4Q-03

遠野物語における複数言語の音声認識モデルを用いたキーワード検出精度向上

飯田英仁† 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‡

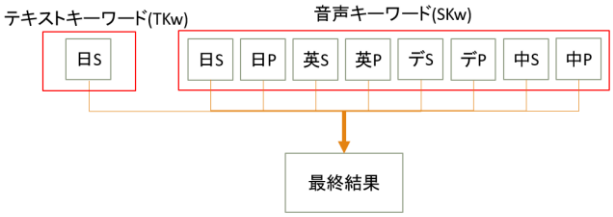
1. はじめに

日本の標準語音声に対しては最先端の自動音声認識システム(ASR：Automatic Speech Recognizer)を用いれば高い認識，検出精度が得られるが，方言や訛りのある音声(以降，方言音声と呼ぶ)の自動音声認識は困難である．方言音声を高い精度で自動音声認識するためには，方言音声の書き起こしが必要となるが，その方言音声の書き起こしは方言を話す人でも困難である．一方，方言音声に対しキーワード(KW)を検出することは既存の STD(Spoken Term Detection)モデルを用いて可能と考える．

遠野物語は岩手県遠野市で語られている昔話で，訛りや方言を用いて風情や面白さを出している．岩手県遠野市では毎週公演が行われている．県外から来た観光者等は遠野物語の公演を視聴した際，方言音声の難しさから，物語の内容を理解することが困難である．語り部(遠野物語の語り手)が，方言の解説を話の途中で行うとリズムや趣を失われるため行われていない．この問題を解決するために，我々は KW 検出技術を用いてオンラインで物語の用語解説を表示するシステムの開発を進行中である．

KW をテキストで与え KW 検出(テキスト KW 検出)する場合，KW が方言であると音声認識システムの未知語となるため，音節等のサブワードを認識単位とした音声認識システムを用いて，音声データ(遠野物語)をサブワード系列，さらに状態系列に変換する．KW も同様に triphone を経由して状態系列に変換する．音声データと KW を状態レベルで照合し，距離が閾値以下になると出力する．KW を音声で与える場合(音声 KW 検出)は音声認識システムを用いて音声 KW をテキストに変換し，テキスト KW 検出と同様に照合を行う方法と音声 KW，音声データのフレームレベルの Posteriorgram で照合する方式が代表的である．我々はこれまでテキスト KW 検出では音節認識を行うことで最も良い精度が得られ，音声 KW 検出よりもテキスト KW 検出の方が高い精度が得られることを確認した[1]．また，音声 KW 検出単体では高い検出精度が得られなかった．

我々は音声 KW 検出で遠野音声と音声 KW を日本語の音声認識システムの音響モデル(以降，音声認識モデル)を用いた．方言には標準語音声にはない音韻があり，外国語にも日本にはない音韻があるため，方言に外国語の音韻モデルを適用すれば有効に機能するのではないかと考える．本稿では複数の外国語の音声認識モデルで方言音声を認識し，それぞれの音声データの認識結果に対し音声 KW 検出を行う．これらの KW 検出結果とテキスト KW 検出結果と統合することにより，テキスト KW 検出と複数の外国語のモデルの音声 KW 検出が補完し合い，検出精度が向上すると考える．



S:状態間照合
P:Posteriorgram照合
日:日本語 英:英語 デ:デンマーク語
中:中国語

図１ 統合のイメージ

2. 提案手法

複数言語の音声認識モデルを用いて音声 KW 検出を行い，それらのスコアとテキスト KW 検出のスコアを統合することにより精度向上を図る．提案方式のイメージを図１に示す．音声 KW は状態間照合[2]と Posteriorgram 照合[3]の２種の照合方法を用いる．図中に略記法を示す．例えば最も右の「中 S」は音声 KW で中国語の Posteriorgram 照合を表す．スコアの統合は式(1)，(2)の 2 種を用いて行った．式(1)ではテキストでのスコアに重みαを与え，残りの重み$(1-\alpha)$をN個の音声 KW 検出のスコアに均等に割り振る方法(均等統合)である．式(2)ではテキストでのスコアに重みαを与え，残りの重み$(1-\alpha)$を音声 KW 検出の検出精度の順位ごとに割り振る方法(ランク統合)である．ランク統合ではより精度が高いスコアに重みが大きくなるように統合を行う．

$$D_{new} = \alpha D_{TKw} + \sum_{i=1}^{N} \frac{1-\alpha}{N} D_i \quad (1)$$

$$D_{new} = \alpha D_{TKw} + \sum_{i=1}^{N} \frac{1-\alpha}{1+2+\cdots+N} (N-i+1) D_i \quad (2)$$

D_{TKw}はテキストの KW 検出のスコア，Nは音声 KW 検出結果の数で最大 8 種(4 言語×2 照合方式)である．D_iはi番目の音声 KW 検出のスコアであり，式(2)

Improving retrieval accuracy for keyword spotting in Tono tales using multi-language speech recognition models.
†Iida Eiji, ‡Lee Shi-wook, †Kojima Kazunori and †Itoh Yoshiaki
†Iwate Prefectural University, ‡AIST

のiは順位を表す．$0 \leq \alpha \leq 1$としαは0.1刻みで精度の評価を行った．今回は予め個々の検出精度を求め，順位を決めた．

3. 評価実験

3.1. 実験条件

キーワード検出するための音声認識システムのツールは Kaldi[4]を用いた．予備実験により Posteriorgram 照合では音声認識システムのツールである ESPnet[5] の方が Kaldi よりも検出精度が高いため ESPnet を用いた．学習データは CSJ 偶数講演の約 287 時間(日本語)，TED-LIUM コーパスリリース 2 の約 250 時間(英語)，Aishell コーパスの約 178 時間(中国語)，Sprakbanken のコーパスの約 305 時間(デンマーク語)を用い，単語認識を行った．

入力特徴量は Kaldi ではフィルタバンク 120 次元を用い，前後 5 フレームを追加し 1320 次元，ESPnet ではフィルタバンク 80 次元を用い，ピッチを追加した 83 次元とした．評価指標には MAP(Mean Average Precision)を用いた．

3.2. 評価データ

評価データを表 1 に示す．遠野物語 3 話，KW は物語毎に計 40 種用意し，音声 KW は 3 人の学生に 2 種類の発話してもらい(物語中の KW の視聴前と視聴後)，計 240(40×3×2)個を用意し評価を行った．

3.3. 実験結果

テキスト KW と各言語それぞれの 2 種の照合方法で検出結果を表 2 に示す．音声 KW では，日本語の検出精度が最も高く，次に英語の Posteriorgram 照合が高かった．各統合方式の結果を図 2 に示す．日本語のみ統合を行った場合(+日 S+日 P)はテキスト KW 検出からランク統合では 1.43pt(74.93%→76.37%)の向上，均等統合では 1.61pt(74.93%→76.54%)の向上した．この日本語の 3 種の結果に英語の Posteriorgram 照合を加えて統合すると，さらにランク統合で 0.98pt(76.37%→77.35%)の向上，均等統合で 0.51pt(76.54%→77.05%)の向上し，英語を統合することで精度向上が確認できた．また，英語を統合に加えた場合，ランク統合(77.35%)の方が，均等統合(77.05%)より高い精度が得られた．一方，中国語，デンマーク語に関しては効果を得られなかった．

表 1　各物語の KW 数と正解数

物語名		A	B	C
話者の性別		女性	男性	女性
物語時間		5:31	4:25	6:59
発話区間数		86	91	115
KW 数		10	11	19
正解数	平均	3.60	3.73	2.53
	最小	1	1	1
	最大	16	17	10

表 2　各言語・照合方法の検出精度

KW	言語	照合方式	MAP[%]	Rank
TKw	日本語	状態間	74.93	
SKw	日本語	状態間	38.15	2
		Posteriorgram	42.78	1
	英語	状態間	18.81	6
		Posteriorgram	21.72	3
	中国語	状態間	20.33	4
		Posteriorgram	16.58	7
	デンマーク語	状態間	19.47	5
		Posteriorgram	8.90	8

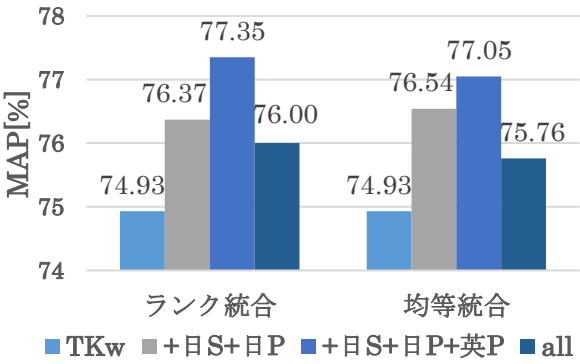


図 2　スコア統合後の検出精度
(S：状態間，P:Posteriorgram)

4. まとめ

本論文では，複数言語の音声認識モデルを用いて音声 KW 検出結果を求め，テキスト KW 検出結果とスコア統合し，精度改善を行った．日本語に外国語を用いて統合する場合，2 種の統合方式ともに効果があり，提案手法の有効性を確認できた．今後は統合方式を再検討するとともに，クロスバリデーション等の評価方法の見直し，及び他の音声認識モデルの検討を行う予定である．

謝辞: 本研究は音声データ収集に協力していただいた遠野文化研究センターの前川さおり様，語り部の会の皆さまに深く感謝いたします．本研究の一部は JSPS 科研費 18K11358 の助成を受けたものです．

参考文献

[1] 飯田英仁 他，”遠野物語方言音声の収録とその理解システムの検討”，音講論(2018).
[2] 高橋仁基 他，”音声中の検索語検出における事前検索・HMM 状態系列照合・リランキングの適用”，情処研報，SLP (2013).
[3] M.Obara, et.al, ”Rescoring by Combination of Posteriorgram Score and Subword-Matching Score for Use in Query-by-Example”, INTERSPEECH, 2016, pp.1918-1922
[4] D Povey,A. et.al, “The Kaldi Speech Recognition Toolkit”, ASRU 2011
[5] S. Watanabe, et.al, “ESPnet: End-to-end speech processing toolkit”, INTERSPEECH, 2018, pp. 2207–2211.