発話の時間変動に着目した音声認識誤り区間推定の検討*

☆舒 禹清，山田 武志，牧野 昭二（筑波大）

1 はじめに

現在の音声認識サービスには，自然発話に対して認識性能が低下するという問題がある．例えば，発話速度が極端に速い・遅い場合，発音が曖昧な場合，言いよどみやフィラーが含まれる場合などである．よって，サービス品質の保証という観点から，音声認識サービスの提供中に認識性能をモニタリングする手法が必要である．

従来，認識結果に付与された信頼度を用いて，各単語を正解または 3 種類の不正解（置換，挿入，欠落）に確率的に分類して認識性能を推定する手法が提案された [1]．しかし，この手法は音声認識を実行する必要があるため，計算コストが高いという問題がある．また，入力発話全体から抽出した統計的な音響特徴量を用いて認識性能を推定する手法が提案された [2]．しかし，発話が短い場合に推定精度が低下するという問題がある．一方，我々は BLSTM（bidirectional long short-term memory）と変調スペクトルを用いて認識誤り区間を推定する手法を提案した [3]．この手法を用いれば，認識誤り区間のフレーム数と発話全体のフレーム数の比率として認識性能を算出できる．また，認識性能のみではなく，認識誤り区間がフレーム単位で分かるので，話者適応の利用や音声対話による認識誤り訂正などが可能となる．

本稿では，この手法の推定精度をさらに高めるために，CNN（convolutional neural network）と RNN（recurrent neural network）を組み合わせた CRNN（convolutional recurrent neural network）[4] を導入する．変調スペクトルはフレーム毎に 2 次元の特徴マップとして表されるので，CNN を用いて分析するのが適していると考えられる．認識誤り区間推定の実験を行うことにより，その有効性を検証する．

2 提案手法

2.1 概要

Fig. 1 に示すように，提案手法ではまず入力発話から MFB（log-mel filterbank energy）をベースとする変調スペクトルを算出する．そして，変調スペクトルのフレーム時系列を CRNN に入力し，フレーム毎に認識誤りか認識正解かを推定する．以下では，CRNN と変調スペクトルについて述べる．

2.2 CRNN

CRNN は CNN と RNN を組み合わせたものであり，画像のような 2 次元の特徴マップに対する分析と時系列データの分析の双方に適している．提案手法では，RNN として BGRU（bidirectional gated recurrent unit）[5] を用いている．

提案手法における CRNN の学習について述べる．まず，学習用の発話を大語彙連続音声認識エンジン Julius[6] に入力し，認識結果のテキストを得る．次に，認識結果のテキストとその発話の書き起こしテキストを音素単位で照合し，フレーム毎に認識誤りラベル，認識正解ラベルを付ける．そして，これを教師ラベルとして CRNN を学習する．

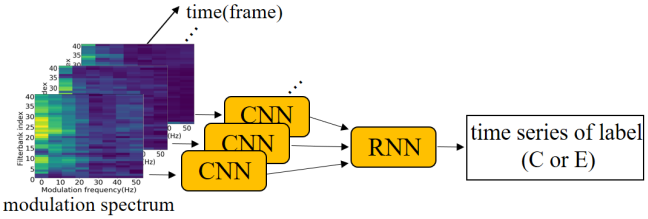

Fig. 1 Proposed method.

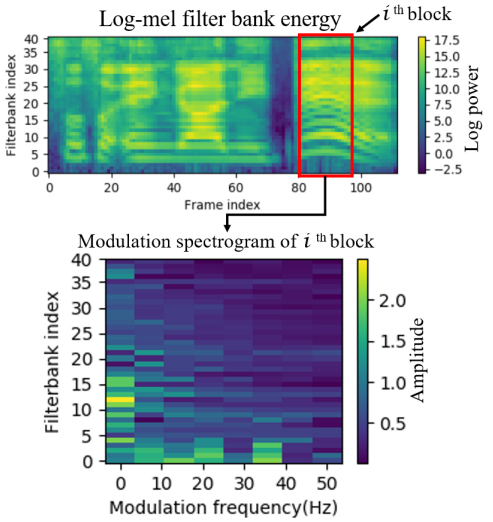

Fig. 2 Modulation spectrum.

2.3 変調スペクトル

変調スペクトルは，対象とする特徴量の時間軌跡のスペクトル表現として定義され，発話速度と強い関係がある [7]．認識誤りの原因は発話速度とその変動に関係することが多い．例えば，極端に早口の場合は発音に曖昧さが生じたり，言いよどみやフィラーにおいては音素遷移に不規則性が現れたりする．よって，変調スペクトルは認識誤り区間の推定に適していると考えられる．

Fig. 2 に変調スペクトルの例を示す．まず，入力発話から MFB を抽出する．次に，連続する t 個のフレームからなるブロック（赤枠部分）において，各フィルタバンクインデックスの時系列信号に対して変調スペクトルを求める．これを全てのブロックに対して行うことにより，変調スペクトルの時系列を得る．

3 提案手法の有効性の検証

3.1 実験条件

音声データの詳細を Table 1 に示す．本実験では，宇都宮大学パラ言語情報研究向け音声対話データベース（UUDB）[8]，重点領域研究「音声対話」対話音声コーパス（PASD）[9]，RWCP 音声対話データベース 96 年版 [10] から男性と女性の各 25 名の計 5415 個の音声データを用いる．ここで，各音声データから認識誤り区間と認識正解区間のフレーム数の比率が 1:1

*Speech recognition error detection based on the time variation of utterances. by Yuqing SHU, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

Table 1　Speech data

Database	UUDB	2 male 12 female
	PASD	8 male
	RWCP-SP96	15 male 13 female
# of utterances	5415	
Training : test	4 : 1	
Sampling rate	16 kHz	
Quantization bit	16 bits	

Table 2　CRNN specifications

# of convolutional layers	3
Filter size	3×3
# of pooling layers	3
Pooling window size	1×2
# of BGRU layers	2
# of BGRU units	200
Loss function	softmax cross entropy
Batch size	10
Learning rate	0.00001，...，0.00010
Optimizer	Adam
Epoch	1，2，...，100

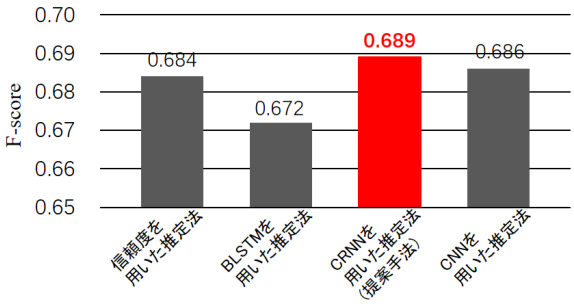

Fig. 3　F-score for each method.

になるように不連続フレームが生じないように切り出しを行う．

本実験では，単語信頼度を用いた手法（Julius が出力した単語信頼度と閾値を比較して単語単位で推定する），BLSTM を用いた手法 [3]，CRNN を用いた手法（提案手法），CNN を用いた手法（提案手法から RNN 部を取り除いてフレーム単位で推定する）を比較する．

Table 2 と Table 3 に CRNN と BLSTM の条件を示す．CNN の畳み込み層の数は 3，フィルタサイズは 3×3，プーリングサイズは 1×1 であり，また BGRU と BLSTM の隠れ層の数は 2，各層のユニット数は 200 である．各推定器の最適化手法は Adam であり，学習率は 0.00001〜0.00010 の範囲で調整する．学習は 100 エポックまで行い，最も推定性能が高い結果を採用する．次に，Table 4 と Table 5 に MFB と変調スペクトルの条件を示す．変調スペクトルを計算する際のブロック長は 4 通り，FFT サイズは 4 通りとし，最も推定性能が高い結果を採用する．

3.2　実験結果

Fig. 3 に各推定手法の F 値を示す．提案手法の F 値は，従来手法である BLSTM を用いた手法よりも高いことが分かる．一方，CNN を用いた手法との差は僅かであることから，提案手法においては RNN 部よりも CNN 部の方が性能改善に寄与していると考え

Table 3　BLSTM specifications

# of layers	2
# of units	200
Loss function	softmax cross entropy
Batch size	10
Learning rate	0.00001，...，0.00010
Optimizer	Adam
Epoch	1，2，...，100

Table 4　MFB specifications

Mel filters	40
Frame length	25 ms
Frame shift length	10 ms

Table 5　Modulation spectrum specifications

Base feature	40-dimensional MFB			
Block shift length(ms)	10			
Block length (ms)	80	160	320	640
(frames)	8	16	32	64
FFT size	8	16	32	64
Dimension	4×40	8×40	16×40	32×40

られる．また，提案手法は音響特徴量のみから推定しているにもかかわらず，単語信頼度を用いた手法を僅かながら上回っていることが分かる．

4　おわりに

本稿では，自然発話音声の時間変動に着目し，CRNN と変調スペクトルを用いた音声認識誤り区間推定法を提案した．認識誤り区間推定の実験を行うことにより，従来手法よりも高い推定性能が得られることを確認した．

謝辞　本研究は JSPS 科研費 17K00224 の助成を受けた．

参考文献

[1] A. Ogawa *et al.*, “Error type classication and word accuracy estimation using alignment features from word confusion network, ” Proc. ICASSP 2012，pp. 4925–4928, 2012.
[2] L. Guo *et al.*, “Performance estimation of spontaneous speech recognition using non-reference acoustic features, ” Proc. APSIPA 2016, Paper ID 239, pp. 1–4, 2016.
[3] 舒禹清ら, “BLSTM を用いた音声認識誤り区間推定の検討, ” 日本音響学会講演論文集, pp. 921–924, Sep. 2019.
[4] E. Cakır *et al.*, “Convolutional recurrent neural networks for polyphonic sound event detection,” IEEE/ACM Trans. ASLP, Vol. 25, issue 6, pp. 1291–1303, 2017.
[5] K. Cho *et al.*, “Learning Phrase Representations using RNN encoder-decoder for statistical machine translation, ” arXiv: 1406.1078, 2014.
[6] https://julius.osdn.jp/
[7] Z. Zhu *et al.*, “Contributions of the temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech,” Acoustical Science and Technology, Vol. 39, No. 3, pp. 234–242, 2018.
[8] http://research.nii.ac.jp/src/UUDB.html.
[9] http://research.nii.ac.jp/src/PASD.html.
[10] http://research.nii.ac.jp/src/rwcp-sp96.html.