4R-06

遠野方言音声理解のためのキーワードスポッティング方式の検討

有賀智広† 皆川玲緒† 小嶋和徳† 李時旭‡ 伊藤慶明†
岩手県立大学† 産業技術総合研究所‡

1. はじめに

岩手県遠野市で語り継がれている遠野物語は，訛りや方言を使用しているため趣がある一方，県外者には理解が困難である．そこで我々は，物語の内容を誰もが理解できるように物語中の重要なキーワード(KW)が話された際に，KW を検出し，解説表示するシステムを提案している．一般に標準語音声用の自動音声認識システムを用いることができない方言音声には書き起こしデータはないため，学習や再学習が困難であり，一方，少資源音声に対する研究にあるように，方言音声に対してもキーワードスポッティング(KWS)は可能と考える．本稿では KWS の手法として，音声 KW を用いた KWS 法及びテキスト KW での KWS 法について検討し，KW をテキストで与えた場合に，フレームレベルのサブワード系列を構成し，詳細な照合を行う手法を提案し，提案法の有効性を検証する．

2. 先行方式

2.1 Posteriorgram 照合方式

音声データの 1 フレーム毎の特徴量を深層学習モデル等に入力することで triphone の状態や音素等の事後確率ベクトルが得られる．この事後確率ベクトルを時系列順に並べた行列を Posteriorgram と呼ぶ．音声クエリと音声データの Posteriorgram 同士で CDP（Continuous Dynamic Programming）等による照合を行い，検索を行う．照合に使用する局所距離は，音声クエリと 音声データ双方の事後確率ベクトルの内積を求め，その負の対数をとることで求められる．

2.2 状態間照合方式[1]

音声データをサブワード系列で音声認識し，結果を triphone の状態系列に変換し保持する．テキスト KW をサブワード系列に変換した後，状態系列に変換し，状態間音響距離[2]を用いて CDP 照合を行う．

Examination of Keyword Spotting Method for Understanding Tono Dialect Speech
†Tomohiro Ariga，Reo Minakawa，Kazunori Kojima，Yoshiaki Itoh ・ Iwate Prefectural University
‡Shi-wook Lee，・National Institute of Advanced Industrial Science and Technology

3. 提案方式

本節では我々の提案方式であるフレームレベル状態系列照合について説明する．まず，テキスト KW をサブワード等の系列番号に変換する．音声データを深層学習モデルに入力し，得られた Posteriorgram を KW の系列番号に一致する事後確率値のみ参照する．図 1 に示すように，KW の系列番号が 4,2,3 の場合 posteriorgram の 4 行，2 行，3 行のみを下図のように並べ替えて考え，KW が発話されている区間では，対角線上に高い確率値が現れる(下図の 3~5 フレーム)．DP を行う上では距離とするため事後確率値の負の対数をとることで距離化しておく．

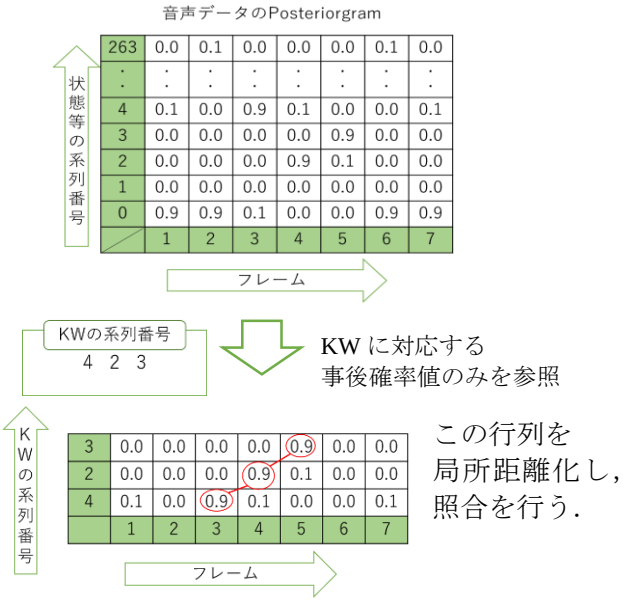

図 1．提案手法の概要図

我々は，DP を行う上では音声データ中の KW が発話されている区間の系列番号数と KW の系列番号数が同程度であることが望ましいと考え，KW の系列番号数が音声データのフレームレベルの系列番号数と同程度となるように，KW の系列番号を複数個並べることで，擬似的に引き延ばす．標準語に対してはテキスト KW の状態等の系列を複数個並べることで精度の向上を確認し，有効性が確認された[3]ため，本方式の遠野方言音声での有効性を検証する．

4. 評価実験

4.1 実験条件

本実験では BLSTM，ESPnet[4]，wav2vec2.0[5]の 3 つの深層学習モデルを用いた．BLSTM，ESPnet の学習条件，wav2vec2.0 の fine-tuning 条件を表 1 に示す．wav2vec2.0 の fine-tuning に用いる事前学習済みモデルには[5]に記述されている base モデルを用いた．

表 1．ESPnet と BLSTM の学習条件

モデル		BLSTM	ESPnet	wav2vec2.0
入力		FBANK		音声波形
出力		triphone 状態	syllable	kana
学習データ		CSJ2702 講演(約 600 時間)		
ノード数	入力層	1,320	83	1
	中間層	256	-	-
	出力層	3,009	264	83
学習係数		0.001	-	0.0001
エポック数		30	26	30

照合に用いる Posteriorgram について，ESPnet, wav2vec2.0 は CTC の出力を softmax 関数を用いて確率値に変換したものを事後確率ベクトルとして用いる．

4.2 テストセット

評価には遠野物語 7 話を用いた．それぞれの物語毎に KW を決定した．各物語および KW 等については表 2 に示す．評価指標には MAP(Mean Average Precision)を用いた．検出時間は最も時間のかかる Posteriorgram 照合でリアルタイムでの検出が行えるため，本実験では考慮しない．

表 2．物語の詳細

物語名		A	B	C	D	E	F	G
発話数		91	115	86	37	40	36	69
KW	種類	11	19	10	9	9	10	11
	正解発話数	41	48	36	29	26	42	57
	話者数	3 人						

4.3 実験結果

まず，音声 KW を用いた Posteriorgram 照合の結果を図 1 に示す．wav2vec2.0 では，1~30 エポックのそれぞれのモデルを用いて Posteriorgram 照合を行い，最も MAP が高い結果(2 エポック)を示す．この場合に，ESPnet よりも高い MAP となった．

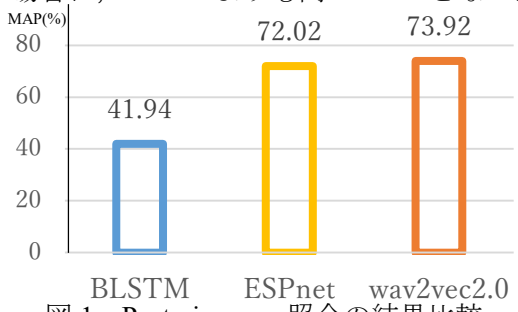

図 1．Posteriorgram 照合の結果比較

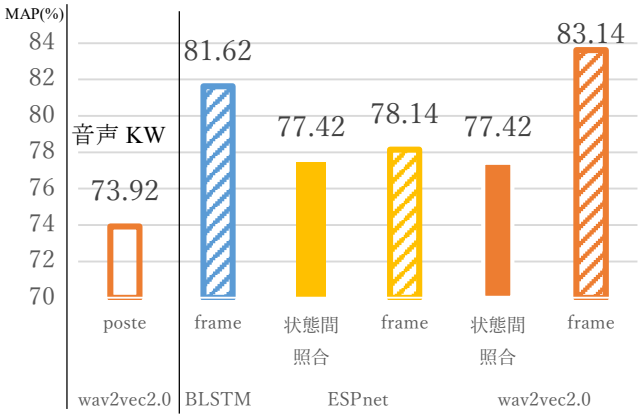

図 2．テキスト KW での結果比較

次に，テキスト KW を用いた結果を図 2 に示す．Posteriorgram 照合を poste，提案方式を frame とした．提案方式においては結果的に 1～4 エポックで高い MAP が得られた．図では Posteriorgram 照合と同じ 2 エポック目の結果を示す．エポック数の設定については今後の課題とする．

フレームレベル状態系列照合を行う際の KW の状態等の系列数は標準語の場合は 3 で最も良い精度が得られたが，遠野音声に関しては BLSTM では 17，ESPnet では 30，wav2vec2.0 では 65 のとき最も MAP が高くなった．この理由についてはより詳細な分析が必要と考える．

図 1，2 より，遠野方言では音声 KW よりもテキスト KW を用いることで，全てのモデルで約 5pt 以上高い MAP が得られた．状態間照合方式に比べフレームレベル状態系列照合方式は高い MAP が得られ，提案方式の有効性を確認できた．wav2vec2.0 を用いた場合，MAP が最も高くなり 83.14%となった

5. 結論

本稿では，テキスト KW におけるフレームレベル状態系列照合方式を用いて遠野方言音声に対して，検出精度の向上を図った．音声 KW よりも，テキスト KW を用いた提案方式により 80%以上の検出精度(MAP)が得られた．従来手法と比べ，検出精度が向上し，提案方式の遠野方言での有効性を確認できた．

参考文献

[1] 紺野良太他，”音声中の検索語検出におけるフレームレベル状態系列間照合方式”，信学技報，Vol. 115，no. 46，SP2015-37，pp 7-12 (2015)
[2] 岩田耕平他,“語彙フリー音声文書検索方式における新しいサブワードモデルとサブワード音響距離の有効性の検証”,情処論文誌, vol.48, no.5, pp.1990-2000 (2007)
[3] 皆川玲緒他，“音声中の検索語検出における検索精度向上のためのフレームレベル照合方式”，情処研究報告，Vol.2022-1R-07，No.2，pp.295-296，(2022)
[4] S. Watanabe, et. al , “ESPnet : End-to-end speech processing toolkit”, INTERSPEECH, 2018, pp. 2207–2211.
[5]A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” arXiv preprint arXiv:2006.11477, 2020.