BLSTM を用い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（筑波大学）

1 はじめに

近年，音声認識の性能は様々な先進技術の導入により著しく改善し，それと共に音声認識サービスは広く一般に普及した．しかし，現在の音声認識技術では自然発話音声に対して認識性能が低下するという問題がある．例えば，極端に速く・遅く話してしまったり，発音が曖昧だったり，言いよどみやフィラーが含まれたりすると正しく認識することが難しい．よって，サービス品質（認識性能）の保証という観点から，音声認識サービスを提供中に自然発話音声に対する認識性能をモニタリングする手法が必要である．

従来，この問題を解決するために，認識結果に対し付与された信頼度を用いて，確率的に誤りタイプ分類を行って認識性能を推定する手法が提案された [1]．この手法では，認識性能は認識結果文中の各単語を正解または 3 種類の不正解（不正解，挿入，欠落）に確率的に分類して認識率を推定する．しかし，この手法は音声認識を実際に行う必要があるため，計算負荷が高いという問題がある．

一方，入力発話から抽出した音響特徴量のみを用いて認識性能を推定する手法が提案された [2]．これは，入力発話全体から各種の統計的な特徴量を抽出し，SVR（support vector regression）[3] を用いて認識性能を推定する．しかし，発話が短い場合に有効な特徴量を抽出することが難しいので，推定精度が低下するという問題点がある．そこで本稿では，Fig. 1 に示すように，時間フレームを単位とする音響特徴量を用いて認識誤り区間を推定するというアプローチを検討する．認識性能は，発話全体のフレーム数と認識誤り区間のフレーム数の比に基づいて算出することができる．また，認識性能のみではなく，認識誤り区間が分かるので，話者適応の適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，これを実現するために，変調スペクトルと BLSTM（bidirectional long short-term memory）[4] を用いた認識誤り区間推定法を提案し，その有効性を検証する．ここで，変調スペクトルは特徴量の時間軌跡のスペクトル表現として定義される．認識誤りの原因の多くは発話速度とその変動に関係していることから，認識誤り区間の推定に適していると考えられる．また，BLSTM は時系列信号の処理に適しており，例えば音響イベント検出における有効性が示されている [5]．

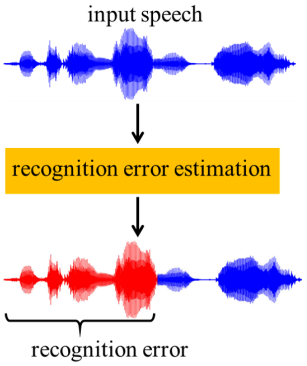

Fig. 1 Approach of estimating recognition error period.

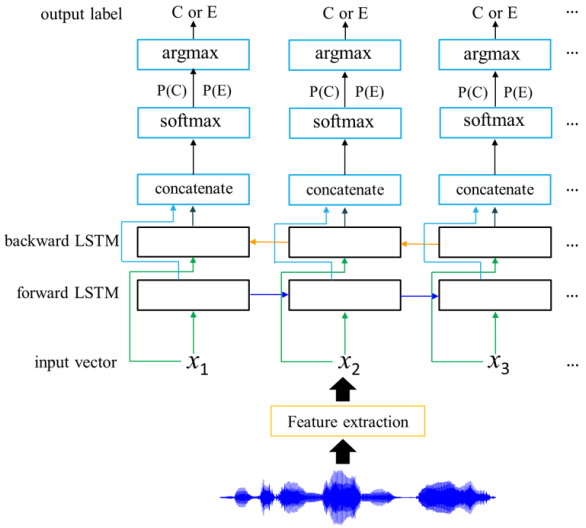

Fig. 2 Process flow of the proposed method

2 提案手法

2.1 提案手法の概要

Fig. 2 に提案手法の処理の流れを示す．提案手法では，まず入力発話から変調スペクトルを求める．ここで，変調スペクトルはメルフィルタバンク特徴量をベースとする．そして，変調スペクトルを BLSTM

* A study on speech recognition error detection using BLSTM by Yuqing SHU, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba).

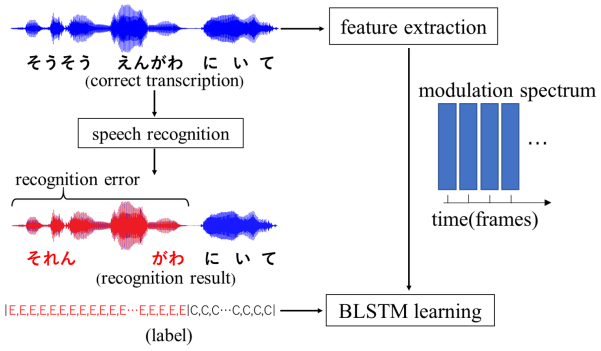

Fig. 3 Learning of BLSTM

に入力し，フレーム毎に認識誤り区間（E）か認識正解区間（C）のラベルを出力する．以下では，BLSTM と変調スペクトルについて詳しく述べる．

2.2 BLSTM

時系列データを扱えるニューラルネットワークとして RNN（recurrent neural network）[6] がある．RNN は再帰型の構造を持つため，過去の隠れ層の情報を保持し，それを併用して現在の処理を行うことができる．RNN には長期依存関係の取り扱いに問題があるが，これを解決するためにメモリセルなどの機能を追加した LSTM が提案された [7]．BLSTM は，LSTM の前向きの隠れ層の情報に加えて，後ろ向きの隠れ層の情報を追加したものである．BLSTM は時系列信号の処理に適しており，例えば音響イベント検出における有効性が示されている [5]．

Fig. 3 に提案手法における BLSTM の学習の概要を示す．まず，学習用の発話を大語彙連続音声認識エンジン Julius [8] 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 BLSTM を学習する．

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る．変調スペクトルは発話速度と強い関係があり，例えば感情識別のタスクにおいて有効であることが報告されている [9]．同様に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，言い淀み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．

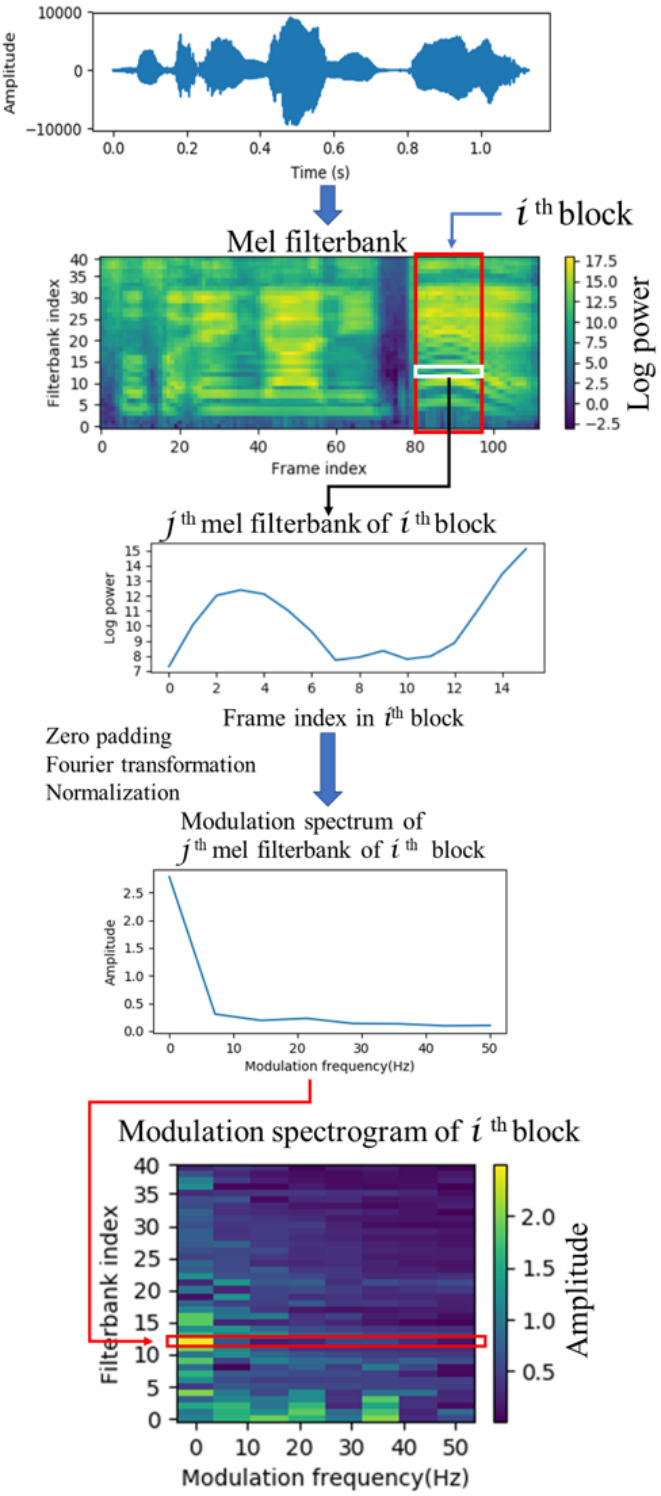

Fig. 4 Calculation of modulation spectrum

Fig. 4 に提案手法における変調スペクトルの求め方を示す．まず，入力発話からメルフィルタバンク特徴量を抽出する．次に，連続する t 個のフレームからなるブロック（Fig. 4 の赤枠）において，各フィルタバンクインデックスにおける時系列信号（Fig. 4 の白枠）に対し，ゼロパディング，フーリエ変換，正規化

Table 1　Speech data

Database	UUDB	2 male 10 female
	PASD	8 male
	RWCP-SP96	10 male 10 female
# of utterances	4724	
Training : test	4 : 1	
Sampling rate	16 kHz	
Quantization bit	16 bits	

Table 2　BLSTM specifications

Number of layers	2
Number of units	220
Loss function	softmax cross entropy
Batch size	10
Learning rate	0.00001, ..., 0.00010
Optimizer	Adam
Epoch	1, 2, ..., 100

を行うことにより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，変調スペクトルグラムの時系列を得る．

3　提案手法の有効性の検証

3.1　実験条件

Table 1 に実験に用いた音声データの詳細を示す．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 [10]），重点領域研究「音声対話」対話音声コーパス（PASD [11]），RWCP 音声対話データベース 96 年版（RWCP-SP96 [12]）を用いる．これらのデータベースから男性 20 名，女性 20 名の 4724 個の認識誤りを含む音声データを選択し，そのうちの 4/5 を学習データセット，1/5 をテストデータセットとする．なお，各音声データに対して認識誤り区間と認識正解区間のフレーム数の比が 1 : 1 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．

Table 2 に BLSTM の条件を示す．予備実験の結果に基づき，隠れ層の数は 2，各層のユニット数は 220 とする．最適化関数は 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用いる．次に，Table 3 と Table 4 にメルフィルタバンク特徴量と変調スペクトルの条件を示す．変調スペクトルは 40 次元のメルフィルタバンク特徴量をベースとして求める．ブロック長は 4 通り，FFT サイズは 2 通りとする．ブロック長はどのくらいの時間長の時間変動

Table 3　Mel filter bank specifications

Sampling rate	16 kHz
FFT sample points	512
Mel filters	40
Frame length	25 ms
Frame shift length	10 ms

Table 4　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	160	320	640	1280

を分析するかを表し，FFT サイズは変調周波数の分解能を決定する．これらのパラメータが推定性能に及ぼす影響を検証する．本実験では，推定性能を測る尺度として以下に示す F 値を用いる．

$$F\text{ 値} = 2 \times \frac{\text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

$$\text{再現率} = \frac{\text{正しく検出した誤認識フレームの数}}{\text{真の誤認識フレームの数}}$$

$$\text{適合率} = \frac{\text{正しく検出した誤認識フレームの数}}{\text{検出した誤認識フレームの数}}$$

本実験では，音声認識の結果から得られる単語信頼度と閾値を比較することにより，単語単位で認識誤りか認識正解かを推定する手法を用意し，提案手法の推定性能と比較する．ここで，本実験では Julius における事後確率に基づく単語信頼度を用いる．

3.2　実験結果

まず，Fig. 5 に比較手法において閾値を変化させたときの F 値を示す．図から，信頼度の閾値を 0.85 に設定したときに最も高い F 値が得られることが分かる．しかし，その F 値は 0.686 であり，音声認識を実際に行ったことにより得られる情報を用いても認識誤りの推定は難しいことが分かる．次に，Table 5 に提案手法においてブロック長と FFT サイズを変化させたときの F 値を示す．図から，ブロック長を 8 フレーム，FFT サイズを 64 ポイントとしたときに最も高い F 値が得られることが分かる．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，

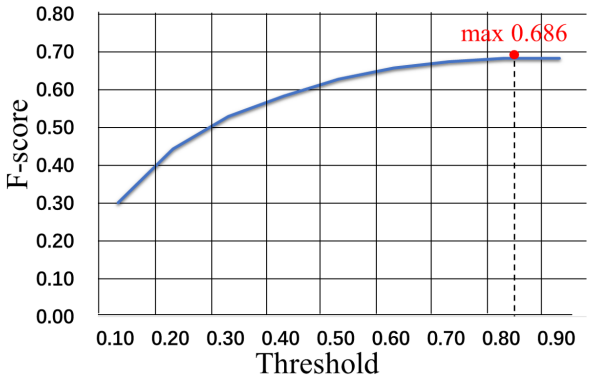

Fig. 5　Result of the comparison method

Table 5　Result of proposed method

		Block length			
		8	16	32	64
FFT size	8	0.642			
	16	0.650	0.637		
	32	0.660	0.647	0.641	
	64	**0.669**	0.651	0.662	0.638

比較手法と同程度の F 値が得られていることが確認できる．

4　おわりに

本稿では，自然発話音声を対象とする認識性能推定を実現するために，変調スペクトルと BLSTM を用いた音声認識誤り区間推定法を提案した．実験により，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，音声認識の結果から得られる単語信頼度に基づく推定手法と同程度の F 値が得られていることを確認した．

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, H. Takaaki, N. Atsushi, “Error type classication and word accuracy estimation using alignment features from word confusion network,” Proc. ICASSP 2012 pp. 4925–4928, 2012.

[2] L. Guo, T. Yamada, S. Makino, “Performance estimation of spontaneous speech recognition using non-reference acoustic features,” Proc. APSIPA 2016, Paper ID 239, pp. 1-4, 2016.

[3] V.N. Vapnik, “Statistical Learning Theory,” A Wiley-Interscience Publication, 1998.

[4] A. Graves, J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” Neural Networks, Vol. 18, pp. 602-610, 2005.

[5] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Roux, K. Takeda, “Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection,” DCASE2016 Challenge, Tech. Rep., 2016.

[6] C. L. Giles, G. M. Kuhn, and R. J. Williams, “Dynamic recurrent neural networks: theory and applications,” IEEE Transactions on Neural Networks, Vol. 5, No. 2, pp. 153–156, 1994.

[7] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” Neural Computation, Vol. 8, No. 9, pp. 1735–1780, 1997.

[8] 河原達也, 李晃伸, “連続音声認識ソフトウェア Julius,” 人工知能学会誌, Vol. 20, No. 1, pp. 41–49, 2005.

[9] Z. Zhu, R. Miyauchi, Y. Araki, M. Unoki, “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.

[10] 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, http://research.nii.ac.jp/src/UUDB.html.

[11] 重点領域研究「音声対話」対話音声コーパス, http://research.nii.ac.jp/src/PASD.html.

[12] RWCP 音声対話データベース 96 年版, http://research.nii.ac.jp/src/rwcp-sp96.html.