advance as in the conventional method, and calculates a feature amount associated with the phoneme start time, a phoneme segmentation positions It was investigated a method to information presentation time against labeling workers.As a result, to develop a mechanism to be efficiently advance the labeling work even in the absence of the worker of knowledge and experience about the dialect voice.

Keywords: Speech information processing, Acoustic-phonetic Speech database, Dialect voice, Labeling

1. はじめに

現在,音声認識技術は,精度や処理速度の向上とともに様々な場面で活用されている．しかし，現在の音声認識技術は標準語を中心としており，方言認識を行ううえでは，多様な方言に対する汎用性，少ない方言言語資源で動作，方言変換・言語理解との接続容易性といった点を考慮しなくてはならず，結果として方言を限定した研究が多く，決定的かつ汎用的な手法はまだ確立していない．

一方で，近年，平均寿命の伸びや出生率の低下により少子高齢化が急速に進んでいる[1]．この少子高齢化により，全国的な福祉サービスの需要増大が見込まれている．筆者らの住む東北地方では，東日本大震災の影響で，被災地へ訪れた都市部の医療・介護従事者がほとんど方言を理解できず支障が出たという実例がある[2]．このことを考慮に入れると，将来的に若者に対する高齢者の比率が極端に大きくなるという予測から，都市部に住む若い従事者が，このような福祉ニーズに対応するケースでこの問題はますます深刻になると想定される．

語学として，新たに習得する言語の発音，意味，アクセント，これらすべてを熟達するのは至難の業である．そこで，筆者らは，発音の理解を進めらるようにするため，高齢者の話し言葉に「カナ文字」で,「字幕を付ける」ことで若い従事者とのコミュニケーションの円滑化を図る情報システムを検討している．

音声認識技術は機械学習により，音声の音響的な特徴（音響モデルと呼ぶ）と言語的な特徴（言語モデルと呼ぶ）を抽出することで，音声を計算機上で解釈をする．上記で示した「カナ文字」への変換は，これらのうち音響モデルが研究対象領域である．方言の音声認識において，音響モデルの学習に用いるラベル付き音声データの不足が課題である．

筆者らは方言音声の機械学習処理に不可欠な良質の方言

†1 (株)日立ソリューションズ東日本
Hitachi Solutions East Ltd.
†2 岩手県立大学 ソフトウェア情報学部
Faculty of Software and Information Science, Iwate Prefectural University.

音声データベースの効率的な構築を目指し，単語単位で発話された大量の音声データに対する音声表記文字列のラベル付け作業の改良を進めている．良質の方言音声データベースの条件として，以下を重視し今後も継続的に拡充する予定である．

(1) 方言特有の発音を含むすべての発音を含むこと．
(2) 話者が多く，性別による声の音色に偏りがないこと．

2. 本研究の背景

本研究においては，岩手県沿岸部を実証フィールドと位置づけ，岩手県宮古市中心部の方言を音声収集対象とした．方言音声データベースには，方言で用いられるすべての発音を含める要件を方言専門家に説明したところ，岩手県宮古市地域の方言には，共通語の発音には見られない音があることを，方言専門家からの助言を受けた．

① 共通語に比べ母音の種類が多く，アとエの中間の母音が流通しており，共通語の音声認識システムに最適化した音響認識モデルでは，この音を含むことで正確に音素が識別されない一因と考えられる．

② 語中の位置により，母音・子音の無音化，ガ行音の鼻濁音化，カ行音の濁音化が起きうる．

①	/a/	/i/	/u/	/e/	/ɛ /	/o/
/_'_/	あ	い	う	え	えぁ	お
	・・・・					
/n/	な	に	ぬ	ね	ねぁ	の
/nj/	にゃ		にゅ			にょ
	・・・・					
/_'j/	や	ゆぃ	ゆ		いぇ	よ
/_'w/	わ				わぇぁ	
/ɴ/	ん					
②	北(きΔ た)：母音が無声化　　蔵(くΔ ら)：子音が無声化 下駄(げだ)：一般的なガ行音 卵(たまご)：語中のガ行音、鼻濁音化 秋祭り(あぎまつり)：カ行が語中にあたるため濁音化					

図 1　宮古市中心部の音韻表(網掛け部が方言特有)

これを受け，宮古市中心部の方言を音響認識するには，共通語にはない発音が存在することから，その発音から派生する音声波形の形状に新たにラベル付けを行い，識別する必要があると考えた．方言専門家に依頼し，収集の対象となる発音 162 パターン（156 異音）を選定し，その発音を含む音声を集めるのに適した単語について整理頂いた．方言専門家のご支援のもと，この単語表に従って宮古の現地で音声データの収録を実施した．

3. 音声へのラベル付け作業の現状と問題点

3.1 従前のラベル付け手法

音声のラベル付け作業とは，実際の音声データに対して，その振幅波形とスペクトログラムの視察結果に基づいて，発話内容に則した文字ラベルの付与と各音素間の境界位置の特定を行う以下の作業のことである．

① 録音された音声データを計算機上に取り込んだ後，単語単位で音声区間を切り出し，個別のファイルに保存．

② 音声データ分析ツール（今回は praat[3]を使用）で音声の波形とスペクトログラムを参照しながら，各音素の境界位置に印を付ける．

③ 個々の音声データに，その発話内容を表す音素ラベル文字列情報を付ける．

従来のラベル付け作業は，音声分析ソフトにて音声波形を図示し，図示されたスペクトログラムの視察による音素継続区間の推定等の分類が作業の大部分を占める．スペクトログラムは，横軸に音の時間的な変化，縦軸に周波数，音の周波数成分の分布を表す．発音が異なる場合，音声の音色の特徴を図示したスペクトログラムの形状も異なる．

スペクトログラムの画像パターン識別に関するノウハウは暗黙知として作業熟練者から未経験者に技術移転することが難しく，作業効率を改善することができないという課題が従来からあった．

共通語では過去にラベル付けされ蓄積された音声資源が存在し，その音声資源の特徴を機械学習した音響モデルを利用することで，新たな音声データにラベル付けの支援をコンピュータシステムで支援することが可能となっている．しかし，方言音声についてはラベル付けされた音声資源は存在しないため，従来からの人手による音声データへのラベル付け作業が必要である．

3.2 方言に対するラベル付けの留意点

宮古市地域の音声に付与する音素ラベル文字列情報は，宮古地域の方言の音の区切りである音節を表記しやすいカナ文字を使用する．カナ文字の終端の発音は，撥音（ん）を除き母音となることから，個々のカナ文字で示されるラベル付け境界は，母音の終端に着目することで識別を行う．

第 1，第 2 ホルマント周波数を音声スペクトログラムから人間が視察することで母音を区別する従来手法 [4] で，音声の波形とスペクトログラム下部（2000Hz 以下）の黒い紋様の継続性に着目し，母音の終端の視察による区切るという視覚的に認識可能な基準を設定した．

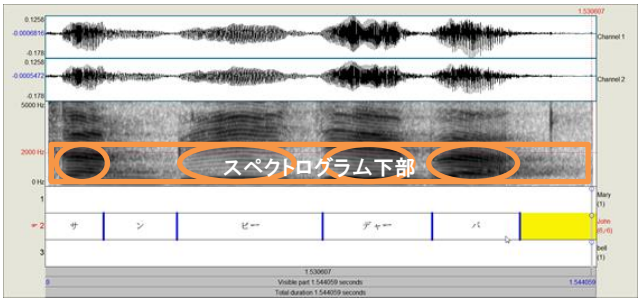

図 2　スペクトログラムの視察基準

3.3 従前のラベル付け手法の作業効率

宮古市地域における 3 回にわたる現地での聞き取り調査の音声データに対して，従前のラベル付け手法で実施した

作業時間の実績を以下に示す.

表 1　現行のラベル付け手法での作業実績

	1 回目	2 回目	3 回目
対象データ数	280 語	155 語	265 語
① 対象データに分割	5 時間（1 分 4 秒/語）	2 時間（46 秒/語）	3 時間（40 秒/語）
② 異音区間の分割 ③ カナ文字ラベル付け	19 時間（4 分 4 秒/語）	10 時間（3 分 52 秒/語）	20 時間（4 分 31 秒/語）
合計	24 時間（5 分 8 秒/語）	12 時間（4 分 38 秒/語）	23 時間（5 分 12 秒/語）

作業実績時間を振り返ると，(2) 異音区間の分割に時間を要している．作業者との事後の対話から，音素の区切り位置の特定作業に時間を要していると推察した.

① スペクトログラムの特徴を視察し異音区間を推定するも，推定した区切り位置ついて音声を再生し検証すると，実際に区切りを入れる場所とずれていた．この現象は，作業者によって音素の境界位置の推定に，ばらつきが発生することを示している.

② 音素区切り位置の検証と修正のため，何度か聞きなおしが発生し, 作業時間を長期化させいる. この現象は，方言音声に関する知識や経験の無い作業者が，音素の境界位置を推定する際に，スペクトログラムの特徴を見落としてしまうことが一因として挙げられる.

4. 提案手法

筆者は，ラベル付け作業の音素開始位置の推定時間の長期化に対する改善策として，方言の音素識別に関する情報が未知なケースで，方言の発音開始時のスペクトル変化量のピークが，個々の音素の開始時間を特定する特徴量として利用可能と仮定し，音素の区切り位置となる時間をラベル付け作業者に対して情報提示することを検討した.

Klapuri[5] らが提案したスペクトル変化量のアクセント（phenomenalaccent）では，音楽から音符の音の開始位置とその強さに関する情報を抽出することができる．これを方言に適用することで方言の音素の開始位置の候補点に関する情報を提示することができるようになると考えられる．アクセント特徴量の分析による音素の境界候補の表示手法を取り入れた試作システムの処理概要図を図 3 に示す.

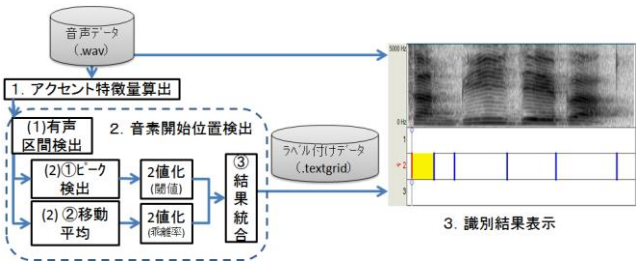

図 3 提案手法の処理概要

アクセント特徴量の信号から，音素の開始時間を抽出する手法は，以下の 3 つのプロセスからなる.

(1) 有声区間検出

算出されたアクセント特徴量は，パワースペクトルをもとに作成されている．パワースペクトルは，ホワイトノイズ部分についても加味されることから，有声区間を識別する必要がある．有声区間，無声区間の判定は，音声データに対し FFT を実施して求めた振幅スペクトルに対し離散コサイン変換を実施して，周波数成分の総和を算出し，その値の符号が正のケースを，有声区間と判断する.

(2) 特徴量のピーク検出

① アクセント特徴量が最小 0，最大が 1 となるように正規化した後，ピーク ピッキング法を適用する．時間軸上で，前のピーク値との差が 0.5 以上をピーク時間として採用する．ここで求めたピーク時間におけるピーク値に対し，細かなピークを排除する目的で，有声区間全体のアクセント特徴量の平均値の 3 倍以上の点を採用とする.

② 音素が変化することにより，アクセント特徴量が顕著に変化する時間は 50ms と仮定する. この時間内で，アクセント特徴量の変化量を評価するため，50ms の幅で，10ms 毎にアクセント特徴量の移動平均を算出する．50ms 移動平均値との乖離が，100 ポイント以上ある点をピークとみなす.

③ 音素が変化することにより，アクセント特徴量が変化する時間は 50ms と仮定しているので，①②のピーク点が 50ms 内に複数存在する場合，そのうちアクセント特徴量が最大の値の点のみを残す.

(3) 閾値の決定

前節で抽出したアクセント特徴量のピーク値に対し，実験的に引数パラメタで変化させながら,「抽出された音素境界点数≧正解の単語の文字数＋1」を満たす点群を導出する. 今回収録した音声データに対して, 上記のアクセント抽出法を適用した例を図4に示す.

算出されたアクセント特徴量（青色の線）に対し，特徴量の平均値の3倍（紫色の線）を閾値とする．アクセント特徴量の50ms移動平均線（赤色の線）との乖離率の推移線（緑色の線）に対し，100ポイント（橙色の線）の閾値とする. この図から，この特徴量が発声開始時や, 音素が変わる瞬間に極値（ピーク）をとり，この点を音素の境界の候補点として提示する.

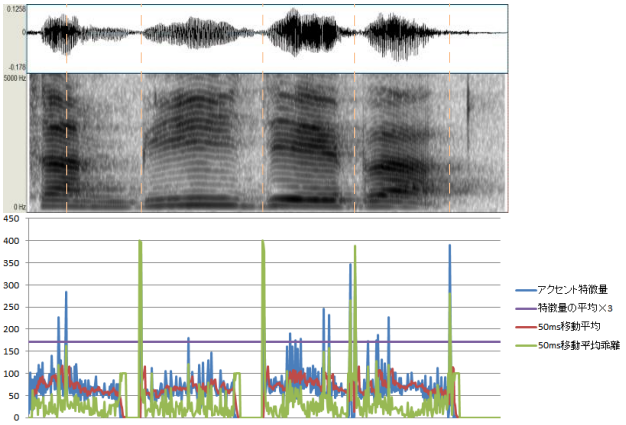

図 4 スペクトログラムとアクセント特徴量の比較
（例：サンビーデャーバ）
上図破線は，ピーク位置に合わせて引いた補助線

5. 評価

5.1 理論上の評価

上述の特徴量が音素の区切り位置を与えると仮定した場合，ラベル付け作業は，以下の２つの作業ステップとなると考えられる.

① アクセント特徴量のピーク位置を，音声分析ソフトウェアにデータを取込み音素の区切り線の表示を行う.

② 区切り位置の検証のため，音声の再生を行う.

よって，これまで人手によって行われてきた視察による音素区切り位置の推定工程が短縮され，作業時間の短縮見込まれる.

本稿で発表したラベル付け支援情報を音声分析ソフトとデータ連携することで，方言音声資源の収集手法を改善し，研究チーム内におけるラベル付け作業の作業時間の 20%以上縮減を見込む.

5.2 評価実験

実際にラベル付け作業未経験者の方に音素の開始位置に関する情報提示を行うこととラベル付け作業時間の効率改善の検証を行うため，情報提示あり，なしの影響を検証する実験を実施する.

表 2 テストパターン

実験種別	1 回目	2 回目
対照実験	情報提示なし	情報提示あり
	情報提示あり	情報提示なし

実験で用いるデータは，宮古市中心部の方言にのみ存在する発音を含む単語から 5 語，共通語にも存在する発音のみで構成される単語から5語のあわせて10語を対象として実験を行う.

表 3 評価対象の音声データ
(下線部の発音が，データベース化対象の発音)

データ種別	テスト対象の単語
方言	えぁーだっこ，めぁーにじ，いぐべす，さんびゃぐえん，くるすー
共通語	あさま，きもの，ぬま，せんとく，もいち

5.3 評価実験

ラベル付け作業の作業効率の改善状況を測定するため，作業時間の計測を行う．また，作業品質の確認のため音素の区切り位置の評価を行う.音素の区切り位置の評価には，熟練作業者と作業未経験者の区切り位置の比較評価を実施する.

表 4 評価基準

評価項目	比較対象	評価方法	評価基準
作業効率	・情報提示あり ・情報提示なし	全作業時間÷ラベル付け語数（訂正を含む）	20%縮減
作業品質	・熟練作業者 ・作業未経験者	時系列上の区切り位置の差[6]	50%以上が同一ラベルの区間に照合

評価実験結果については，現在実験を推進中のため後報とさせていただきます.

6. 今後の課題

今後の課題として，今回の手法の効果検証のため，単語表全体に渡り区切り位置の視察調査した際に確認された，以下の音素の区切り位置を特定できないケースを確認している.

① 母音から母音に渡る音の区切りが認識されない.

② 母音から撥音「ん」に渡る時，母音側に音素区切り情報の提示時間が前倒しになる.

これらについて，韻律情報を周波数帯域で細分化して視察することによって分離した事例[7]があることから，今回の手法で算出したアクセント特徴量を周波数帯域ごとに分析処理を行う改善を行う考えである.

謝辞 本研究は総務省 SCOPE(No.: 152302005)の支援により行われています．本稿を作成するにあたり宮古地域の方言音声収録にご協力いただいた田中宣廣教授（岩手県立大学宮古短期大学部）ならびに話者の方々に感謝の意を表します.

参考文献

[1] "第 2 部 ICT が拓く未来社会第 1 部第 1 節我が国経済の将来課題と ICT". 情報通信白書，平成 27 年版．総務省，(2015)，p. 251-256.http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/n5100000.pdf
[2] 竹田晃子,"東北方言オノマトペ用例集".国立国語研究所, (2012) https://www.ninjal.ac.jp/pages/onomatopoeia/
[3] Boersma, Paul and David Weenink 2015 Praat: Doing phonetics by computer (version 5.4.15) [computer program].http://www.fon.hum.uva.nl/praat/
[4] 板橋秀一,"音声工学",森北出版
[5] Klapuri, A., Eronen, A. and Astola, J.: Analysis of the meter of acoustic musical signals, Audio, Speech, and Language Processing, IEEE Transactions on, Vol.14,No.1, pp.342 – 355 (2006).
[6] 武田一哉,匂坂芳典,片桐滋，"視察に基づく音韻ラベルの性質". 日本音響学会講演論文集, (1987)
[7] 北澤茂良,北村達也,伊藤敏彦,"日本語 MULTEXT における韻律情報の分析と収録".平成 12–16 年度　文部科学省 科研費特定領域研究「韻律に着目した音声言語情報処理の高度化」