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　元 年　 ６ 月 １６ 日現在

機関番号：３２６８９
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２４６５

研究課題名（和文）人の発声機構を考慮した話者固有の情報の抽出と話者照合への応用に関する研究

研究課題名（英文）A study on speaker-specific information extraction in consideration of vocalization mechanism and its application to speaker verification

研究代表者

小川　哲司（Ogawa, Tetsuji）

早稲田大学・理工学術院・准教授

研究者番号：７０３８６５９８

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：話者性と音韻性は分離可能であると仮定し，音韻の影響を受けない話者表現を得るためのニューラルネットワークを構築することを試みた．その成果として，音響特徴量から音韻性と話者性をフレーム単位で分離・抽出するディスエンタングリング・ニューラルネットワークの構築に成功した．発話単位で表出する話者情報をフレーム単位の特徴量に反映させるために統計的プーリングを導入し，特に識別の直前にプーリングを行うことの重要性を明らかにした．さらに，分離・抽出された各特徴量が各々話者および音韻の情報のみを含むように特徴抽出器を最適化するために，識別器のエントロピーに基づく損失を新たに導入しその有効性を明らかにした．

研究成果の学術的意義や社会的意義
本研究成果は，発話内容の違いの影響による話者照合性能劣化に対する本質的な解法を与えるもので，音声によるバイオメトリクス認証などアプリケーションとしての期待は高いものの依然として実用のレベルに達していない，数秒程度の短い発話に対する話者照合の性能を抜本的に改善することを可能とする．また，本研究を通じて，これまでほとんど議論されてこなかった「真の話者性」を工学的に明らかにするための新たな研究領域の開拓が期待できる．これは話者認識研究における本質的な問いであり，当該研究分野において日本のプレゼンスを示す好機ともなる．

研究成果の概要（英文）：An attempt was made to develop a neural network to learn speaker representations that are not affected by phoneme information under the assumption that speaker and phoneme information are separable on acoustic features. As the achievement, the disentangling neural network was successfully developed to extract the phoneme and speaker information separately from each frame of acoustic features. The present study introduced statistical pooling, which aims at reflecting the utterance-by-utterance speaker information to the frame-by-frame features, and demonstrated that the pooling just before classification (i.e., late pooling) performed well. In addition, a loss function based on the entropy of classifiers was introduced to optimize feature extractors such that the extracted features could contain only the desired speaker-specific and phoneme-specific information and shown to be effective in speaker verification.

研究分野：音声情報処理，パターン認識

キーワード：話者照合　特徴抽出　深層学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

話者認識（音声から個人を特定する技術）は，話者内の音響的変動の影響，特に発話内容（音韻系列）の違いの影響を受けて性能が劣化する．その逆に，音声認識（発話内容を特定する技術）では，発話者の違いの影響で性能が劣化する．そのため，音声認識では，話者の違いに頑健な特徴抽出について検討がなされている．例えば PLP（Perceptual Linear Prediction）やボトルネック特徴（深層学習により構築した音素識別器の隠れ層の出力）を用いることで，発話内容の認識に寄与しない発話者などの影響を抑圧する試みがなされている．しかし，原理的には発話者の情報が削減されているはずのこれら音声認識用特徴量は，話者認識性能をも向上させるという知見があり，申請者も追実験などを通じて同様の体験をしてきた．これは，本来異なるはずの「音韻を決めるための特徴」と「話者らしさを決めるための情報」を区別できていないことを意味する．

以上より，発話内容の影響を受けない話者の表現方法を検討することが，発話内容の差の影響で実用レベルに達していない短い発話に対する話者照合（登録データと照合データが同一人物によるものか否かを特定する技術）の性能を抜本的に改善するためにも，さらに「話者らしさとは何か？」という本質的な問いに工学的に答えるためにも，最も重要であると確信するに至った．

２．研究の目的

音声信号が持つ個人性（話者特徴）を発話内容（音韻特徴）の影響を受けずに抽出する技術を開発し，発話内容の違いの影響が原因で実用化の域に達していない短い発話に対する話者照合の性能を抜本的に改善することを目指した．そのために，人の発声機構に着目し，発話内容の情報が混入しない声帯からの信号を用いて話者特徴を抽出する技術，深層学習を用いて音響的な情報から話者を認識するための情報と発話内容を認識するための情報を分離する技術，音声認識と音声合成技術を駆使して話者特徴に含まれる発話内容の影響を直接的に取り除く技術を確立することを試みた．また，声の情報から発話者と発話内容の情報を厳密に分離する方法論を検討することで,「話者らしさとは何か？」という本質的な問いに工学的に答えることを目指した．

３．研究の方法

本研究では，以下のワークパッケージ（WP）について検討を行った．

（WP1）人の発声機構に着目し，発話内容の情報が混入しない声帯からの信号を用いて話者特徴を抽出する技術

（WP2）深層学習を用いて音響的な情報から話者を認識するための情報と発話内容を認識するための情報を分離する技術

（WP3）音声認識と音声合成技術を駆使して話者特徴に含まれる発話内容の影響を直接的に取り除く技術

以上の 3 項目を検討していく過程で，敵対的学習に基づく深層ニューラルネットワークが音韻の影響を受けにくい話者特徴抽出において有望であることが明らかになり，研究期間全体にわたり（WP2）に重きを置きながら検討を行った．

（WP2）では，話者性と音韻性は分離することが可能であると仮定し，音韻の影響を受けない話者表現を得るためのニューラルネットワークを構築することを試みた．具体的には，音響特徴量から音韻性と話者性をフレーム単位で分離・抽出するディスエンタングリング・ニューラルネットワークを提案した．提案モデルにより，音韻の変動に対して不変な特徴量が得られることが期待できる．実験では，発話単位で表出する話者情報をフレーム単位の特徴量に反映させるために統計的プーリングを導入し，特に識別の直前にプーリングを行うことの重要性を明らかにした．さらに，分離・抽出された各特徴量が各々話者および音韻の情報のみを含むように特徴抽出器を最適化するために，識別器のエントロピーに基づく損失を新たに導入し，その有効性も明らかにした．

このように，提案する話者特徴表現の学習法は音素非依存な話者表現をフレーム単位で得るために極めて有効であると言え，研究期間全体として概ね順調に研究が推移したと言える．

４．研究成果

（成果の概要）

本研究の主な成果は，以下の 2 点である．

①　音響特徴量から音韻性と話者性をフレーム単位でそれぞれ分離・抽出するニューラルネッ

トワークの構造およびそれを学習するための損失関数を明らかにしたこと．

② 音響特徴量から表出単位がフレーム単位・発話単位のように異なる音韻性と話者性を分離・抽出するための技術と知見を得たこと．

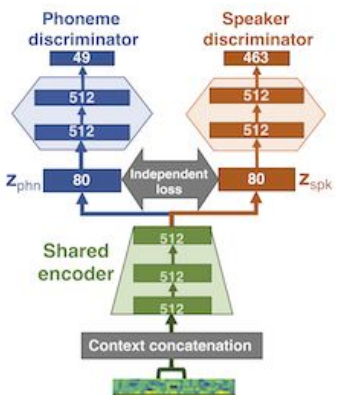
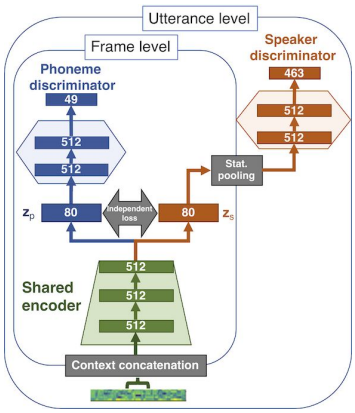
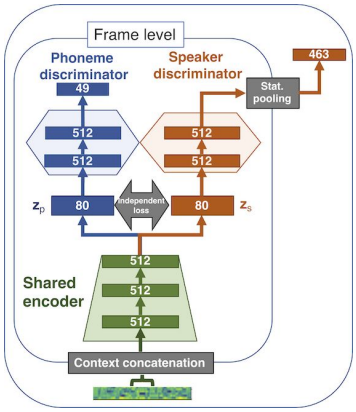


図 1: ディスエンタングリング・ニューラルネットワークの基本構造



(a) アーリープーリング



(b) レイトプーリング

図 2：統計的プーリング

（成果の詳細）

本研究で提案したディスエンタングリング・ニューラルネットワークの基本構造を図 1 に示す．本ネットワークは話者識別器と音素識別器，そして両者に共有のエンコーダから構成される．当該フレームの前後 10 フレームの音響特徴ベクトルを連結し共有エンコーダに入力すると，話者ボトルネック特徴量と発話ボトルネック特徴量が得られる．各ボトルネック特徴を音素識別器と話者識別器に入力したときに，中心フレームに付与された音素ラベルと話者ラベルに対する音素識別損失と話者識別損失が最小となるようにネットワーク全体を学習した．このとき，話者性と音韻性をより明確に分離できるように，一方のボトルネック特徴量をもう一方の識別器に入力したときのエントロピーを損失として新たに導入した．

ここで，音韻性はフレーム単位で表れるのに対し，話者性は複数のフレームからなるセグメント単位で表れることが知られている．それに対し，セグメントレベルで話者識別損失を算出するため統計的プーリング層を導入した．統計的プーリング層は入力されたベクトルの平均と分散をそれぞれ算出しその結合ベクトルを出力する関数であり，可変長の入力から固定長のベクトルを得るために用いられる．

一般に話者認識では，セグメント単位で話者表現を得ることが目的であるため，ボトルネック特徴量をプーリングしてから話者識別器に入力するアーリープーリング（図 2(a)）が用いられる．しかし，本研究ではフレーム単位で話者表現を得ることが目的であるため，話者識別損失算出時にプーリングを行うことで，ボトルネック特徴量に話者性が残ることが期待されるレイトプーリング（図 2(b)）を新たに導入した．

提案するディスエンタングリング・ニューラルネットワークから抽出されたボトルネック特徴量の話者性と音韻性を，話者識別および音韻識別実験によりそれぞれ評価したところ，以下の知見が得られた．

① ディスエンタングリング・ニューラルネットワークは有効である．
話者ボトルネック特徴量と音素ボトルネック特徴量を別々に抽出する構造が有効であった．

② 統計的プーリングは有効である．
統計的プーリングを用いずフレーム単位で得られた話者識別損失を用いてエンコーダを最適化した場合，未知話者に対する話者識別率が著しく低下した．その一方で，統計的プーリングを導入し，発話単位で得られた話者識別損失を用いることで，未知話者に対して高い識別性能が得られた．

③ レイトプーリングは有効である．
話者認識において一般的に用いられる，ボトルネック特徴量に対するアーリープーリングよりも，識別誤差を算出する直前にプーリングを行うレイトプーリングの方が高い性能を与えた．これは，従来の話者認識では発話単位で話者表現を得ることが目的であるため，早い段階でのプーリングが有用であったのに対し，本研究で対象とするフレーム単位の話者表現を得るためには，ボトルネック層にできるだけ話者情報が残るように，識別損失算出の直前までプーリングを行わないことが有効であったためと考えられる．フレーム単位

での話者表現を得ることに注力したモデルであるため，従来話者照合性能が低い短い発話においても頑健に高い性能が得られることが期待できる．

④ エントロピー損失は有効である．
話者識別器および音韻識別器各々のボトルネック特徴量が所望の情報のみを持つよう導入したエントロピーに基づく損失により，話者識別性能も音素識別性能も改善した．つまり，エントロピーに基づく損失は，各ボトルネック特徴量における音韻情報と話者情報の分離性能の改善に効果があることが明らかになった．

⑤ 直交制約は効果がない．
④と同様の目的でボトルネック特徴量に対して直交制約を導入したが，効果は得られなかった．有効性が示されていた類似研究では隠れ層を持たない線形識別器が用いられていたのに対し，本研究では多層の識別器を用いているため，直交性の制約が識別器に吸収されてしまい有効に機能しなかったためと考えられる．

５．主な発表論文等

〔雑誌論文〕（計 4 件）

① Naohiro Tawara, Hikari Tanabe, Tetsunori Kobayashi, Masaru Fujieda, Kazuhiro Katagiri, Takashi Yazu, Tetsuji Ogawa, ``Postfiltering using an adversarial denoising autoencoder with noise-aware training," Proc. ICASSP2019, pp.3282-3286, May 2019. [doi: 10.1109/ICASSP.2019.8682684]
② Naohiro Tawara, Tetsunori Kobayashi, Masaru Fujieda, Kazuhiro Katagiri, Takashi Yazu, Tetsuji Ogawa, ``Adversarial autoencoder for reducing nonlinear distortion," Proc. APSIPA2018, pp.1669-1673, Nov. 2018. [doi: 10.23919/APSIPA.2018.8659540]
③ Taira Tsuchiya, Naohiro Tawara, Tetsunori Kobayashi, Tetsuji Ogawa, ``Speaker invariant feature extraction for zero-resource languages with adversarial training," Proc. ICASSP2018, pp.2381-2385, April 2018. [doi: 10.1109/ICASSP.2018.8461648]
④ Tetsuji Ogawa, Harish Mallidi, Emmanuel Dupoux, Jordan Cohen, Naomi Feldman, Hynek Hermansky, ``A new efficient measure for accuracy prediction and its application to multistream-based unsupervised adaptation," Proc. ICPR2016, pp.2222-2227, Dec. 2016. [doi: 10.1109/ICPR.2016.7899966]

〔学会発表〕（計 10 件）

⑤ 樋口陽祐，俵直弘，小林哲則，小川哲司，``DPGMM と敵対的学習に基づく話者の違いに頑健な特徴抽出とゼロリソース音声認識での評価，" 情報処理学会研究報告 (SLP)，July 2019.（発表予定）
⑥ 田辺ひかり，俵直弘，小林哲則，藤枝大，片桐一浩，矢頭隆，小川哲司，``敵対的デノイジングオートエンコーダを用いた拡散性雑音除去，" 電子情報通信学会技術研究報告(SP), SP2018-87, pp.155-160, March 2019.
⑦ 俵直弘，田辺ひかり，小林哲則，藤枝大，片桐一浩，矢頭隆，小川哲司，``noise-aware 学習を用いた敵対的デノイジングオートエンコーダによるポストフィルタリング，" 日本音響学会講演論文集，pp.159-162, March 2019.
⑧ 樋口陽祐，俵直弘，小川哲司，小林哲則，``ゼロリソース言語音声認識のための発話者の違いに頑健な特徴抽出，" 日本音響学会講演論文集，pp.923-924, March 2019.
⑨ 俵直弘，小林哲則，小川哲司，``音韻・話者特徴抽出のためのディスエンタングリングニューラルネットワークの実現にむけて，" 日本音響学会講演論文集, pp.1003-1004, March 2019.
⑩ 俵直弘，小林哲則，藤枝大，片桐一浩，矢頭隆，小川哲司，``敵対的デノイジングオートエンコーダによる非線形ひずみ除去フィルタリング，" 日本音響学会講演論文集，pp.159-160，Sept. 2018.
⑪ 俵直弘，小林哲則，藤枝大，片桐一浩，矢頭隆，小川哲司，``非線形ひずみ除去のための敵対的 denoising autoencoder," 情報処理学会研究報告，2018-SLP-123(1), pp.1-5, July 2018.
⑫ 俵直弘，土屋平，小川哲司，小林哲則，``敵対的学習に基づく話者特徴抽出，'' 日本音響学会講演論文集, pp.141-144, March 2018.
⑬ 島田拓也，俵直弘，小川哲司，小林哲則，``話者正規化における言語非依存性とゼロリソース音声認識における効果，'' 日本音響学会講演論文集, pp.109-112, March 2018.
⑭ 土屋平，俵直弘，小川哲司，小林哲則，``敵対的学習を用いた話者の違いに頑健な特徴抽出とゼロリソース音素識別による評価，'' 日本音響学会講演論文集，pp.9-12，March 2018.

６．研究組織

(2)研究協力者
研究協力者氏名：俵 直弘
ローマ字氏名：Naohiro Tawara