論文 / 著書情報
Article / Book Information

論題(和文)	音韻レベルの話者情報を用いた音声認識における話者適応
Title(English)	Speaker adaptation in speech recognition using phonological level speaker information
著者(和文)	伊藤光一, 篠田浩一
Authors(English)	Koichi Ito, Koichi Shinoda
出典(和文)	日本音響学会第153回(2025年春季)研究発表会講演論文集, , , pp. 991-992
Citation(English)	, , , pp. 991-992
発行日 / Pub. date	2025, 3

音韻レベルの話者情報を用いた音声認識における話者適応*

☆伊藤光一, 篠田浩一 (東京科学大)

1 はじめに

音声認識は音声をテキストに変換する技術であり，スマートスピーカーや会議記録システム，音声翻訳などのベースとなっている。近年の深層学習ベースの音声認識は，モデルとデータの大規模化に伴い高い精度を記録するようになった。しかし，雑音下や複数話者条件下などで課題が残り，話者適応が重要である。従来は深層学習における話者情報の利用では発話全体に対する特徴が利用されてきたが，話者の違いは音韻レベルにも現れる。本研究では音声認識における話者適応について，深層学習ベースの音韻レベルの細かい話者情報を用いたマルチタスク学習手法を提案する。話者情報の利用方法について複数の手法を比較検討するための実験を行い，その結果を示す。

2 関連研究

音声認識モデルとして現在 Hybrid CTC/Attention Architecture [1] が高い精度を記録している。Encoder と Decoder で協調して学習と推論が行われ，推論の際は言語モデルも使用されることがある。Encoder は Transformer を音声向けに改良した Conformer や Branchformer などが使用される。しかし，人の発話する音声は年齢や性別，構音障害の有無といった発話者の性質によって特性が変化し [2, 3]，音声認識結果へ影響を与えることが知られる。複数話者や雑音下環境などの対象話者の声が聞き取りにくくなる状況でも悪影響を与える [4]。このような場合，話者情報を利用することで音声認識の精度を向上させてきた。

話者情報として使われる話者埋め込みは以前は i-vector などが用いられた。近年の深層学習ベースでは ECAPA-TDNN などの CNN ベースのモデルや，Conformer を利用した MFA-Conformer [5] などのモデルが知られる。

話者情報を利用して音声認識の精度向上を図るのが，話者適応である。Wagner ら [4, 6] は，対象となる発話あるいは対象話者全ての発話の平均の i-vector を用いて Encoder に投入する話者適応手法を提案した。Geng ら [3, 7] は，事前に深層学習ベースのモデルで話者 ID や年齢で訓練された特徴量を用いた話者適応手法を提案した。しかし，話者の特性の表れ方は音韻ごとに異なり [8]，より細かい単位での話者適応による精度向上の可能性がある。フレームごとの i-vector を使用した話者適応に関して，Peddinti ら [9] は学習時にフレームごとの i-vector を用いて学習したが，推論時は平均した情報を使用した。変動の多いフレームごとの i-vector を使用することで少ない話者数での多様性の確保を図ったとしている。また，Soni ら [7] は発話単位の x-vector に加えてフレーム単位での i-vector を使用する実験を行った。

i-vector は発話全体の特徴としては変動の平均を取るため安定するが，そのフレーム単位で見た場合には発話系列における内部相関を考慮した特徴ではないため話者埋め込みとして問題がある可能性がある。また，近年では発話内相関を考慮できる深層学習モデルによる発話埋め込みが，i-vector による発話埋め込みよりも優れた値を記録している。

3 提案手法

本研究では音声認識において音韻性を明示的に考慮するために，発話内相互関係を考慮した深層学習ベースによるフレーム単位での話者情報を用いた話者適応手法を提案する。Fig. 1 が提案手法の概要である。話者埋め込み，音韻埋め込みが共に深層学習ベースであり，マルチタスク学習が可能な構造である。

話者 Encoder の出力から Adapter 層への入力は，全ての層の出力を用いる。Adapter 層では各フレームの話者埋め込みを 2 層の線形層を通して，各フレームの話者埋め込みに加算する。ここで最小化する損失関数 L は，話者 Encoder の損失 L_{spk}，音韻 Encoder の CTC ロス L_{ctc}，音韻 Decoder の損失 L_{att} を用いて，

$$L = \alpha L_{\mathrm{ctc}} + (1-\alpha) L_{\mathrm{att}} + \beta L_{\mathrm{spk}}$$

である。

3.1 話者埋め込みの投入方法

話者の安定性のために以下の 4 つを比較検討する。

- 分類機直前の埋め込みを利用する (Utterance;U)
- フレーム単位の話者 Encoder の出力のみを利用する (Frame;F)
- フレーム単位の話者 Encoder の出力に対して幅 w で平均を取って利用する (Mean(w);M(w))
- フレーム単位の話者 Encoder の出力と分類機直前の埋め込みを利用する (U+F)

*Speaker adaptation in speech recognition using phonological level speaker information. by ITO, Koichi, SHINODA, Koichi (Institute of Science Tokyo)

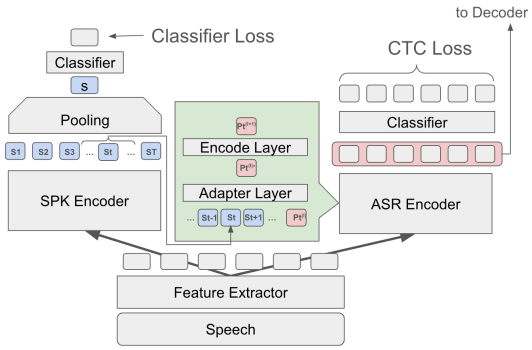

Fig. 1　提案手法概要

4　実験

4.1　実験設定

実験コードは ESPnet をベースに実装した。データセットは LibriSpeech (960h) である。特徴量は WavLM の出力を用い，話者 Encoder は 2 層の MFA-Conformer，音韻 Encoder は 2 層の Conformer，音韻 Decoder は 6 層の Transformer である。学習時の損失について，$\alpha = 0.3$，$\beta = 0.1$ である。推論時は，CTC 出力と Decoder 出力に加えて言語モデルも使用する。

評価として比較のため音声認識タスクのみ (ASR) および，話者認証タスク (SPK) のみの学習モデルの結果も掲載する。音声認識の精度比較は単語誤り率 (Word Error Rate;WER) を用い，Librispeech の test-other(O)，test-clean(C) を用いる。また，話者認証の評価は，等価誤り率 (Equal Error Rate;EER) を用いる。LibriSpeech における話者埋め込みを評価するため，LibriSpeech の開発，テストセットを用いて話者認証用データセットを作成した。なお，開発，テストセットの話者は学習セットと重複がなく，今回作成したデータセットは特定話者に偏りすぎないようにしてある。

4.2　実験結果

実験結果を Table 1 に示す。最もシンプルな話者適応手法である発話平均を用いた手法 U について，話者適応手法を使用しない手法よりやや悪化した。また，フレーム単位の情報を用いた話者適応手法 F について，発話埋め込みの精度，音声認識結果がいずれも大きく悪化した。フレーム単位の情報として，周囲を用いた平均プーリングを行う場合，長さが長くなるにつれて WER が改善した。

また，深層学習ベースの話者埋め込みとその長さに関する関係を調べるため，発話全体を入力後，プーリング直前に特定のフレーム数に切り取ってプーリング層に入力した場合の評価結果 Table 2 に示す。使用したモデルは今回実験で用いた SPK モデルである。長さが特に短い場合，話者識別評価は悪化する。

話者適応を試みた場合に性能が下がる問題について，Mean(w) と Table 2 の結果から話者埋め込みの同時最適化では音声認識性能を上げるには難しい可能性が考えられる。さらに，手法 U や F になるにつれて，損失の推移に過学習の傾向が見られた。話者埋め込みとして与える情報量が過多な可能性があり，より緻密な制御が必要な可能性がある。

Table 1　手法ごとの音声認識と話者識別の評価

Model	WER(O)[%]	WER(C)[%]	EER[%]
ASR	3.2	1.7	-
SPK	-	-	0.60
F	5.9	3.3	6.10
M(3)	4.1	2.0	5.11
M(7)	3.9	2.0	32.70
M(15)	3.5	1.8	35.16
M(21)	3.5	1.8	44.27
U	3.4	1.7	4.16
U+F	6.5	3.4	6.96

Table 2　切り取り長ごとの話者埋め込み評価結果

Length	3	6	12	25	50	100
EER[%]	14.78	7.35	2.71	0.98	0.65	0.57

5　おわりに

本研究では音声認識における話者適応について，深層学習ベースの音韻レベルの細かい話者情報を用いたマルチタスク学習手法を提案した。

参考文献

[1] Watanabe *et al.*, IEEE Journal of Selected Topics in Signal Processing, 11(8) , 1240–1253, 2017.
[2] 粕谷 他, 日本音響学会誌, 24(6), 355–365, 1968.
[3] Geng *et al.*, TASLP, 30, 2597–2611, 2022
[4] Wagner *et al.*, ASRU, 1–6, 2023
[5] Yang *et al.*, Interspeech, 306–310, 2022
[6] Saon *et al.*, IEEE Workshop on Automatic Speech Recognition and Understanding, 55–59, 2013
[7] Soni *et al.*, APSIPA ASC, 833-839, 2022
[8] Siyuan *et al.*, Computer Speech & Language, 84, 101567, 2024.
[9] Peddinti *et al.*, ASRU, 539–546, 2015