様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 ３０ 日現在

機関番号：３３９１０
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０２７

研究課題名（和文）深層学習を用いた音声認識を最適化する音響モデル単位の自動獲得に関する研究

研究課題名（英文）Automatic acquisition of optimized acoustic model unit for automatic speech recognition using deep learning

研究代表者

山本　一公（Yamamoto, Kazumasa）

中部大学・工学部・教授

研究者番号：４０３２４２３０

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、日本語音声認識の性能向上のために、音響モデル単位を最新の深層学習技術を用いて自動獲得することを目指した。研究は「（1）深層学習を用いたクラスタリングによるモデル単位自動獲得」と「(2)多言語単音モデル群による単音-音素マッピングの曖昧さ解消」に分かれている。(1)では、DNN-HMM音響モデルにおいて、従来の文脈依存音素クラスタリングでは得られない状態クラスタリングにより、認識精度を向上させることができた。(2)では、多言語同時音声認識において、言語別の音素モデル単位よりも、話者や言語の違いを吸収するような音響モデリングを行うことで、音声認識精度が改善することが分かった。

研究成果の学術的意義や社会的意義
最近の深層学習技術の発展により、自動音声認識の性能は大きく向上し、音声AIアシスタントの入力インタフェースとして広く実用化されるに至った。しかしながら、英語音声認識と比べて日本語音声認識はやや性能が悪く、英語圏に比べて日本語の音声入力システムの活用頻度が低い理由のひとつとなっていると考えられる。本研究では、日本語音声認識システムの基本的な性能向上を目指すことが学術的な意義であり、ディジタルディバイドの影響を受けやすい高齢者に対しても高い音声認識精度を持つ音声入力システムを提供できるようになることが社会的な意義である。

研究成果の概要（英文）：In this research, we aimed to acquire acoustic model units automatically using the latest deep learning technology in order to improve the performance of Japanese speech recognition. This research is divided into two sub-themes: "(1) automatic acquisition of model units by clustering using deep learning," and "(2) disambiguation of phone-phoneme mapping by using groups of multilingual phone models". In the sub-theme (1), in DNN-HMM acoustic model, recognition accuracy could be improved by state clustering, which cannot be obtained by conventional context-dependent phonetic clustering. In the sub-theme (2), it was found that in multilingual (code-switching) speech recognition, the speech recognition accuracy is improved by performing acoustic modeling that absorbs differences in speakers and languages rather than the phonetic model unit for each language.

研究分野：音声言語情報処理

キーワード：音声認識　音響モデル　深層学習　モデル単位　音響クラスタリング　多言語

１．研究開始当初の背景

ここ数年、音声認識や音声合成を始めとする音声情報処理技術は、大量の学習データ（Google 等では数万時間に及ぶ）を用いた深層学習の導入によってその性能が大幅に向上し、これまでの人工知能研究の成果と相まって、Google アシスタントや Apple Siri、Amazon Echo 等の「音声 AI アシスタント」「スマートスピーカー」として、一気に実用化のステージへと登ってきた。しかしながら、YouTube の動画に付けられている音声認識結果を用いた字幕（「自動字幕」）を見ると、一見して英語字幕の精度に比べて日本語字幕の精度が低いと感じられ（例：図 1。YouTube の字幕は Google 音声認識エンジンによるものであり、世界最高水準の精度である）、音声認識はまだ実用上の精度が十分でない場面があり、精度の改善が継続して求められていることが分かる。

図 1: （左図）YouTube 英語自動字幕の例：かなり早口のニュース音声だが、認識誤りは多くは見られない。（右図）日本語自動字幕の例：「イノベーション」→「預言書」と誤認識。全体に誤認識が多いと感じる。

現在、世界中で用いられている一般的な音声認識のフレームワークは、DNN-HMM（Deep Neural Network - Hidden Markov Model; 深層学習により尤度計算を行う隠れマルコフモデル）による音響モデルと N-gram（単語連鎖確率モデル）による言語モデルを組み合わせて用い、最も尤度の高くなる単語列を探索して認識結果とする手法を採っている。これは英語音声認識で培われた研究成果のフレームワークをほぼそのまま利用したものであり、その言語特有の音声事象を扱ったり、英語と当該言語の音声言語的な違いを組み込んだりすることで音声認識性能を向上させる仕組みは、ほとんど入っていない。例えば、音声の特徴を表す音響モデルの単位として、英語では音声学的な単位である「単音」が使用されている。物理的な音がそのままモデル単位となっているため、パターン認識の単位として非常に扱い易いものになっていると考えられる。一方、日本語音声認識では、ラベル付けの容易な言語学的な単位である「音素」が音響モデルの単位として用いられている。音素は、言語として人間が知覚・弁別する単位であり、通常複数の単音がひとつの音素にマッピングされるため、パターン認識の単位として用いる場合、音響モデルが物理音響特徴的に曖昧になっていると言える。また、英語は声の強さでアクセントを表すため、日本語のように声の高さでアクセントを表す言語のための特徴量も不足している。このような曖昧な音響モデルや音響特徴量の不足が日本語音声認識の性能を英語音声認識より悪くしているのではないか？この曖昧性を解消することで、日本語音声認識の性能を向上することが可能ではないか？と考えたことが、本研究を行う動機となった。

２．研究の目的

本研究の目的は、適切な音響モデル単位を自動的に獲得するフレームワークを構築し、日本語音声認識精度を向上することである。これまでにも音響単位の自動獲得が試みられたことはあった（例えば、引用文献①）が、計算機の処理能力が低いために小さな次元数の特徴量しか扱えていなかったり、データベースの規模が十分でなかったりと、考察としては不十分であった。また、引用文献①の試みは従来のガウス混合モデル型 HMM（GMM-HMM）が主流であった時代の研究であり、現在の音声認識技術の重要部分を成す深層学習が取り入れられていない。現在、音声認識の性能改善に関する研究は様々行われているが、言語毎の最適な音響モデル単位を考える視点からの研究はほとんど行われていない。しかし、深層学習により実用的な音声認識技術の枠組みが出来上がった今だからこそ、この点を再考する意味があると考えた。

３．研究の方法

本研究の目的に対して、「(a)従来行われてきた自動音響モデル単位獲得の研究と深層学習の組合せによって、より高精度に適切な音響モデル単位を獲得すること」と、「(b)多言語単音モデルを日本語音声認識に適用できるように単音から音素への適切なマッピングを獲得すること」の 2 つのサブテーマを設定した。

(1) サブテーマ(a)では、実際の音声データから共通の音声単位を獲得することを目指すが、このために深層学習技術を用いる。DNN の隠れ層の一部のノード数を少なくし、そこから取り出した特徴はボトルネック特徴量と呼ばれ、入力層を音響特徴量、出力層を音声認識に使われる音素ラベルに対応させてネットワークを学習すれば、ボトルネック特徴は音声認識に適した音声圧縮表現となり得る。この圧縮された音声表現を用いて、圧縮表現同士を何らかの距離尺度でクラスタリングし、言語情報と統合することで、音声単位を自動的に獲得する枠組みを構築する。深層学習の枠組みと大規模なデータベースを活用すれば、大規模な次元数を持つ特徴量を直接的に扱うことができ、引用文献①で行われたものとは異なる視点で音響単位の自動獲得に関する研究を行うことができる。また、ボトルネック特徴はその他の音響クラスタリング（例えば、音響イベント認識）への応用も可能であるため、関連してこのような研究も行った。

(2) サブテーマ(b)では、複数言語の単音モデルを混合して DNN-HMM を学習することで、言語切り替えをすることなく多言語音声を認識する研究を行う。これまで、複数言語が認識可能な多言語音声認識システムでは、多言語の音素セットをまとめた大きな音素セットを用いて大きな音響モデルを学習していたが、例えば引用文献②では共通に用いることが可能な音素を言語間でシェアするだけで、発音変形を含む物理音モデルが自動的に共通化されて音声認識性能向上に貢献するわけではなかった。サブテーマ(b)の研究はこの点を狙っているものであり、多言語単音モデルから日本語音声認識に適した音素単位への自動マッピングを獲得する研究を行った。

(3) サブテーマ(a)およびサブテーマ(b)を行う中で、より新しい音声認識技術である End-to-End(E2E)音声認識フレームワークを用いることで、目指していたことが自然と実現できる可能性に気が付いたため、最終年度において研究方針の修正を行い、E2E 音声認識での音声認識精度向上を目指した特徴抽出フレームワークによる高速話者適応や E2E 多言語音声認識等を行った。

４．研究成果

(1) サブテーマ(a)では、DNN-HMM を用いた大語彙連続音声認識システムの DNN 部の途中出力を特徴量として用い、これにディリクレ過程混合モデルを用いて triphone の状態毎にクラスタリングすることにより音響モデル単位を自動的に獲得するベースラインフレームワークの構築を行った。実験では、図 2 に示すように、従来の文脈依存音素クラスタリング(当該音素の前後音素環境によるクラスタリング)では分割できない triphone 音響モデルの状態を、音響特徴量により更に分割することが可能である（この更に分割した状態を「疑似状態」と呼ぶ)。このより詳細化した状態分割の結果を用いて、図 3 に示す手順で新たに DNN-HMM 音響モデルを学習することにより、音声認識精度の向上を図る。

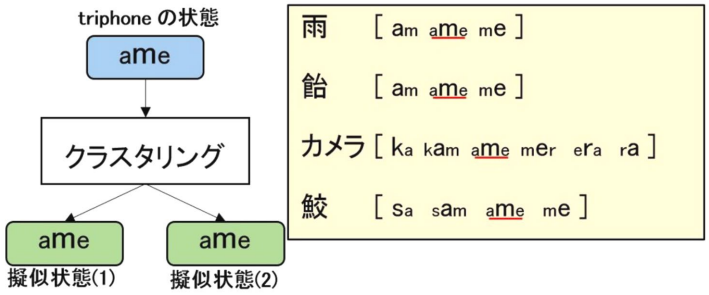

図 2 音素クラスタリング例（サブテーマ(a)）

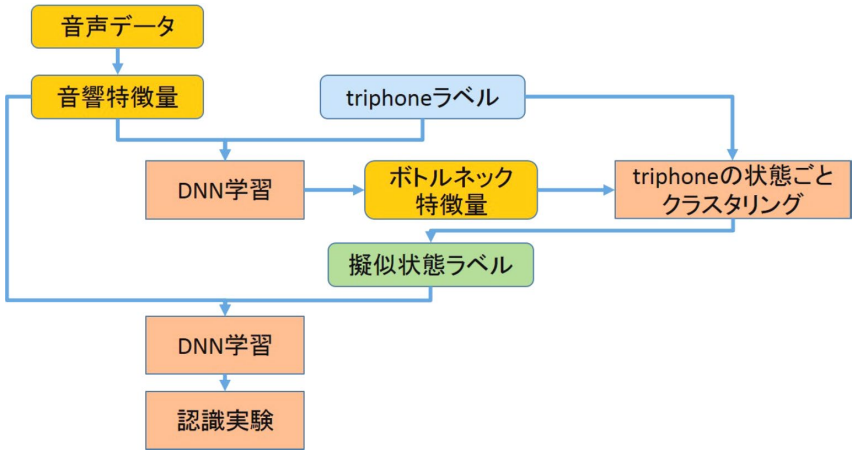

図 3 サブテーマ(a)の提案手法における音響モデル学習の流れ

日本語話し言葉コーパス（CSJ）を用いた実験結果を表 1 に示す。学習データは学会講演音声（男女混合の 967 講演、257 時間）、評価データは学習データに含まれない男性 10 講演である。「状態数」は triphone 音響モデルの共有状態数を示しており、824 が分割前の状態数、1126 が提案手法により状態数を増やした状態数であるまた、「採用法」の“max”は提案手法により再分割された状態の中での最大の確率値を音声認識時に用いる手法、“sum”は再分割された状態の確率を全て加算した値を音声認識時に用いる手法を示す。実験結果より、音素コンテキストでは状態分割できない状態を更に分割して状態数を増やすことで、音声認識精度を改善できていることが分かる。紙面の都合上割愛するが、状態分割結果の詳細を調査したところ、従来法では状態分割が起こらない同一の前後音素環境の音素においても状態分割が行われていることが分かり、音響特徴の変動が音素環境という言語情報環境だけでなく、韻律等の情報によっても起きていることが分かった。また、再分割した状態の確率値は、状態毎に用いるよりも、加算して用いる方が音声認識精度の改善に寄与することから、音素に対する単音のバラツキを音響モデルで捕らえることが重要であることが分かった。

表 1 サブテーマ(a)の提案手法における音声認識精度の改善

状態数	採用法	単語認識率 (%)	挿入誤り	削除誤り	置換誤り
824		75.56	221	1362	2725
1126	max	75.42	226	1336	2751
1126	sum	76.42	204	1251	2656

(2) サブテーマ(b)では、CTC(Connectionist Temporal Classification)（引用文献③）をベースとした E2E 音声認識の枠組みを用いて、複数言語（本研究では 6 ヶ国語（チェコ語、英語、フランス語、ドイツ語、日本語、スペイン語)）の音声を同時に用いてニューラルネットワークで複数言語の音声認識器を学習するのと同時に、言語識別器や話者識別器、スタイル識別器を学習させるマルチタスク学習を行い、各言語の音声認識精度改善を行った。また、各サブタスクに Gradient Reversal Layer (GRL)を導入することで、言語や話者に依存しない特徴変換の実現を目指した。図 4 に本サブテーマで使用したニューラルネットワークの構成を示す。左半分が CTC ネットワーク、右半分が補助タスクネットワークであり、左半分のネットワークを単体で事前学習を行った後で右半分のネットワークを結合して追加学習を行っている。

表 2 に使用した各言語のデータ量を、表 3 に実験結果を示す。提案手法では、すべてのマルチタスク学習モデルにおいて、CTC 事前学習モデルの音素誤り率を低下させることができた。これにより、CTC 単体で学習するよりも、補助タスクを加えた方が認識性能をより改善できることを示した。また、GRL を適用することで、適用しない場合よりも誤り率を抑えることができており、GRL で言語不変・話者不変・スタイル不変の特徴抽出ができていると考えられる。また、提案手法であるマルチタスク学習モデルがベースラインである DNN-HMM 音声認識システムの認識性能に匹敵し、複数言語の音声認識において、マルチタスク学習モデルが有効であることを確認した。

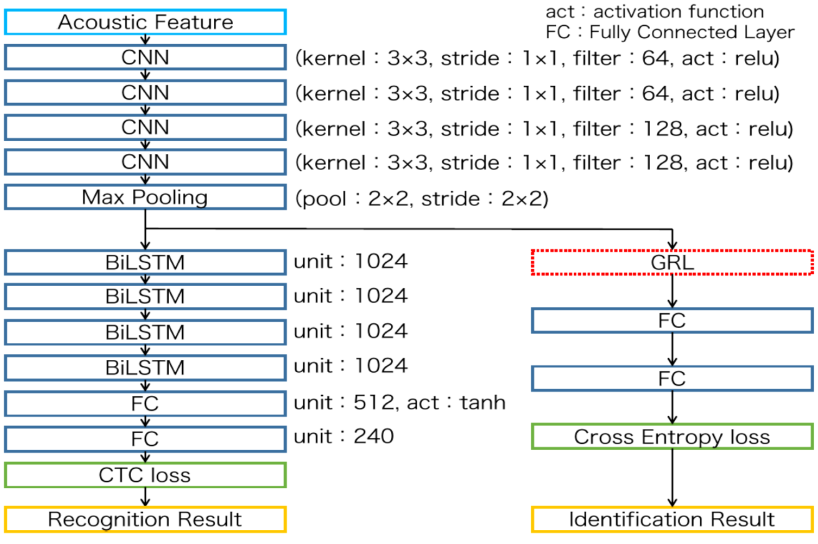

図 4 サブテーマ(b)で使用したニューラルネットワークの構成

表 2 サブテーマ(b)で使用した多言語データ

言語	学習時間長	評価時間長	音素種類数
チェコ語	26 [hour]	2.7 [hour]	41
英語	11 [hour]	2.6 [hour]	39
フランス語	23 [hour]	2.0 [hour]	38
ドイツ語	17 [hour]	1.5 [hour]	41
日本語	19 [hour]	5.1 [hour]	39
スペイン語	18 [hour]	1.7 [hour]	40

表 3 サブテーマ(b)の提案手法における音素誤り率

言語	DNN-HMM (Baseline)	事前学習 (CTC)	CTC and 言語識別器		CTC and 話者識別器		CTC and スタイル識別器	
			GRL×	GRL○	GRL×	GRL○	GRL×	GRL○
チェコ語	11.6	10.0	8.7	8.5	8.7	8.8	8.7	8.7
英語	26.5	34.9	33.7	32.6	31.2	36.7	35.1	34.7
フランス語	13.1	9.2	8.2	8.2	7.9	8.2	8.3	8.1
ドイツ語	19.0	17.7	15.3	15.2	14.9	15.6	15.5	15.3
日本語	14.7	20.3	16.3	16.8	19.7	16.0	17.7	16.1
スペイン語	10.8	12.1	9.7	9.5	9.2	9.6	9.5	9.3
全言語	15.7	18.1	15.7	15.6	16.3	16.1	16.4	15.7

(3) 最終年度では、E2E 音声認識での音声認識精度向上を目指し、各サブテーマに対応する形で特徴抽出フレームワークによる高速話者適応や E2E 多言語音声認識等を行った。

サブテーマ(1)への対応として、従来から行っていた頑健な音声認識のためのガンマトーンフィルタバンクによる自動的な特徴抽出手法を、音声認識が難しい超高齢者音声認識の少量音声による高速な話者適応手法として利用し、超高齢者音声認識の精度を改善する手法について研究を行った。この音声特徴抽出は、E2E 音声認識においても音声特徴抽出部として利用が可能であり、且つ、話者適応手法としても同様に動作するため、今後行う予定の研究の基礎として、有効利用可能である。

サブテーマ(2)への対応として、E2E の音声認識のフレームワークにおいて、言語および話者非依存の音声特徴抽出を行うことで音声認識の性能を向上させる手法を考案した。具体的には、音声認識器に加えて、言語認識器と話者認識器を組み合わせたマルチタスク学習法である。提案手法を多言語音声認識タスクで評価した結果、提案手法は各言語の音声認識モデルよりも高い精度を達成することができた。

＜引用文献＞

① 中川, 斎藤, “エルゴディック HMM に基づく音声の自動獲得単位を用いた音声認識”, 電子情報通信学会技術報告, 音声(SP), 97(64), pp.55-62, May 1997.

② T. Schultz, A. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition”, Speech Communication, Vol.35, No.1-2, pp.31-51, Aug. 2001.

③ A. Graves, S. Fernandez, F. J. Gomez, J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”, Proc. of ICML, 2006.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Kazumasa Yamamoto, Akinori Ishiki, Seiichi Nakagawa	４．巻 -
２．論文標題 Improvement of Elderly Speech Recognition Using Gammatone Filterbank Adaptation	５．発行年 2021年
３．雑誌名 Proceedings of 2020 IEEE 10th Global Conference on Consumer Electronics (GCCE)	６．最初と最後の頁 327-328
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/GCCE53005.2021.9622086	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Wang Yu, Chee Siang Leow, Akio Kobayashi, Takehito Utsuro, Hiromitsu Nishizaki	４．巻 -
２．論文標題 ExKaldi-RT: A Real-Time Automatic Speech Recognition Extension Toolkit of Kaldi	５．発行年 2021年
３．雑誌名 Proceedings of 2020 IEEE 10th Global Conference on Consumer Electronics (GCCE)	６．最初と最後の頁 346-350
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/GCCE53005.2021.9621992	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tomoaki Hayakawa, Chee Siang Leow, Akio Kobayashi, Takehito Utsuro, and Hiromitsu Nishizaki	４．巻 -
２．論文標題 Language and Speaker-Independent Feature Transformation for End-to-End Multilingual Speech Recognition	５．発行年 2021年
３．雑誌名 Proceedings of the 22th Annual Conference of the International Speech Communicationn Association (INTERSPEECH2021)	６．最初と最後の頁 2431-2435
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/Interspeech.2021-390	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kazumasa Yamamoto, Ryo Yamamoto, Seiichi Nakagawa	４．巻 -
２．論文標題 Effectiveness of Fine Linear Frequency Spectral Feature for Acoustic Event Detection	５．発行年 2020年
３．雑誌名 Proceedings of 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)	６．最初と最後の頁 923-924
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/GCCE50665.2020.9291954	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Masaki Okawa, Takuya Saito, Naoki Sawada, Hiromitsu Nishizaki	４．巻 -
２．論文標題 Audio Classification of Bit-Representation Waveform	５．発行年 2019年
３．雑誌名 Proceedings of the 20th Annual Conference of the International Speech Communicationn Association (INTERSPEECH2019)	６．最初と最後の頁 2553-2557
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/Interspeech.2019-1855	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Meiko Fukuda, Ryota Nishimura, Hiromitsu Nishizaki, Yurie Iribe, Norihide Kitaoka	４．巻 -
２．論文標題 A New Corpus of Elderly Japanese Speech for Acoustic Modeling, and a Preliminary Investigation of Dialect-Dependent Speech Recognition	５．発行年 2019年
３．雑誌名 Proceedings of the 22nd Conference of the Oriental COCOSDA (International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA 2019)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 福田芽衣子，西村良太，西崎博光，入部百合絵，山本一公，北岡教英
２．発表標題 超高齢者音声コーパスEARS における超高齢者の音響的特徴
３．学会等名 日本音響学会2021年秋季研究発表会
４．発表年 2021年

１．発表者名 早川友瑛，西崎博光，山本一公，小林彰夫，宇津呂武仁
２．発表標題 End-to-End複数言語音声認識モデルにおける様々なマルチタスク学習の検討
３．学会等名 日本音響学会 2020年秋季研究発表会
４．発表年 2020年

１．発表者名
Yu Wang, Hiromitsu Nishizaki , Akio Kobayashi , Takehito Utsuro, Chee Siang Leow

２．発表標題
Development and Evaluation of Kaldi Extension Tools with Python

３．学会等名
情報処理学会，音声言語情報処理研究会，2019-SLP-130(5)

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	西崎　博光 (Nishizaki Hiromitsu) (40362082)	山梨大学・大学院総合研究部・准教授 (13501)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関