TR-I-0327

ニューラルネットワークを用いた話者適応化および不特定話者連続音声認識の研究
- 発表論文リスト -

A Study of Speaker Adaptation and Speaker-Independent Continuous Speech Recognition Using Neural Networks
- A List of Papers -

福沢圭二
Keiji FUKUZAWA

1993.03

概要

著者が1990年4月1日から1993年3月31日まで、ＡＴＲ自動翻訳電話研究所で行なったニューラルネットワークを用いた音声認識に関する研究発表論文リストを載せた報告書である。研究発表の内容は、「ニューラルネットワークを用いた話者適応」「ニューラルネットワークを用いた不特定話者音声認識」「ニューラルネットワークを用いた単語予備選択」に関するものである。また、本テクニカルレポートの最後に、不特定話者FPM-LRを用いた機能試験文600文(F_set)に対する評価結果を載せる。

目次

1 はじめに

本テクニカルレポートに、ニューラルネットワークを用いた音声認識に関する著者が発表した論文のリストを載せる。著者は、ニューラルネットワークを用いた大語彙連続音声認識システムを構築することを目的とし研究を行なってきた。また、不特定話者を対象としたより高い性能を得るために、ニューラルネットワークによる特徴量写像による話者適応、多数話者音声資料を用いたニューラルネットワークの音素識別学習の研究を行い、それぞれの方法により不特定話者を対象とした連続音声認識の高性能化が図れることを示してきた。

今までの主な研究の概要を以下に示す。

1. ニューラルネットワークを用いた話者適応化方式

   ニューラルネットワークを用いた話者適応化方式により、未知話者に対する認識システムの性能向上を目指した。従来の話者適応化方式ではフレーム単位にスペクトル構造の適応を行なっていたが、本研究ではセグメントを単位とし話者間の時間構造も含めた適応を検討した。未知話者に対する音素認識実験の結果、本手法の有効性が確認された。 [1, 15]

2. 話者適応ニューラルネットワークとTDNN-LRを組み合わせた連続音声認識

   セグメントベース話者適応ニューラルネットワークとTDNN-LR連続音声認識システムを組み合わせ、未知話者に対する単語及び文節認識の性能を評価した。未知話者に対する評価の結果、本適応化方式の単語及び文節認識に対する有効性が確認された。 [4, 5, 7, 12, 16]

3. ニューラルネットワークを用いた教師なし話者適応化方式

   教師あり話者適応化は、教師信号に基づいて適応化の写像を構成するため強力であるが、あらかじめ決められた発声を行なわなければならないと言う点で柔軟性に乏しい。これに対して、教師なし話者適応化は、任意の発声を基に適応を行なえるため柔軟性に富みその用途が広い。本研究ではVQとニューラルネットが共に歪み最小化の基準で動作している点に着目し、VQとニューラルネットを組み合わせた歪み最小化の基準によるアルゴリズムと階層的クラスタリングによるアルゴリズムを組み合わせた教師なし話者適応方式の有効性を評価した。音素識別による評価実験の結果本手法の有効性が確認された。 [2, 6, 11]

4. 多数話者学習TDNN-LRによる不特定話者連続音声認識

   TDNNを用いた連続音声認識システムの不特定話者に対する性能向上を目指す。多数話者の音声データにより学習を行なったTDNNを用いてTDNN-LRを構成し、単語及び文節認識の性能を評価した。その結果、セグメントベース話者適応ニューラルネットワークと標準話者学習TDNN-LRの組み合わせとほぼ同等の性能が確認され、本手法が有効であることが判った。 [3, 7]

5. FPM-LRを用いた不特定話者連続音声認識

   素子出力の和が定数になるような拘束条件を持つニューラルネットワーク (FPM: Fuzzy Partition Model) を適用した不特定話者連続音声認識システム (FPM-LR) の性能を評価した。その結果、FPMは学習時間がTDNNの2分の1以下で済み認識性能の面でもTDNNを上回っていること、複数のFPM-LRを用いることで不特定話者に対する認識性能の向上が図れることが示され、男女話者に対する文節認識率80.0%が達成された。また、文節発声文音声認識においても良好に動作することが確認された。 [8, 9, 13, 14, 17, 18]

6. ニューラルネットワークの発火パターンを単語候補の予備選択に利用した大語彙単語認識

   ニューラルネットワークの発火パターンを利用した単語予備選択法を提案した。予備選択はニューラルネットワークによる音素スキャンニングにより得られる単語全体の音素ベクトルを用いて行ない、認識は予備選択された候補を対象として発火パターン列とLRパーザを用いて行なう。評価の結果、大語彙単語認識における処理時間の短縮が図れるこが示された。 [10]

2 発表論文リスト

2.1 日本音響学会 - 研究発表会

[1] 福沢 圭二, 沢井 秀文, 杉山 雅英: “ニューラルネットワークによる恒等写像を用いた話者適応,” 日本音響学会平成2年度秋季研究発表会講演論文集, 1-8-16, pp.31-32 (1990.09).

[2] 杉山雅英, 福沢圭二, 沢井秀文, 嵯峨山茂樹: “ニューラルネットワークによる集合間写像の教師なし話者適応,” 日本音響学会平成 2 年度秋季研究発表会講演論文集, 2-P-10, pp.149-150 (1990.09).

[3] 沢井秀文, 中村 悟, 福沢圭二, 杉山雅英: “ニューラルネットワークによる不特定話者音声認識へのアプローチ法について,” ’日本音響学会平成 3 年度春季研究発表会講演論文集, 1-5-17, pp.37-38(1991.03).

[4] 小森 康弘, 福沢 圭二, 杉山 雅英, A. H. WAIBEL, 嵯峨山 茂樹: “ニューラルファジー学習法の連続音声認識における効果,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 2-5-11, pp.69-70 (1991.10).

[5] 福沢 圭二, 小森 康弘, 沢井 秀文, 杉山 雅英: “セグメントベース話者適応ニューラルネットワークを用いた文節音声認識,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 3-5-10, pp.109-110 (1991.10).

[6] 福沢 圭二, 杉山 雅英: “ニューラルネットワークによる教師なし話者適応法とその評価,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 3-5-11, pp.111-112 (1991.10).

[7] 福沢 圭二, 小森 康弘, 杉山 雅英: “TDNN-LR 連続音声認識における不特定話者 TDNN と話者適応ニューラルネットワークの性能比較,” 日本音響学会平成 4 年度春季研究発表会講演論文集, 2-Q-21, pp.199-200 (1992.03).

[8] 福沢 圭二, 加藤 喜永, 杉山 雅英: “FPM-LR を用いた不特定話者連続音声認識,” 日本音響学会平成 4 年度秋季研究発表会講演論文集, 3-7-3, pp.167-168 (1992.10).

[9] 山口 耕市, 永井 明人, 鷹見 淳一, 大倉 計美, 小坂 哲夫, 福沢 圭二, 他: “ATREUS: ATR における連続音声認識諸方式の比較,” 日本音響学会平成 4 年度秋季研究発表会講演論文集, 2-Q-5, pp.181-182 (1992.10).

[10] 福沢 圭二, 杉山 雅英: “ニューラルネットワークを利用した予備選択による大語彙単語音声認識,” 日本音響学会平成 5 年度春季研究発表会講演論文集, 2-Q-20, (1993.03).

2.2 電子情報通信学会 - 大会

[11] 福沢 圭二, 杉山 雅英: “階層的クラスタリングと Neural Network を用いた教師なし話者適応法,” 電子情報通信学会秋季大会講演論文集, Vol. 1, pp.271-272 (1992.09).

2.3 電子情報通信学会 - 音声研究会

[12] 福沢 圭二, 小森 康弘, 沢井 秀文, 杉山 雅英, “セグメントベース話者適応ニューラルネットワークと TDNN-LR を用いた文節音声認識,” 電子情報通信学会技術研究報告, SP91-105, pp.23-29 (1992.01).

[13] 福沢 圭二, 加藤 喜永, 杉山 雅英: “FPM-LR を用いた不特定話者連続音声認識の実現,” 電子情報通信学会技術研究報告, SP92-107, pp.31-38 (1992.12).

[14] 永井 明人, 山口 耕市, 鷹見 淳一, 大倉 計美, 小坂 哲夫, 福沢 圭二, 他: “ATR における連続音声認識システム ”ATREUS” の諸方式と性能,” 電子情報通信学会技術研究報告, SP92-122, pp.51-58 (1993.01).

2.4 国際会議

[15] Keiji Fukuzawa, Hidehumi Sawai, Masahide Sugiyama: “Segment-based Speaker Adaptation by Neural Network,” IEEE workshop on Neural Networks for Singnal Processing, pp.442-451, (1991.09).

[16] Keiji Fukuzawa, Yasuhiro Komori, Hidehumi Sawai, Masahide Sugiyama: “A Segment-based Speaker Adaptation Neural Network Applied to Continuous Speech Recognition,” Proc. of 1992 International Conference on Acoustics, Speech, and Signal Processing, 55.1, Vol.1, pp.433-436 (1992.03).

[17] Keiji Fukuzawa, Yoshinaga Kato and Masahide Sugiyama: “A Fuzzy Partition Model Neural Network Architecture for Speaker-independent Continuous Speech Recognition,” Proc. of 1992 International Conference on Spoken Language Processing, Vol.2, pp.1383-1386 (1992.10).

[18] S. Sagayama, M. Sugiyama, K. Ohkura, J. Takami, A. Nagai, H. Singer, H. Hattori, K. Fukuzawa, Y. Kato, K. Yamaguchi, J. Murakami, and A. Kurematsu: “ATREUS: Continuous Speech Recognition Systems at ATR Interpreting Telephony Research Laboratories,” Proc. of 1992 Speech Science and Technology, pp324-329 (1992.12).

3 テクニカルレポート一覧

- TR-I-0192 福沢圭二、中村雅巳: “ニューラルネット・ワークベンチシステム - 操作マニュアル -,” (1991.02)

  概要：操作性、可視性の優れたニューラルネット開発環境を提供することを目的として、ニューラルネット・ワークベンチシステムが開発された。本報告ではこのワークベンチシステムの機能、操作方法について述べる。

- TR-I-0222 山本 雅章、福沢圭二、杉山雅英 : “VQ Neural Network による教師なし話者適応,” (1991.07).

  概要：教師あり話者適応化は、教師信号に基づいて適応化の写像を構成するため強力であるが、あらかじめ決められた発声を行なわなければならないと言う点で柔軟性に乏しい。これに対して、教師なし話者適応化は、任意の発声を基に適応を行なえるため柔軟性に富みその用途が広い。本研究では VQ とニューラルネットが共に歪み最小化の基準で動作している点に着目し、VQ とニューラルネットを組み合わせた教師なし適応化方式の有効性を評価した。5 母音による評価実験の結果、72.0% であった適応化前の認識率が、適応化によりの認識率 94.4% となり、本手法の有効性が確認された。

4 出願特許一覧

- 福沢圭二、沢井秀文、 杉山雅英 :“ニューラルネットワークによる話者適応化方式,” (1990.09).

  概要：入力された音声から音声特徴パターンを抽出する音声特徴抽出部と、ニューラルネットワークを用いて未知話者の音声特徴パターンを標準話者の音声特徴パターンへ写像する適応化部と、適応化された音声特徴パターンの認識を行なう認識部を備え、ニューラルネットワークの話者適応化学習の前段階で標準話者の音声特徴パターンを用いて恒等写像の学習を行なうことを特徴とする話者適応化方式。

- 福沢圭二、杉山雅英 :“ニューラルネットワークの発火パターンを用いた単語予備選択方式,” (1993.02).

  概要：ニューラルネットワークの発火パターンを用いて、認識対象とする単語の中から候補を絞り、候補となる単語のみを対象として認識を行なうため認識時の処理時間を短縮可能な方法であり、大語彙単語を対象とした音声認識の分野に適用可能である。

Appendix

A 機能試験文 600 文 (F_set) に対する不特定話者 FPM-LR の性能

不特定話者 FPM-LR の機能試験文 600 文に対する認識性能を以下に示す。

A.1 実験条件

実験条件を以下に示す。

- **音響特徴量：**

  ニューラルネットワークに入力する音響特徴量として、メル 16 チャネル対数 FFT スペクトラム×7 フレーム (70ms) にパワーとデルタスペクトラムを加えた特徴量用いる [13]。

- **学習話者と評価話者の設定：**

表 1: 評価話者設定

話者設定	学習	評価
男性話者	男性 8 名 (MHT MMS MMY MSH MTK MTM MTT MXM)	男性 2 名 (M01 M02)
女性話者	女性 8 名 (FAF FFS FKN FKS FMS FSU FYM FYN)	女性 2 名 (F01 F02)
男女混合	男女各 8 名 (MHT MMS MMY MSH MTK MTM MTT MXM FAF FFS FKN FKS FMS FSU FYM FYN)	男女 各 2 名 (M01 M02 F01 F02)

- **FPM の学習：**

  各学習話者発声の重要単語 5240(A-set) の偶数番目の単語を用いる。学習に用いる音素サンプルは、音素ラベル情報に基づいて取り出す。男性話者、女性話者の学習には 50,000 音素サンプルを用い、男女混合の学習においては 100,000 音素サンプルを用いる。

- **認識タスク：**

  機能試験文 600 文 (F_set) を用いる。但し、実際の評価においては LR 文法を通らない 41 文を省いた 559 文に対して評価した。また、ここに含まれる文節数は 2,166 であり 1 文あたりの平均文節数は 3.87 である。

- **LR 文法のサイズ：**

  認識に用いた LR 文法のサイズを表 2に示す。

表 2: 文法のサイズ

	文節内文法	文節間文法
語彙数	1,656	110 文節カテゴリ
規則数	2,758	421

- **LR パーザのビーム幅：** 100 に設定。

A.2　評価結果

Single-FPM-LR および Multi-FPM-LR[13] による評価結果を以下に示す。

- **Single-FPM-LR による評価：**

  男性話者、女性話者、男女混合の各評価話者設定における Single-FPM-LR の文節および文認識性能を表 3 に示す。文認識は、ラベル情報を用いて文節単位に音声を切り出し文節毎の認識を行い、それらの文節候補列に基づいて、文節間文法を用いて行なう。

表 3: Single-FPM-LR による 2166 文節認識率および 559 文認識率

話者設定	認識率 (%)	
設定	文節	文
男性話者	72.8 (89.1)	29.5 (44.4)
女性話者	67.3 (84.1)	22.3 (33.6)
男女混合	61.6 (82.1)	19.7 (31.0)

( )：認識候補第 5 位までの累積認識率

- **Multi-FPM-LR による評価：**

  男性話者、女性話者、男女混合の FPM-LR を組み合わせた Multi-FPM-LR による文節および文認識性能を表 4 に示す。評価話者は、男女各 2 名 (M01, M02, F01, F02) である。

表 4: Multi-FPM-LR による 2166 文節認識率および 559 文認識率

認識率 (%)	
文節	文
68.5 (85.8)	24.6 (37.4)

( )：認識候補第 5 位までの累積認識率