BLSTM と変調スペクトルを用いた発話特徴識別の検討 *

☆サントソ ジェニファー, 山田武志, 牧野昭二 (筑波大)

1 はじめに

近年，携帯機器の普及に伴い，音声認識システムが広く一般に使われるようになってきている．現在の音声認識システムは，音声アシスタントのように短い音声コマンドを認識することから，講義音声の書き起こしのように長時間の発話を認識することまで，様々なニーズに応えている．このように音声認識システムの性能は向上しているが，ユーザからの発話を正しく認識できない場合が依然としてある．

従来，誤認識の可能性が高い単語を検出し，音声対話により聞き直す手法 [1] が提案された．しかし，ユーザは何故誤認識されたのか，すなわちどのような種類の発話，また発話のどの部分が誤認識を引き起こしたのかが分からない．それにもかかわらず，システムはユーザに誤認識原因についてのフィードバックを与えずに単に発話を繰り返すように要求する．有益なフィードバックの欠如は，音声認識システムのユーザビリティを低下させ，その結果，ユーザは不親切であるとみなして使用することを諦めてしまう．従って，音声認識システムのユーザビリティを向上させるためには，誤認識原因を特定し，ユーザが理解しやすいように通知することによってユーザに次の発話を誤認識されないように発話させることが極めて重要であると言える．

音声認識における誤認識には主に 3 つの原因がある． 1 つ目は環境条件である．例えば，雑音，反射，残響などの外部からの干渉である．2 つ目は辞書に登録されていない未知語などのシステム要因である． 3 つ目は発音，発話速度，フィラー，言い淀みなどの発話特徴である．

環境条件による誤認識に関する対策として，雑音の多い環境での音声認識に必要な発話音量をユーザに知らせる手法が提案された [2]．これは雑音を含む入力信号から適切な発話音量を予測し，それをユーザに分かりやすく通知することにより，潜在的な誤認識を減らすことに成功した．

また，発話特徴による誤認識に関する対策としては，入力音声から発話特徴を識別し，それをユーザにフィードバックすることが考えられる．発音については，computer-assisted language learning（CALL）

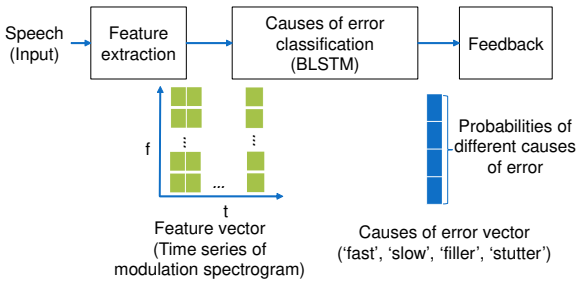

Fig. 1 Process flow of the proposed method flow

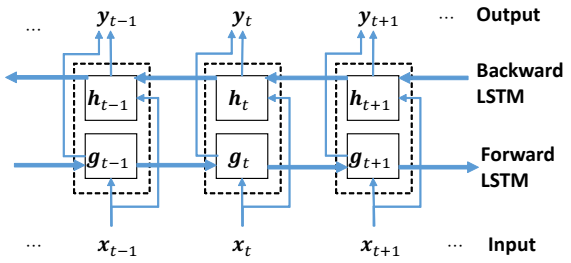

Fig. 2 BLSTM network architecture

システムにおいて発音誤りを検出し，ユーザに改善方法をフィードバックする方法 [3] がある．しかし，そのフィードバックはユーザにとって複雑で難しく，それゆえ発音をすぐに改善することは難しい．

一方，発話速度，フィラー，言い淀みなどの発話特徴については，ユーザにフィードバックしやすく，またユーザにとって比較的容易に改善できると考えられる．そこで我々は，‘fast’，‘slow’，‘filler’，‘stutter’という誤認識原因に注目し，音響特徴量として変調スペクトル（modulation spectrum; MS），識別器として bidirectional long short-term memory（BLSTM）を用いた発話特徴識別手法を提案した [4]．本稿では，この手法の識別精度を更に向上させるために，MS を求める際の分析時間長や変調周波数分解能が識別精度に及ぼす影響を調査する．

2 提案手法 [4]

2.1 提案手法の概要

提案手法の処理の流れを Fig. 1 に示す．まず，音声を入力し，入力音声から特徴量ベクトルの時系列を得る．次に，特徴量ベクトルを識別器に入力し，誤認

* A study on classification of utterance characteristics using BLSTM and modulation spectrum.
by Jennifer SANTOSO, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

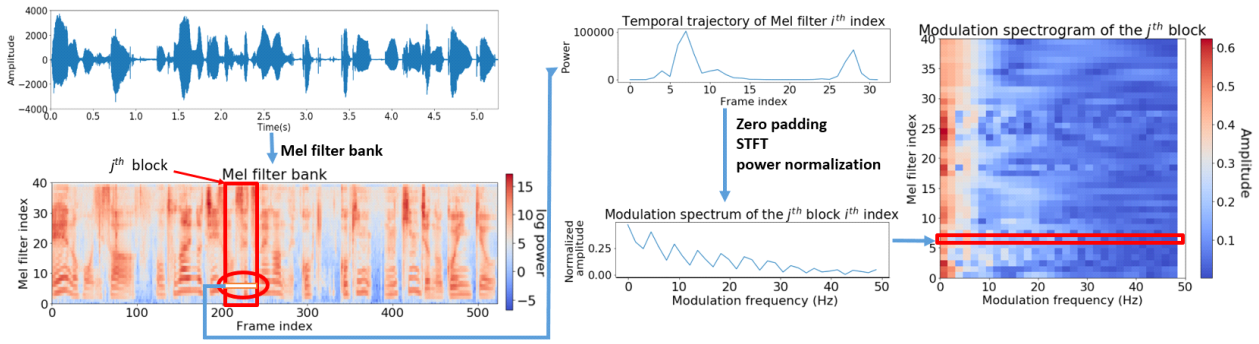

Fig. 3　Process flow to obtain modulation spectrum

識原因となり得る各発話特徴の存在確率を出力する．この出力に基づいてフィードバックを行う．

提案手法では，BLSTM [5] を識別器として使用する．BLSTM のネットワーク構造を Fig. 2 に示す．BLSTM は recurrent neural network [6] の拡張であり，時系列情報を格納しながら前後に移動する 2 つの LSTM [7] ネットワークから成る．BLSTM は音関連のタスクを処理するのに適し，例えば音響イベント検出 [8] に適用されている．

2.2　音響特徴量抽出

従来，音声認識システムでは，メルケプストラム係数やメルフィルタバンク（mel filter bank; MFB）などのスペクトログラムベースの特徴量が一般に使用されている．スペクトログラムベースの特徴量には音素情報が含まれているため，音素を識別するのに有効である．しかし，本稿で対象とする発話特徴を識別するのに必要とされる情報は，発話速度に関連する情報である．従って，音響特徴量は時系列信号の時間変化に関する情報を表現できることが重要である．

提案手法では，このような音響特徴量として MS [9] を使用する．MS 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．例えば音節の不規則な時間遷移を捕えることができ，また MS は音声感情認識タスク [10] において有効であることが示されている．

Fig. 3 に MS を求める処理の流れを示す．まず，音声信号に短時間フーリエ変換（short time Fourier transform; STFT）を適用してパワースペクトログラムを計算する．次に，このパワースペクトログラムをメルフィルタバンクに適用して MFB を得る．そして，連続した t 個のフレームから成る各ブロックにおいて，各メルフィルタインデックスの時系列信号に対して，Q ポイントの STFT を適用し MS を得る．ここで，変調周波数の分解能を上げるため，$Q(\geq t)$ ポイントまでゼロパディングを行う．最後に，各メルフィルタインデックスの変調スペクトルを正規化して組み合わせることにより変調スペクトログラムを形成する．

Table 1　Speech data specifications

Database	PASD	UUDB
Speakers	8 male 2 female	2 male 8 female
Dataset ID	kyo0121, kyo0221, kyo0321, osa0910, osa0918, uec0001, uec0002, uec0003, uec0004	C001, C002, C021, C022, C023, C024, C031, C032, C033, C051, C052, C053, C061, C062, C063, C064
Utterance count	1400 utterances	
Sampling rate	16 kHz	16 kHz
Quantization	16 bits	16 bits
Length	1–10 s	1–10 s

変調スペクトログラムの重要なパラメータとして，ブロック長と FFT ポイント数がある．ここで，ブロック長はどのくらいの時間長の時間変動を分析するかを決定し，FFT ポイント数は変調周波数分解能を設定する．

3　実験

本章では，発話特徴識別実験を行い，提案手法の有効性を検証する．その際，MS のブロック長と FFT ポイント数が識別精度に及ぼす影響を調べる．また，MFB と MS の識別精度を比較する．

3.1　音声データ

Table 1 に，本実験で使用した音声データの仕様を示す．本実験では，日本語の対話音声データベースである重点領域研究「音声対話」 対話音声コーパス

Table 2　Mel filter bank specifications

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

Table 3　Modulation spectrum specifications

Base feature	40-dimensional MFB				
Block length (ms)	320	80	160	320	640
(frames)	32	8	16	32	64
FFT points	32	64			
Dimension	640	1280			

(PASD)[11] と宇都宮大学 パラ言語情報研究向け音声対話データベース（UUDB）[12] を用いる．話者は全部で男性 10 名，女性 10 名であり，これらの話者の音声データの中から Julius [13] によって誤認識された 1,400 個の音声データを用いる．ここで，音声データの長さは 1〜10 秒である．誤認識された音声データには発話特徴のラベルを付ける．ここで，‘filler’ と ‘stutter’ については，各音声データベースに付属の書き起こしテキストに従ってラベルを付ける．一方，‘fast’ と ‘slow’ については実際に聴取し，手動でラベルを付ける．

3.2　特徴量抽出

Table 2 と Table 3 はそれぞれ MFB と MS の音響特徴量抽出条件を示す．MFB については，サンプリング周波数 16 kHz，FFT ポイント数 512，フレーム長 25 ms，フレームシフト長 10 ms, メルフィルタバンク数 40 である．MS については，上記の 40 次元 MFB をベース特徴量として用いる．ブロック長は 4 通り（8, 16, 32, 64 フレーム），FFT ポイント数は 2 通り（32, 64 ポイント）に変化させ，識別精度を比較する．その結果，1 ブロックの MS の次元は 640(=40×16) または 1280(= 40×32) になる．

3.3　識別器

識別器の条件を Table 4 に示す．識別器については発話特徴のクラスごとに個別に学習する．隠れ層の数は 1 または 2，各層のユニット数は 64 または 128 である．また，ドロップアウトは 0.0 から 0.2 に設定する．最適化関数は Adam[14] であり，学習率は 0.001, 0.0005, 0.0001, 0.00005 である．各識別器の学習は 50 エポックまで実施し，最も認識精度が高い結果を選択

Table 4　Classifier specifications

Classifier	BLSTM
Layers and units	640–(128–128)–256–2 1280–(128–128)–256–2
Optimizer	Adam
Learning rate	0.001–0.00005
Dropout	0.0–0.2
Loss function	Softmax cross entropy
Epoch	1, 2, ..., 50
Cross-validation	5-fold speaker open test
training	8 male+8 female
test	2 male+2 female

Table 5　F-score comparison of different modulation spectrum block length

Classifier	BLSTM			
Block length	8	16	32	64
FFT points	64	64	64	64
fast	**0.596**	0.568	0.594	0.584
slow	0.590	0.585	0.580	**0.595**
filler	0.649	0.667	0.669	**0.675**
stutter	0.627	0.634	**0.666**	0.618

する．各クラスについて 5 フォールドの交差検定を行い，以下の式で示す F 値を平均して評価尺度として利用する．

$$F\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{2}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{3}$$

3.4　実験結果と考察

まず，提案手法において変調スペクトルのブロック長を変化させた時の F 値を Table 5 に示す．各発話特徴によって変調スペクトルの適切なブロック長が異なることが分かる．‘fast’ については，8 フレームのブロック長が最も有効である．一方，‘slow’，‘filler’，‘stutter’ については，特徴量を捉えるためにある程度の時間長が必要であるために，32 フレーム，あるいは 64 フレームのブロック長が適していると考えられる．

次に，提案手法において FFT サンプル数を変化させた時の F 値を Table 6 に示す．FFT サンプル数を 64 としたときに最も F 値が高いことが分かる．これは，変調周波数の分解能を十分高くすることが重要であることを示している．

Table 6　F-score comparison of FFT points in the proposed method

Classifier	BLSTM	
Block length	32	32
FFT points	32	64
fast	0.577	**0.594**
slow	0.576	**0.580**
filler	0.641	**0.669**
stutter	0.650	**0.666**

Table 7　F-score comparison of Mel filter bank and modulation spectrum

Classifier	BLSTM	
Feature	MFB	MS
Block length		8, 16, 32, 64
FFT points	n/a	64
fast	0.572	**0.596**
slow	0.576	**0.595**
filler	0.601	**0.675**
stutter	0.601	**0.666**

最後に，音響特徴量として MFB と MS を用いたときの F 値を Table 7 に示す．Table 7 から，MS の方が MFB よりも F 値が 0.019–0.074 高く，MS の有効性を確認できる．

4　おわりに

本稿では, BLSTM と変調スペクトルを用いた発話特徴識別手法を提案した．実験を行った結果，各発話特徴によって変調スペクトルの適切なブロック長が異なること，および変調スペクトルの周波数分解能が高い方が識別精度が高くなることを確認した．また，変調スペクトルはメルフィルタバンクより有効であることを示した．

謝辞　本研究は JSPS 科研費 17K00224 の助成を受けた．

参考文献

[1] E. Pincus, S. Stoyanchev, J. Hirschberg, “Exploring features for localized detection of speech recognition errors,” Proc. SIGDIAL 2013, pp. 132–136, 2013.

[2] T. Goto, T. Yamada, S. Makino, “Novel speech recognition interface based on notification of utterance volume required in changing noisy environment,” Proc. NCSP’18, pp. 192–195, 2018.

[3] R. Ai “Automatic pronunciation error detection, feedback generation for CALL applications,” Proc. LCT 2015, pp. 175–186, 2015.

[4] J. Santoso, T. Yamada, S. Makino, “Categorizing error causes related to utterance characteristics in speech recognition,” Proc. NCSP’19, pp. 514–517, 2019.

[5] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, “Bidirectional LSTM networks for improved phoneme classification and recognition,” Proc. ICANN 2005, Vol. 2, pp. 799–804, 2005.

[6] C. L. Giles, G. M. Kuhn, 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] 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.

[9] H. Hermansky, “Should recognizers have ears?” Speech Communication, Vol. 25, No. 1–3, pp. 3-27, 1998.

[10] 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.

[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, Accessed: 2019-06-12.

[12] 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, Accessed: 2019-06-12.

[13] 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.

[14] D. P. Kingma, J. Ba, “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980, 2014.