超多言語事前学習による低資源音声認識の検討 *

☆侯汶昕，董越，庄佰融，楊龍飛（東工大），史嘉彤（JHU）, 篠崎隆宏 (東工大)

1 はじめに

End-to-end 音声認識モデルは構成・拡張の柔軟性が高いだけでなく、近年では認識性能の点でも従来の隠れマルコフモデル (HMM) を用いたシステムを超える性能が得られつつある [1]。しかし、データが少ない言語（低資源言語）で高い認識性能を得ることは依然として困難である。Cho 等は 10 言語のデータを用いて事前学習したモデルを 4 言語に適応する転移学習の有効性について報告している [2]。本研究では 42 言語を含む計約 5000 時間の学習データを用いて音声認識モデルを事前学習する、超多言語事前学習を行う。得られたモデルを初期モデルとし 14 の低資源言語への転移学習を行い、言語による違い等について比較検討する。

2 多言語音声認識

2.1 言語非依存認識器

多言語認識を行うために、認識対象のすべての言語においてネットワーク構造とパラメーターを共有する言語非依存認識器 [3] を用いる。すべての認識対象言語の文字集合の和が出力語彙である。認識発話中で不用意に言語が切り替わってしまう可能性を減らすために、学習時に発話ラベルの先頭に言語 ID（例えば [en]、[fr] など）を挿入する。認識時にモデルは最初に言語を識別した後に出力テキストを予測する。すなわち言語識別を補助タスクとして用いる。

2.2 低資源言語への転移学習

事前学習済みモデルをもとに、低資源言語への適応化を行う。本研究では特定の低資源言語への適応化（単一言語転移）とともに、複数の低資源言語に同時に適応化（多言語転移）する場合についても検討する。言語間転移学習は事前学習したネットワークの出力層を転移先言語用に置き換えた上で、転移先言語の少量のラベル付きデータを用いてネットワークを再学習することにより行う。多言語転移する場合は、事前学習時と同様に言語 ID を出力ラベルの先頭に追加する。

Table 1 低資源言語データ

言語	継続時間
Arabic	7
Breton	5
Hakka Chins	2
Chuvash	0.96
Dhivehi	6
Esperanto	35
Estonian	10
Indonesian	3
Interlingua	1
Kinyarwanda	0.25
Kyrgyz	11
Latvian	4
Sakha	3
Slovenian	3

3 評価実験

3.1 実験設定

事前学習には、42 言語を含む 11 種類のコーパス（AISHELL, Aurora4, Babel, Common Voice, CSJ, CHiME4, Fisher Callhome, Fisher Switchboard, Voxforge, WSJ, HKUST）の音声データを用いた。総発話数は 600 万以上で、音声データは約 5000 時間である。転移先低資源言語として、Table 1に示す Common Voice データベース [4] の 14 言語を用いる。14 言語中の 12 言語の音声データ量は 10 時間以下である。一番データの少ないのは Kinyarwanda の 0.25 時間であり、一番多いのは Esperanto の 35 時間である。実験では、各言語においてランダムに 80 %を訓練セット、20 %を評価セットとして用いた。

認識に用いる音響特徴量は 80 次元のフィルターバンクと 3 次元のピッチ特徴の計 83 次元のベクトルである。認識器には Transformer を用いた Hybrid CTC/Attention アーキテクチャを用いる [5]。Transformer は、[6] で説明されている *big model* と同じ構成である。42 言語のデータを用いて事前学習した認識器を初期モデルとする場合と、乱数初期化したモデルから低資源言語を直接学習する場合について認識性能を比較する。ニューラルネットの学習・評価には

* A Study on Low-resource Speech Recognition by Multilingual Pre-training. by HOU, Wenxin and DONG, Yue and ZHUANG, Bairong and YANG, Longfei (Tokyo Institute of Technology) and SHI, Jiatong (Johns Hopkins University) and SHINOZAKI, Takahiro (Tokyo Institute of Technology)

Table 2　単一言語転移学習

言語	事前学習なし	事前学習あり
Arabic	42.0	**14.2** (66.2%↓)
Breton	39.5	**16.3** (58.7%↓)
Hakha Chin	34.2	**11.0** (67.8%↓)
Chuvash	**51.3**	68.4 (33.3%↑)
Dhivehi	31.1	**10.3** (66.9%↓)
Esperanto	5.2	**2.3** (55.8%↓)
Estonian	30.2	**9.9** (67.2%↓)
Indonesian	33.2	**11.3** (66.0%↓)
Interlingua	35.1	**18.2** (48.1%↓)
Kyrgyz	21.8	**7.7** (64.7%↓)
Latvian	30.6	**9.4** (69.3%↓)
Kinyarwanda	**84.8**	169.5 (99.9%↑)
Sakha	36.4	**13.3** (63.5%↓)
Slovenian	26.1	**23.2** (11.1%↓)
加重平均	23.0	**9.6** (58.3%↓)

Table 3　多言語転移学習

言語	事前学習なし	事前学習あり
Arabic	19.7	**15.7** (20.3%↓)
Breton	21.1	**15.8** (25.1%↓)
Hakha Chin	14.7	**10.0** (32.0%↓)
Chuvash	19.2	**14.4** (25.0%↓)
Dhivehi	13.7	**10.8** (21.2%↓)
Esperanto	3.8	**2.7** (28.9%↓)
Estonian	14.5	**10.1** (30.3%↓)
Indonesian	14.5	**10.1** (30.3%↓)
Interlingua	14.4	**10.5** (27.1%↓)
Kyrgyz	11.1	**8.0** (28.2%↓)
Latvian	13.7	**10.2** (25.5%↓)
Kinyarwanda	45.3	**31.7** (30.0%↓)
Sakha	15.3	**11.9** (22.2%↓)
Slovenian	10.7	**8.5** (20.6%↓)
加重平均	11.1	**8.2** (26.1%↓)

TSUBAME3.0 スーパーコンピュータ [1] を用いた。

3.2　実験結果

Table 2に単一言語転移タスクにおける評価結果を示す。次元学習を行わない場合、ごく少量のデータのみを用いて End-to-End 認識システムを学習することになるため、文字誤り率 (CER) は大きな値となっている。データ量の特に少ない Chuvash と Kinyarwanda では、CER は 50%以上の値となった。他方 Esperanto については、少ないデータで学習しているにもかかわらず 5.2%と比較的低い CER となった。事前学習したモデルを初期モデルとした場合、概ねどの言語においても CER が大きく低下し言語間転移学習が有効であることが確認できた。ただし学習データ量の特に少ない Chuvash と Kinyarwanda の 2 言語に関しては、次元学習を行ったときの方が誤り率が高くなる現象が見られた。14 言語における平均では、事前学習を行わない場合の CER が 23%に対して事前学習を行った場合は 9.6%であり、58.3%の大きな相対誤り削減率が得られた。

Table 2に多言語転移タスクにおける評価結果を示す。事前学習無しのモデルは、乱数初期化状態から 14 の低資源言語データを同時に用いて学習している。複数言語のデータを使用するため、単一言語転移学習の事前学習なしの場合と比較して全ての言語において大幅に低い CER が得られた。事前学習なしの場合とありの場合を比較すると、Chuvash と Kinyarwanda を含め 14 言語すべてで事前学習ありの場合の方が低い CER となった。転移学習学習を行う際に複数言語データを同時に用いることで適応用データが特に少ない場合においても安定した言語適応化ができたためと考えられる。一方で、事前学習モデルを用いた単一言語転移学習と多言語転移学習の結果を比べると、Arabic 等後者の方がやや大きな CER となっている言語が複数存在している。これは、識対象以外の言語を同時学習しているための副作用と考えられる。

[1] https://www.t3.gsic.titech.ac.jp

4　おわりに

超多言語事前学習による低資源音声認識を検討した。事前学習により、低資源言語での認識性能が大きく向上することを確認した。今後の課題として、副作用を抑えながら多言語転移学習の効果を高めることが挙げられる。

参考文献

[1] C.C. Chiu, *et al.*, ICASSP, 2018.
[2] J. Cho, *et al.*, IEEE SLT, 2018.
[3] S. Watanabe, *et al.*, ASRU, pp. 265 - 271, 2017.
[4] R. Ardila, *et al.*, arXiv preprint arXiv:1912.06670, 2019.
[5] S. Kim, *et al.*, ICASSP, pp. 4835 - 4839, 2017.
[6] L. Dong, *et al.*, ICASSP, pp. 5884 - 5888, 2018.