Vol.2020-NL-246 No.1
Vol.2020-SLP-134 No.1
2020/12/2

ゼロ資源言語の音声検索に用いる特徴量および学習方法の検討

水落 智[1,a)]　伊藤 彰則[1,b)]　能勢 隆[1,c)]

概要：本稿では，資源が豊富な言語を使用して，ゼロ資源言語の音声検索に用いるための特徴量，およびその抽出器となるモデルの学習方法について検討する．検討した特徴量は音素 Posteriorgram とボトルネック特徴量，モデルの学習方法は単一言語および複数言語を用いるものである．その結果，単一言語の音素 Posteriorgram を連結することで得られた特徴表現を用いた場合で，音声検索精度が最も高くなった．

Investigation of Features and Learning methods Used for Spoken Term Detection of Zero-Resource Language

1. はじめに

現在，多くの少数言語が絶滅の危機に瀕している．世界中には 4000〜6000 の言語があると推定されているが，それらの多くは 100 年程度で絶滅すると考えられている [1]．このような言語の消滅を止めることは困難であるが，絶滅危惧言語の話者がいる間に記録を残すために多くの試みがなされている [2–6]．その中でも，文字言語だけではなく，音声言語を保存してデータベースを作成することが重要であり，実際に多くの言語を保存する努力が行われている．世界中のすべての言語をアーカイブする Human Language Project [7] のような野心的な試みもあり，このように作成されたコーパスを利用しようとする試みもある [8].

このようなデータベースでは，話し言葉を単に記録してカタログ化するだけではなく，単語でデータベースを検索できることが望ましい．音声データに対して検索を行うことは「音声ドキュメント検索」と呼ばれ，特に特定の単語を検出するタスクは“Spoken Term Detection（STD）”と呼ばれる [9]．STD を実行するためには，いくつかの方法がある．1 つの方法は，音声認識によって音声データベースを書き起こし，テキスト検索を実行することである．この方法は，高速に検索が実行できるが，この方法を利用するには，目的言語の正確な音声識別器が必要である．絶滅危惧言語のように，言語資源が利用できない言語の音声認識システムの開発に関する研究もあるが [10]，一般に，このような言語に対して高精度の音声認識を実現することは困難である．他の方法として，検索される音声データベースを認識せずに検索を行う方法がある．この方法は，検索キーとして音声を利用し，信号処理的な方法でデータベースを検索する [11–13]．音声対音声の検索手法は言語に依存しないため，システム開発のための言語資源の量に依存せずに実現することができる．

これまでに我々は，対象言語とは異なる単一言語の音素識別器から得られた Posteriogram およびボトルネック特徴量を音声特徴量とした連続 DP マッチング [14] を用いる音声検索手法を検討した [15–18]．文献 [15] では，日本語 Posteriorgram を用いることで，MFCC よりも高い精度で単語検出できることを示したものの，言語の不一致が性能を低下させることが示唆された．文献 [16, 18] では，日本語 Posteriorgram と英語 Posteriorgram をフレーム単位で結合した特徴量を用いることで，言語の不一致による性能の低下を低減できることが示唆された．文献 [17] では，日本語および英語でそれぞれ学習したモデルから得られたボトルネック特徴量の使用を検討したが，単語検出精度は向上しなかった．

本研究では，ゼロ資源言語のデータベースにおける音声を用いた STD の精度の向上を目的とし，検討手法として，複数言語で学習したモデルから得られた音素 Posteriorgram

[1] 東北大学
Tohoku University, Sendai, Japan
[a)] satoru.mizuochi.p3@dc.tohoku.ac.jp
[b)] aito@spcom.ecei.tohoku.ac.jp
[c)] takashi.nose.b7@tohoku.ac.jp

Vol.2020-NL-246 No.1
Vol.2020-SLP-134 No.1
2020/12/2

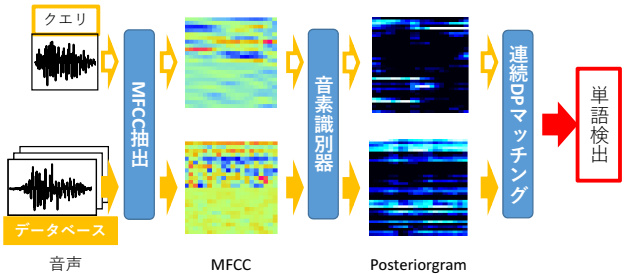

図 1　音素 Posteriorgram を用いた STD の概略
Fig. 1　Outline of STD using phonemic posteriorgram

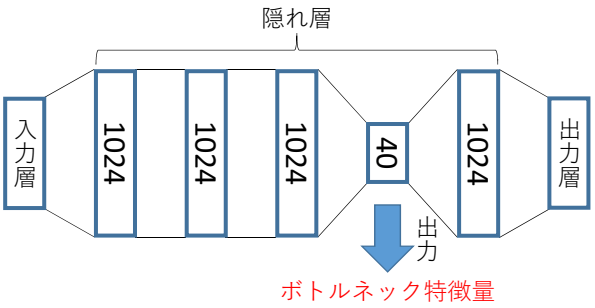

図 2　ボトルネック特徴量
Fig. 2　Bottleneck feature

およびボトルネック特徴量を音声特徴量とした連続 DP マッチング [14] を用いて STD を行う．また，目的言語はカクチケル語 [19,20] を用いる．

カクチケル語はグアテマラで話されているマヤ語族の言語で，約 450,000 人の話者がいると推定されている．実験に使用できる言語資源はごく僅かなので，基本的にこれらの言語資源は評価にのみ用い，システムの開発には他の言語（日本語，英語）のみを用いる．

2.　音声検索の概要および使用する特徴量

2.1　音素 Posteriorgram を用いた音声検索

音声をキーとした音声データベース検索 (Query by Example) は広く研究が行われている．現在は，話者性の影響を軽減するため，特徴量を Posteriorgram のような話者非依存な表現に変換してから距離を測る方法が提案されている [21]．しかしながら，ゼロ資源言語の音声検索では，対象言語の音素識別器を用意することはできない．そこで我々は，日本語音素識別器から得られた Posteriorgram を対象言語の音声検索に用いる手法を検討した [15]．文献 [15] で検討した STD の手法の概略を**図 1** に示す．データベース中の音声とクエリ音声の両方に対して，まず音響特徴量を抽出する．ここで，音響特徴量は MFCC12 次元とその Δ パラメータの計 24 次元である．次に，音響特徴量系列を音素識別器を用いて音素 Posteriorgram に変換する．その後，データベース中の音声とクエリ音声から得られた音素 Posteriorgram 間のユークリッド距離を算出し，連続 DP マッチングを行うことで単語の検出を行う．連続 DP マッチングによる STD では DP スコアの極小値の位置を検出候補点とし，設定した閾値より小さいスコアの検出候補点を検出結果とする．文献 [15] では，検討手法によって MFCC を特徴量とした場合よりも高い検索精度が得られることが確認できたが，言語の不一致が単語検出精度を低下させることが示唆された．

2.2　ボトルネック特徴量を用いた音声検索

文献 [22,23] ではボトルネック特徴量を用いた音声認識および音声検索が提案されている．ボトルネック特徴量は，DNN の隠れ層のうち他の隠れ層よりもノード数を小さくした（ボトルネックにした）隠れ層から得られる特徴量である（**図 2**）．文献 [22] は，特定の言語に偏らない特徴量を抽出するために，複数言語で学習したモデルから得られたボトルネック特徴量を用いて音声認識が行われた．提案されたモデルは出力層が言語ごとに分かれている．モデルの学習時にある言語が入力された場合には，その言語に対応する出力層のノードの誤差のみを用いてパラメータを更新する．結果として，クロスリンガルな音声認識において，複数言語で学習したモデルから得られたボトルネック特徴量を用いることで認識精度が向上することが示された．また，文献 [23] では，Stacked ボトルネック特徴量を用いて学習されたモデルから得られた特徴量をクロスリンガルな音声検索に用いる手法が提案された．この手法では，まず異なる話者が発話した同じ単語を文献 [22] と同様に学習したモデルを用いてボトルネック特徴量 に変換する．次に，それぞれ変換したボトルネック特徴量を入力と出力に用いて autoencoder を学習する．そして，学習された autoencoder の隠れ層から得られた特徴量を用いてスコアを算出し，単語検出を行う．クロスリンガルな単語検出において，上記の手法を用いることで MFCC および単一言語で学習したモデルから得られたボトルネック特徴量を autoencoder の学習に用いた場合に比べ，精度の向上が示された．文献 [22,23] より，複数言語で学習したモデルから得られたボトルネック特徴量を用いることで，クロスリンガルな音声認識および音声検索において高い精度を示すことが分かる．そこで我々は，単一言語のボトルネック特徴量であっても話者の情報と言語の種類に関する情報を取り除いて音韻の情報のみをスコアの算出に使用できることを期待し，単一言語で学習したモデルから得られたボトルネック特徴量を音声検索に用いる手法を検討した [17]．文献 [17] で検討した STD の手法の概略を**図 3** に示す．単語検出の流れは 2.1 節の音素 Posteriorgram を用いる場合とほとんど同様であり，音素識別器のボトルネック層から特徴量を得る点のみ異なる．以上の検討手法では，音素 Posteriorgram を用いた場合と比較して単語検出精度の向上を確認することができなかった．

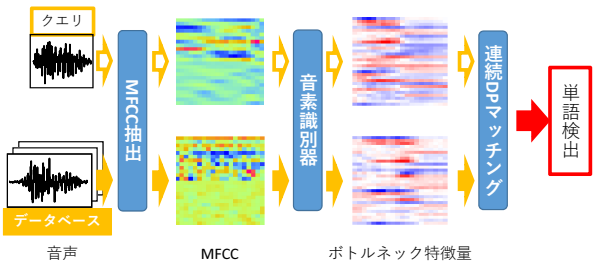

図 3　ボトルネック特徴量を用いた STD の概略
Fig. 3　Outline of STD using bottleneck feature

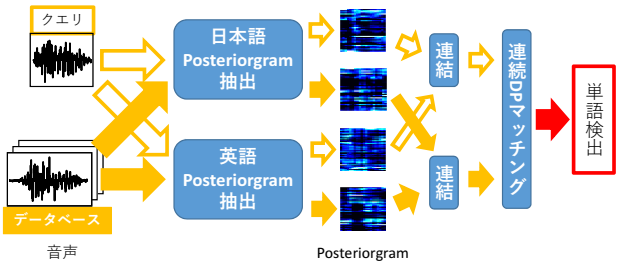

図 4　複数の音素 Posteriorgram を用いた STD の概略

2.3　複数言語の特徴量を用いた音声検索

2.1 節および 2.2 節の手法は，単一言語の特徴量を使用するものである．そこで，単語検出における言語の差異の影響を低減するため，複数言語の特徴量を使用する方法を検討する．

2.3.1　特徴量の連結

文献 [16, 18] では，言語の不一致による単語検出精度の低下を低減するために，複数言語の音素 Posteriorgram を用いる方法を検討した．文献 [16, 18] で検討した STD の手法の概略を**図 4** に示す．ここでは，英語と日本語の音素 Posteriorgram を距離の計算に用いる．まず，2.1 節と同様に，それぞれの言語の音素識別器を用いて，データベース中の音声とクエリ音声を音素 Posteriorgram に変換する．次に，各言語の音素 Posteriorgram をフレーム単位で連結する．その後，データベース中の音声とクエリ音声から得られた特徴表現間のユークリッド距離を算出し，連続 DP マッチングを行うことで単語の検出を行う．以上の検討手法によって，言語の不一致による単語検出性能の低下を低減できることが示唆された．

2.3.2　複数言語を用いた抽出器の学習

我々は，文献 [15–18] で単一言語で学習したモデルから得られる特徴量を用いた単語検出実験を行ったが，2.2 節で述べた通り，文献 [22, 23] では複数言語で学習したモデルから得られたボトルネック特徴量を用いた音声認識および音声検索が提案されている．そこで本稿では，複数言語で学習したモデルから得られた音素 Posteriorgram およびボトルネック特徴量を単語検出に用いる手法を検討する．複数言語でモデルを学習することによって，言語に依存しづらい特徴量が得られることを期待する．

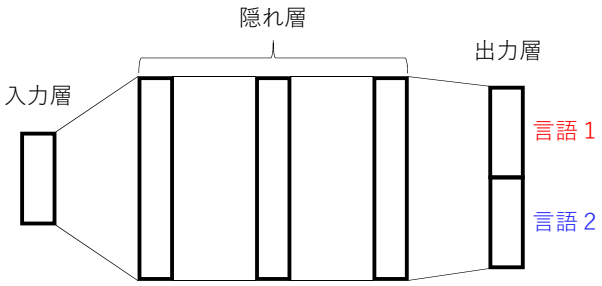

図 5　複数言語を用いたモデルの学習方法（ALL）
Fig. 5　Method of learning model using multiple languages (ALL)

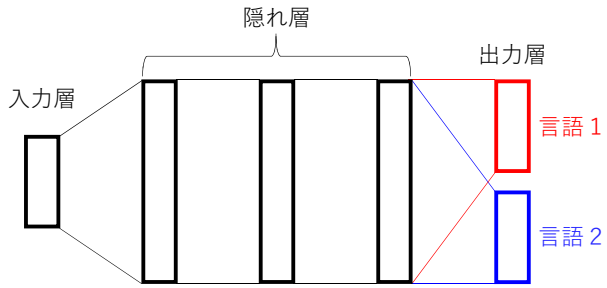

図 6　複数言語を用いたモデルの学習方法（DIV）
Fig. 6　Method of learning model using multiple languages (DIV)

複数言語を用いたモデルの学習にあたって，モデルのパラメータの更新の方法として以下の 2 通りの手法を検討する．

ALL　出力層の全てのノードの出力で Loss を計算してパラメータを更新（**図 5**）
DIV　入力された言語に対応するノードの出力のみで Loss を計算してパラメータを更新（**図 6**）

方法 ALL では，日本語音素の/a/と英語音素の/a/を明確に識別するように学習を行うことになる．一方で，方法 DIV では，日本語音素の/a/と英語音素の/a/が明確に識別されないように学習を行う．

3.　単語検出実験

本稿では，従来法として MFCC，日本語と英語それぞれの抽出器から得られた音素 Posteriorgram およびボトルネック特徴量，日本語と英語それぞれの抽出器から得られた特徴量を連結した特徴表現，検討手法として複数言語で学習したモデルから得られた特徴量をスコアの算出に使用した場合の単語検出性能を比較する．

3.1 特徴抽出器の学習

3.1.1 学習データ

日本語音素識別器の学習には JNAS を用いた．JNAS の音素バランス文のうち，男女各 127 名が発話した 2794 文を学習データ，男女各 13 名が発話した 286 文を開発データ，男女各 13 名が発話した 286 文を評価データとした．

英語音素識別器の学習には TIMIT を用いた．TIMIT の音声データベースのうち，男性 344 名，女性 146 名が発話した 4900 文を学習データ，男性 27 名，女性 13 名が発話した 400 文を開発データ，男性 27 名，女性 13 名が発話した 400 文を評価データとした．

複数言語の音素識別器の学習には，日本語音素識別器および英語音素識別器の学習に用いたデータセットの両方を使用した．

3.1.2 音素 Posteriorgram の抽出器

本稿では，音素 Posteriorgram の抽出器として，日本語音素識別器，英語音素識別器および複数言語の音素識別器を用いる．ここで，各言語の音素識別器には全結合の Multilayer Perceptron（MLP）を利用した．出力は各言語における音素のクラスとし，当該フレームの MFCC に加えて，前後数フレームの MFCC を結合したものを入力とした．MFCC は 12 次元とその Δ パラメータの計 24 次元とし，標本化周波数 16 kHz, 窓長 25 msec, フレームシフト 10 msec の条件で抽出した．各言語で共通の条件を**表 1** に示す．隠れ層数，隠れ層のノード数は表 1 の条件の中から最適なハイパーパラメータを探索するツールである Optuna [24] を用いて最適化した．複数言語の音素識別器のモデルのパラメータの更新は，2.3.2 項に示した 2 通りの手法で行った．

各言語およびパラメータの更新方法おいて，**表 2〜5** に示す入力フレーム数の異なる 4 つの音素識別器をそれぞれ学習し，単語検出に用いた．

3.1.3 ボトルネック特徴量の抽出器

本稿では，ボトルネック特徴量の抽出器として，日本語音素識別器および英語音素識別器を用いる．ここで，各言語の音素識別器には全結合の MLP を利用し，モデルの構造は文献 [22] を参考とした．入出力は 3.1.2 項と同様のものを用いた．各言語で共通の条件を**表 6** に示す．ボトルネック層以外の隠れ層のノード数は表 6 の条件の中から最適なハイパーパラメータを探索するツールである Optuna [24] を用いて最適化した．複数言語における学習の条件は 3.1.2 項と同様とした．

各言語およびパラメータの更新方法おいて，**表 7〜10** に示す入力フレーム数の異なる 4 つの音素識別器をそれぞれ学習し，単語検出に用いた．

3.2 特徴量の連結条件

単語検出実験においては，2 言語の音素 Posteriorgram

表 1 言語ごとに共通な学習条件（音素 Posteriorgram）
Table 1 Common condition of learning in each language (phoneme posteriorgram)

隠れ層数	2〜7
隠れ層のノード数	512, 1024, 2048, 4096, 8192, 16384
活性化関数	ReLU
ドロップアウト確率	0.5

表 2 日本語の学習条件（音素 Posteriorgram）
Table 2 Condition of learning in Japanese (phoneme posteriorgram)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層数	4		3	5
隠れ層のノード数	1024			
出力	36 クラス			

表 3 英語の学習条件（音素 Posteriorgram）
Table 3 Condition of learning in English (phoneme posteriorgram)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層数	4	3		4
隠れ層のノード数	1024	2048	1024	
出力	46 クラス			

表 4 複数言語の学習条件 : ALL（音素 Posteriorgram）
Table 4 Condition of learning in multiple languages : ALL (phoneme posteriorgram)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層数	3			
隠れ層のノード数	1024			
出力	81 クラス			

表 5 複数言語の学習条件 : DIV（音素 Posteriorgram）
Table 5 Condition of learning in multiple languages : DIV (phoneme posteriorgram)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層数	3			
隠れ層のノード数	2048	1024		2048
出力	82 クラス			

表 6　言語ごとに共通な学習条件（ボトルネック特徴量）
Table 6　Common condition of learning in each language (bottleneck feature)

隠れ層数	3
隠れ層のノード数	512, 1024, 2048, 4096, 8192, 16384
ボトルネック層	隠れ層 2 層目
ボトルネック層のノード数	30
活性化関数	ReLU
ドロップアウト確率	0.5

表 7　日本語の学習条件（ボトルネック特徴量）
Table 7　Condition of learning in Japanese (bottleneck feature)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層のノード数	8192	4096		
出力	36 クラス			

表 8　英語の学習条件（ボトルネック特徴量）
Table 8　Condition of learning in English (bottleneck feature)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層のノード数	4096			2048
出力	46 クラス			

表 9　複数言語の学習条件：ALL（ボトルネック特徴量）
Table 9　Condition of learning in multiple languages : ALL (bottleneck feature)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層のノード数	4096	2048		4096
出力	81 クラス			

表 10　複数言語の学習条件：DIV（ボトルネック特徴量）
Table 10　Condition of learning in multiple languages : DIV (bottleneck feature)

入力フレーム数	1 (当該のみ)	9 (前後 4)	15 (前後 7)	17 (前後 8)
隠れ層のノード数	4096		2048	4096
出力	82 クラス			

およびボトルネック特徴量を連結したものをスコア算出の特徴量として用いる．ここで，それぞれの言語における単語検出精度は音素識別器の入力フレーム数によって異なる．そのため，3.1 節で検討した入力フレーム数の条件に関して，すべての組み合わせで検討を行う．異なる入力フレーム数の音素識別器の出力を連結する時は，中心フレームの時刻が同じになるように結合した．

3.3　単語検出実験の条件

距離の計算には，以下の特徴量を使用した．

MFCC　MFCC12 次元と 1 次動的特徴量 12 次元
PPG　音素 Posteriorgram
BNF　ボトルネック特徴量
_JP　日本語で学習したモデルから得られた特徴量
_EN　英語で学習したモデルから得られた特徴量
_CONC　日本語，英語でそれぞれ学習したモデルから得られた特徴量を連結した特徴表現
_MULTI_ALL　手法 ALL を用いて複数言語で学習したモデルを使用
_MULTI_DIV　手法 DIV を用いて複数言語で学習したモデルを使用

日本語の評価データには，JNAS の新聞記事読み上げ音声のうち，音素識別器の学習，評価に用いていない発話文からランダムに選択した 400 発話を利用した．ここから，評価データに存在しない話者による 6 発話をクエリ音声とした．英語の評価データには，TIMIT の音声データベースのうち，音素識別器の学習，評価に用いていない発話文からランダムに選択した 400 発話を利用した．クエリ音声として評価データに存在しない話者が発話した 6 発話を用いた．カクチケル語の評価データには，静寂な環境で収録したカクチケル語会話音声 16 対話 347 発話（約 14 分）を利用した．クエリ音声は “achike”, “matyöx”, “peraj”, “richin” の 4 発話であり，これらの話者は評価データには含まれていない．クエリ音声には，連続発話から切り出したものを用いた．実際のシステムではクエリ音声は前後に無音を含んだ孤立発話であることが多いため，本稿では切り出された音声区間の前後に量子化 bit 数 16 bit で振幅 1 の白色雑音を，音素識別器の入力時に結合するフレーム数分だけ挿入した．

ここで，検出単語の評価は発話単位で行う．したがって，クエリ音声がその音声を含む評価音声の中で一度でも検出されれば正解とした．評価指標には Mean Average Precision (MAP) を用いた．MAP は，各クエリに対して求めた平均適合率の全クエリ平均であり，数値が大きいほど精度が高いことを示す．AP, MAP はそれぞれ式 (1), (2) で表される．

$$AP(q) = \frac{1}{c_q}\sum_{i=1}^{N}\delta_i \times precision(q,i) \quad (1)$$

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q}AP(q) \quad (2)$$

ここで，q はあるクエリ，c_q は検出された発話の中でクエリ q を含む発話数，N は検出された発話数，$precision(q,i)$ はクエリ q において発話 i が検出されたときの適合率，Q はクエリの総数を表す．

3.4 実験結果

単語検出実験の結果を**図 7** および**図 8** に示す．これらのグラフに示した結果は，距離の計算に使用した各特徴量の中で最も単語検出精度が高くなった結果である．図 7 から，カクチケル語における単語検出では，手法 DIV を用いて複数言語で学習したモデルから得られた音素 Posteriorgram を用いることで，単一言語の音素 Posteriorgram を用いた場合と比べて単語検出精度の向上が確認された．しかしながら，日本語と英語の音素 Posteriorgram を連結した特徴表現を用いた場合に比べると，単語検出精度が僅かに低くなった．一方で，手法 ALL を用いて複数言語で学習したモデルから得られた音素 Posteriorgram を用いても，単一言語の音素 Posteriorgram を用いた場合と比べて単語検出精度は向上しなかった．これは，手法 ALL では音素数が多い単一言語の特徴抽出器を学習していることと同じ条件になったことが原因だと考えられる．また，図 8 を見ると，ボトルネック特徴量を用いる場合は複数の言語を用いても単語検出精度は向上しないことが確認された．さらに，図 7 と図 8 を比べると，どの条件においてもボトルネック特徴量よりも音素 Posteriorgram を用いた場合で単語検出精度が高くなった．全体の結果として，全ての対象言語において最も単語検出精度が高いのは，日本語と英語の音素 Posteriorgram を連結した特徴表現を用いた場合であった．

4. まとめ

本稿では，ゼロ資源言語の音声検索を目標として，目的言語とは異なる複数の言語から得られた音素 Posteriorgram およびボトルネック特徴量を用いる単語検出法を検討した．実験の結果，音素 Posteriorgram を用いる検討手法では，単一言語の音素 Posteriorgram を連結した特徴表現を用いた場合と近い精度を確認することはできたが，さらなる向上は認められなかった．また，ボトルネック特徴量を用いる検討手法では，単語検出精度の向上は確認できなかった．今後は，モデルの学習に使用する言語の追加や特徴系列の比較方法の変更を検討していく予定である．

謝辞 本研究は文科省概算要求（プロジェクト分）「ヨッタインフォマティクス研究センターの設立事業」および科研費 19H05589 の支援を受けた．

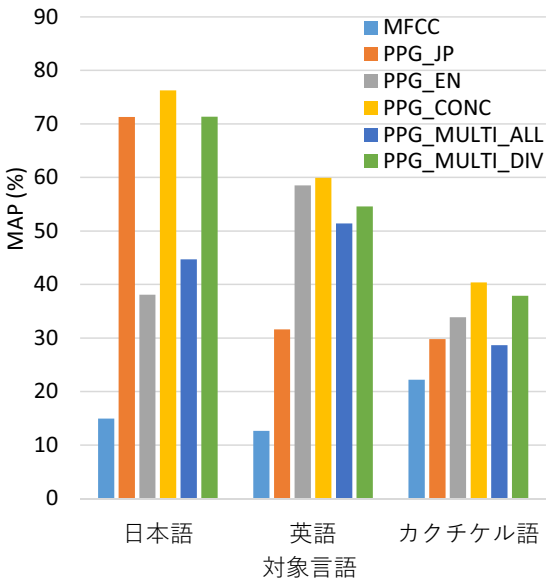

図 7 音素 Posteriorgram を用いた場合の結果
Fig. 7 Result using phoneme posteriorgrams

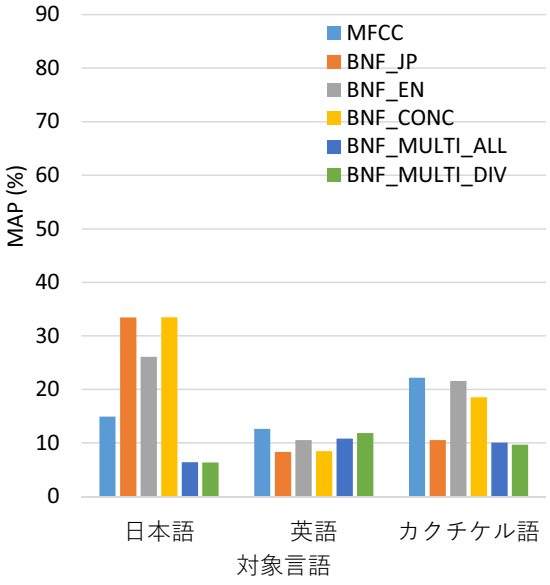

図 8 ボトルネック特徴量を用いた場合の結果
Fig. 8 Result using bottleneck features

参考文献

[1] M. Krauss, *The world's language in crisis,* Language, vol. 68, no. 1, pp. 4–10, 1992.
[2] M. Janse and S. Tol., eds., *Language Death and Language Maintainance: Theoretical, practical and descriptive approaches,* John Benjamins Publishing Co., 2003.
[3] A.C. Woodbury, *Language documentation and description,* vol. 1, chapter Defining documentary linguistics, pp. 140–153, London: SOAS, 2003.
[4] H. Johnson, *Language documentation and description,* vol. 1, chapter Language Documentation and Archiving or How to Build a Better Corpus, pp. 140–153, London: SOAS, 2003.

[5] S.N. Laoire, *Scottish gaelic speech and writing: Register variation in an endangered language,* Scottish Language, vol. 27, pp. 115–119, 2008.

[6] M.C. Yang and D.V. Rau, *An integrated framework for archiving, processing and developing learning materials for an endangered aboriginal language in taiwan,* Proc. the Fifth Workshop on Asian Language Resources (ALR-05) and First Symposium on Asian Language Resources Network (ALRN), pp. 32–39, 2005.

[7] S. Abney and S. Bird, *The human language project: Building a universal corpus of the world's languages,* Proc. 48th Annual Meeting of the Association for Computational Linguistics, pp. 88–97, 2010.

[8] T. McEnery, P. Baker, and L. Burnard, *Corpus resources and minority language engineering,* Proc. the Second International Conference on Language Resources and Evaluation (LREC'00), p., 2000.

[9] A. Mandel, K.P.P. Kumar, and P. Mitra, *Recent development in spoken term detection: a survey,* Int. J. of Speech Tech., vol. 17, no. 2, pp. 183–198, 2014.

[10] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, *Automatic speech recognition for under-resourced language: A survey,* Speech Communication, vol. 56, pp. 85–100, 2014.

[11] X. Anguera, L.J. Rodriguez-Fuentes, A. Buzo, F. Metze, I. Szöke, and M. Penagarikano, *Quesst2014: Evaluating query-by-sample speech search in a zero-resource setting with real-life queries,* In *Proc. ICASSP,* pp. 5833–5837, 2015.

[12] E. Barnard, J.Schalkwyk, C. vanHeerden, and P.J. Moreno, *Voice search for development,* In *Proc. INTERSPEECH,* pp. 282–285, 2010.

[13] F. Metze, X. Anguera, F. Barnard, M. Davel, and G. Gravier, *Language independent search in mediaeval's spoken web search task,* Computer Speech & Language, vol. 28, no. 5, pp. 1066–1082, 2014.

[14] S. Nakagawa, *Connected spoken word recognition algorithms by constant time delay dp, o(n) dp and augmented continuous dp matching,* Information Science, vol. 33, no. 1–2, pp. 63–85, 1984.

[15] 水落智，千葉祐弥，能勢隆，伊藤彰則. 日本語の Posteriorgram を用いたゼロ資源言語の音声検索の検討, 日本音響学会 2019 年秋季研究発表会講演論文集, pp. 841–842, 2019.

[16] 水落智，千葉祐弥，能勢隆，伊藤彰則. 複数言語を学習に用いたモデルによるゼロ資源言語の音声検索の検討, 日本音響学会 2020 年春季研究発表会講演論文集, pp. 993–994, 2020.

[17] 水落智，伊藤彰則，能勢隆. ボトルネック特徴量を用いたゼロ資源言語の音声検索, 日本音響学会 2020 年秋季研究発表会講演論文集, pp. 849–852, 2020.

[18] S. Mizuochi, Y. Chiba, T. Nose, and A. Ito, *Spoken Term Detection Based on Acoustic Models Trained in Multiple Languages for Zero-Resource Language,* In Proc. 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE 2020), pp. 447–448

[19] R.M. Brown, J.M. Maxwell, and W.E. Little, *La ütz awäch?: introduction to Kaqchikel Maya language,* University of Texas Press, 2010.

[20] M. Koizumi, Y. Yasugi, K. Tamaoka, S. Kiyama, J. Kim, J.E.A. Sian, and L.P.O.G. Matzer, *On the (non) universality of the preference for subject-object word order in sentence comprehension: a sentence-processing study in Kaqchikel Maya,* Language, vol. 90, no. 3, pp. 722–736, 2014.

[21] H. Wang, T. Lee, C. Leung, B. Ma, and H. Li, *Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection,* In *Proc. ICASSP,* pp. 8545–8549, 2013.

[22] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, *The language-independent bottleneck features,* In *Proc. SLT,* pp. 336–341, 2012.

[23] Y. Yuan, C. Leung, L. Xie, H.Chen, B. Ma, and H. Li, *Pairwise learning using multilingual bottleneck features for low-resource query-by-example spoken term detection,* In *Proc. ICASSP,* pp. 5645–5649, 2017.

[24] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Koyama, "Optuna: A next-generation hyperparameter optimization framework", In Proc. the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp.2623-–2631, July 2019.