4R-08

LSTM に基づいた L2 学習者の日本語発音スコアリング (Scoring for L2 Japanese learner based on LSTM)

YANG TINGCHENG† 豊橋技術科学大学†
細田侑也‡ 豊橋技術科学大学‡
若林佑幸§ 豊橋技術科学大学§
北岡教英¶ 豊橋技術科学大学¶

1 はじめに

日本語を学んでいる L2 学習者にとって，発音練習は言語習得において重要な訓練である．ただし，日本語を学んでいる外国人は 379 万人に達している一方で, 日本語教師は 8 万人未満である．もし自動的に発音レベルをスコアリングできれば，学習効率の改善に貢献できる．先行研究 [1] では，音声認識モデルで取得した特徴量を用いて発音スコアリングモデルを構築した．本研究では，音声特徴量を直接解析することで日本語発話スコアリングモデルを構築する．また，初級から上級までの日本語能力レベルの中国人話者の発話データセットを新たに作成して，提案法の有効性を検証する．

2 発話データセット

本稿では, JRF データセット [2] に加えて, 日本語能力レベルが初級な話者も収録している発話データセットを使用する．このとき，8:1:1 の比率で学習セット，検証セット，評価セットに分割する．

JRF [2] は，外国人話者の日本語発話が収録されたデータセットである．男性が 72 人，女性が 69 人の話者で構成され，母国語は 26 言語にわたる．話者の日本語能力レベルは中級から上級である．そのため，スコアの差小さく，スコアリングモデルを構築しにくい恐れがある．

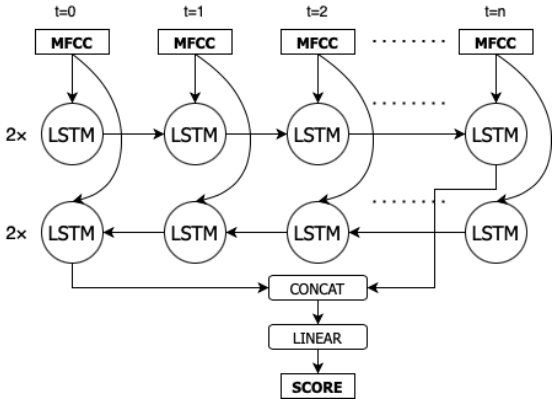

Figure 1 BiLSTM architecture for scoring

本稿では，初級から上級までの日本語能力レベルをもつ中国人学生 26 人の発話を収集したデータセット (ND) を新たに収集した．JRF から 28 文選択して合計 113 分収録した．このとき，日本人教師三名が発音の流暢さについて 1 から 5 までの整数スコアで採点した．Table 1は, 3 人の教師が評価したスコアについて全ての組み合わせで算出した Pearson's correlation coefficient (PCC) を示す．Table 1 より，教師ごとにスコアが完全に一致しておらず差があることがわかる．

3 日本語発音スコアリングモデル

3.1 音響特徴量

本研究では，Kaldi を用いて音声から抽出された四種類の音響特徴量を用いる. ここで, サンプリングレートは 16 kHz, 解析フレーム長は 25 ミリ秒, フレームシフト長は 10 ミリ秒である. 音響特徴量として，音声の周波数特徴を捉える

† YANG TINGCHENG, Toyohashi University of Technology
‡ Hosoda Yuya, Toyohashi University of Technology
§ Wakabayashi Yukoh, Toyohashi University of Technology
¶ Kitaoka Norihide, Toyohashi University of Technology

Table 1 PCC between different teachers

Score type	PCC
Teacher1-teacher2	0.55
Teacher1-teacher3	0.55
Teacher2-teacher3	0.62

Table 2 Performance of LSTM model on JRF overall score

Input feature	MSE	PCC
MFCC	0.67	0.34
FBANK	0.68	**0.48**
SPEC	0.58	0.46
PLP	0.67	0.40

Mel-Frequency Cepstrum Coefficients (MFCC), 音声信号のエネルギーを捉える Filter bank features (FBANK) を扱う. このとき，メルフィルタバンク数は 23 で設定する. また，音声信号の時間周波数成分を表す Spectrogram (SPEC), 音声信号のスペクトルを表す Perceptual Linear Predictive (PLP) も使用する．このとき，PLP では三角メルフィルタの数は 23 で設定する.

3.2 深層学習モデル

本稿では，長期的な依存関係を学習しやすい Long Short-Term Memory (LSTM) [3] を用いて発話スコアリングモデルを構築する．ここで，音声特徴量を入力したモデルの出力を線形層に与えて予測スコアを取得する．このとき，二層で LSTM を構成する．加えて，順方向と逆方向からの情報を結合して学習する Bidirectional LSTM (BiLSTM) [4] も使用する. BiLSTM では，長い文脈を必要とするタスクでも効果的に機能することが期待される. Figure 1は, BiLSTM の構成図を表す．このとき，順方向および逆方向の LSTM はそれぞれ二層で構成する.

4 実験結果

Table 2に，JRF における LSTM で構築したスコアリングモデルの評価結果を示す．この結果と，3 人の教師の平均値との間の評価指標 Mean Square Error (MSE) は最小 0.58，PCC は最大 0.48 を記録した．話者間の日本語能力レベルの差が小さいとき，LSTM モデルではその差を学習できていないことが示された.

Table 3 Performance of LSTM based models on ND overall score

	LSTM	BiLSTM
Input feature	PCC	PCC
MFCC	0.60	0.53
FBANK	0.53	**0.63**
SPEC	0.46	0.58
PLP	0.56	0.59

Table 3に，ND における評価結果を示す. LSTM と BiLSTM はどちらも，JRF よりも ND の方がパフォーマンスが優れている．PCC の最大は 0.63 に達し，教師間の PCC と同程度であった．したがって，日本語能力の差がより大きいデータセットでは，発音スコアを正常に評価できるモデルを構築できた.

5 おわりに

本研究では，音響特徴量のみを用いて日本語の発音レベルを評価するスコアリングモデルを構築した．JRF と ND の二種類のデータセットにおける実験では，BiLSTM モデルを用いた手法は L2 学習者の日本語の発音レベルを評価できた．今後は，Transformer をスコアリングモデルに導入することを検討する.

参考文献

[1] Y Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland. Bidirectional LSTM-RNN for Improving Automated Assessment of Non-Native Children's Speech. In *INTERSPEECH*, pages 1417–1421, 2017.

[2] S. Nakagawa. UME Japanese Speech Database Read by Foreign Students (UME-JRF). *Speech Resources Consortium, NII*, 2007.

[3] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[4] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681, 1997.