音声認識における誤認識原因通知のための印象評定値推定の検討*

☆後藤孝宏，山田武志，牧野昭二（筑波大)

1 はじめに

近年，音声認識技術の向上やスマートフォンの普及により，音声認識を用いたサービスはより一般的なものとなり，様々なシーンで活用されている．しかし音声認識サービスの発展のためには，解決しなければならない課題が存在する．その課題の一つとして，周囲の雑音や話者の発話様式によって起こる認識性能の低下が挙げられる．現在の音声認識サービスでは，誤認識の可能性が高いと判断された場合 [1] に，「もう一度話してください」などと通知し，再発話を促すことが多く，誤認識の原因を通知することはほとんどない．そのため，ユーザは誤認識に対してどのように対処すればよいか分らず同じ誤認識を繰り返してしまい，音声認識を利用する意欲を失ってしまうケースもある．

この問題を解決する方法の一つは，音声認識の認識性能を十分高いレベルに引き上げることである．一方，我々はこの問題をユーザインタフェース上の問題と捉え，誤認識の原因をユーザに対して分かりやすく通知する手法を検討している．通知を受けたユーザは，誤認識の原因を知り対策を講じることができるので，何度も同じ誤認識を繰り返すことはなくなり，音声認識サービスのユーザビリティ向上につながる．雑音に起因する誤認識については，現在の雑音環境において得られるであろう認識率を予測することで適切な発話音量を判断し，ユーザに通知する手法を提案した [2][3]．一方で，話者の発話様式に起因する誤認識についてユーザに原因を通知する手法はあまり検討されていない．

そこで本稿では，誤認識の原因となり得る発話特徴を，ユーザに伝わりやすい形で定量化し，音声からその発話特徴を推定する手法を提案する．ここで誤認識の原因となり得る発話特徴として，CSJ（日本語話し言葉コーパス）[4] の印象評定項目 [5] に着目し，認識性能との相関が高い印象評定項目を誤認識の原因となり得る発話特徴とする．CSJ のコアデータを用いて実験を行い推定性能を検証する．

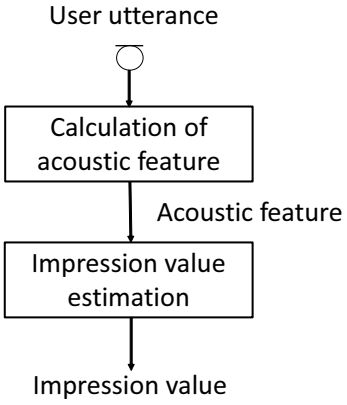

Fig. 1: Process flow of the proposed method.

2 提案手法

提案手法の処理の流れを Fig. 1 に示す．まずユーザの発話から音響特徴量を抽出し，次に誤認識の原因となり得る発話特徴を推定する．

推定対象である誤認識の原因となり得る発話特徴は，CSJ のコアデータにつけられている印象評定項目である．人が発話に対して持つ印象の評定値と認識率との相関を分析することで，誤認識の原因になり得る印象評定項目を選定する．これについては 3 章で述べる．

また音響特徴量は，時間フレームごとに求めた音響特徴量の統計量，および時間フレームごとに求めた音響特徴量の時系列の 2 種類を検討する．推定モデルは，前者については SVR(Support Vector Regression)[6] と MLP（Multi Layer Perceptron）[7]，後者については BLSTM（Bidirectional Long Short Term Memory）[8] の適用を検討する．これについては 4 章で述べる．

3 印象評定項目の選定

3.1 実験の目的と条件

誤認識の原因となり得る発話特徴として，CSJ の印象評定項目に着目する．人が発話に対して付けた印象評定値と認識率との相関を分析することで，誤

*Impression rating score estimation for error cause notification in speech recognition, by Takahiro Goto, Takeshi Yamada, Shoji Makino (University of Tsukuba).

講演音声評定項目

好悪(a1-4)	上手さ(a5-8)	速さ感(a9-12)	活動性(a13-16)	スタイル(a17-20)
好きな 嫌いな	流暢な たどたどしい	速い 遅い	声の大きい 声の小さい	礼儀正しい 無礼な
心地よい 不快な	話し慣れた 話し慣れていない	スピード感のある ゆったりした	力強い 弱々しい	まじめな ふまじめな
感じのいい 感じの悪い	なめらかな しどろもどろな	せわしげな のんきな	元気のある 元気のない	丁寧な ぞんざいな
親しみやすい 親しみにくい	上手い 下手な	落ち着きのない 落ち着きのある	積極的な 消極的な	上品な 下品な

単項項目(a21-26)

あらたまった くだけた	きまじめな 奔放な	きちんとした くつろいだ	甘えた そっけない	その場で考えて話している 原稿を読み上げている	聞き取りやすい 聞き取りにくい

Fig. 2: Impression item.

認識の原因になり得る印象評定項目を選定する．

日本語話し言葉コーパスとは，日本語の自発話音声を集め各音声に種々の情報を付与したものである．その中でもコアデータと呼ばれるデータには各音声がどのような印象を与えるかを複数人の評定者により評価した印象評定データが含まれる．今回実験に用いた印象評定項目の一覧を Fig. 2 に示す．これらの印象評定項目は，心理尺度構成法に基づいて作成されている．それぞれの音声について異なる 10 人の評定者が聞き，7 段階で評定値を付ける．7 に近いほど上の評定語によくあてはまり，1 に近いほど下の評定語によくあてはまる．この印象評定値がつけられた音声を実際に認識し，評定値と認識率の相関を分析することにより誤認識の原因となる発話特徴を選定する．

認識実験の条件について説明する．CSJ のコアデータのうち学会講演，模擬講演の音声計 171 個を使用する．それぞれの音声から印象評定値の付けられている区間のうち前半の 1 分間を切り出し，認識を行う．認識器には大語彙連続音声認識エンジン Julius4.4，音響モデルには ASJ-JNAS コーパスと CSJ 模擬講演音声で学習した 3 状態 DNNHMM，言語モデルには現代日本語書き言葉均衡コーパスで学習した Trigram モデルを用いた [9].

3.2 印象評定項目の選定結果

認識率と各印象評定値の相関を分析した．印象評定項目ごとの相関係数の値を Fig. 3 に示す．「a26 : 聞き取りやすい–聞き取りにくい」の項目が最も相関が高くなり相関係数は約 0.6 であった．本稿では以下に

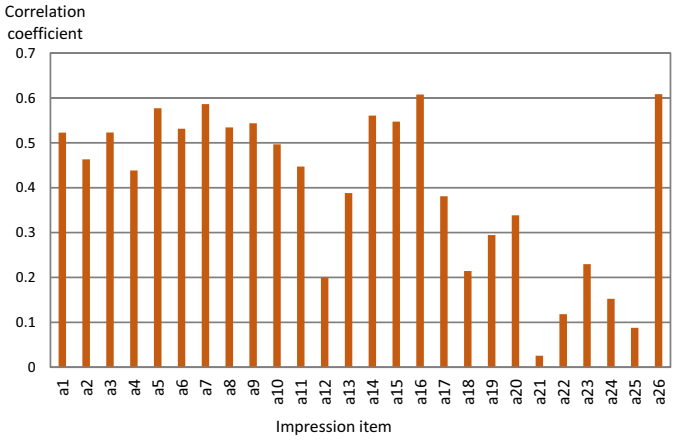

Fig. 3: Result of the experiment.

示す上位 4 つの項目を誤認識の原因となり得る発話特徴として選定することとした．

- 「a5 : 流暢な–たどたどしい」
- 「a7 : なめらかな–しどろもどろな」
- 「a16 : 積極的な–消極的な」
- 「a26 : 聞き取りやすい–聞き取りにくい」

4 印象評定値の推定

4.1 実験の目的と条件

誤認識の原因となる発話特徴として 3 章の実験で選定した印象評定値を音響特徴量から推定する．時間フレームごとに求めた音響特徴量の統計量，および時間フレームごとに求めた音響特徴量の時系列の 2 種類を検討する．時間フレームごとに求めた音響特徴量の統計量としては，OpenSMILE で抽出した

Table 1: Descriptor and functional for the feature set.

Descriptor	Functionals
ZCR	mean
RMS Energy	standard deviation
F0	kurtosis, skewness
HNR	extremes: value, rel, position, range
MFCC 1-12	linear regression: offset, slope, MSE

INTERSPEECH 2009 Emotion Challenge で使用された音響特徴量セット [10](以下 IS09 特徴量と呼ぶ)を使用する．IS09 特徴量は感情認識等に使われているものである．この特徴量は発話の印象と関係すると期待できる．この特徴量セットには 384 次元の音響特徴量が含まれており，IS09 特徴量は Table 1 に示す 16 種類の descriptor と 12 種類の functional の組み合わせとして表される．これらの特徴量は，韻律，スペクトル，音声品質などを表している．次に時間フレームごとに求めた音響特徴量の時系列についてはメルフィルタバンク出力 40 次元を用いる．ここで分析フレーム長は 25ms，フレームシフト幅は 10ms である．

印象評定値の推定手法には SVR，MLP，BLSTM を用いる．ここで BLSTM は RNN（Recurrent Neural Network）の手法の一つである．RNN はニューラルネットワークの手法で時系列データを扱うために考案されたモデルである．現在の時刻の入力として前の時刻の隠れ層の情報を利用することで学習を行う．LSTM は RNN にメモリセルなどを追加し長い系列データに対しても対応できるよう改良された手法である．BLSTM は LSTM に逆方向の隠れ層の情報を追加したモデルであり，これを推定に適用する．時間フレームごとに求めた音響特徴量の統計量である IS09 については SVR，MLP を用いて推定し，時間フレームごとに求めた音響特徴量の時系列であるメルフィルタバンク出力 40 次元については BLSTM を用いて推定を行う．

推定する印象評定項目は 3 章の実験により選定した 4 種類である．これらの項目の印象評定値を評定者 10 人の値で平均し学習の際の教師データとした．実験に用いた音声データは CSJ のコアデータのうち学会講演，模擬講演の音声計 525 である．

Table 2: Condition of the NN.

	MLP	BLSTM
Feature	384th-order IS09	40th-order mel filter bank outputs
Optimization method	Momentam SGD	Adam
Hidden layer number	2	2
Hidden layer size	1024, 512	100,100
Epoch	500	100
Batch size	10	50
Dropout rate	0.3	0.3

SVR のハイパーパラメータの最適化は学習データに対するクロスバリデーションテストにより行った．また MLP と BLSTM の条件を Table 2 に示す．MLP は，隠れ層 2 層で次元は 1024，512，とし出力層の次元は 4 としそれぞれの項目に推定する．最適化手法には MomentumSGD を用いた．BLSTM は隠れ層 2 層で次元数は 100，100 とし逆方向の出力と結合させその後の 4 次元の出力層を用いて各項目を推定する．最適化手法には Adam を用いた．なお BLSTM では 1 分すべてのフレームに教師データとして同じ値を与えた．また入力には 1 分の音声を 10 秒に分割し，それぞれ学習を行った．

4.2 発話特徴推定の結果

推定結果を Fig. 4 に示す．図は 4-fold クロスバリデーションにより得られた真値と推定値の RMSE と相関係数の平均値である．BLSTM については，各フレームの出力を平均し，さらに分割した 10 秒ごとのデータに対して平均したものを最終的な推定値とした．

まず項目ごとに推定した結果を見ると，RMSE が 0.5 から 0.7 の精度で推定できていることが分かる．また「a16 : 積極的な–消極的な」以外の 3 つの項目については BLSTM が僅差ではあるが最もよい結果を示した．今後はさらにより良い特徴量や推定手法を検討する．

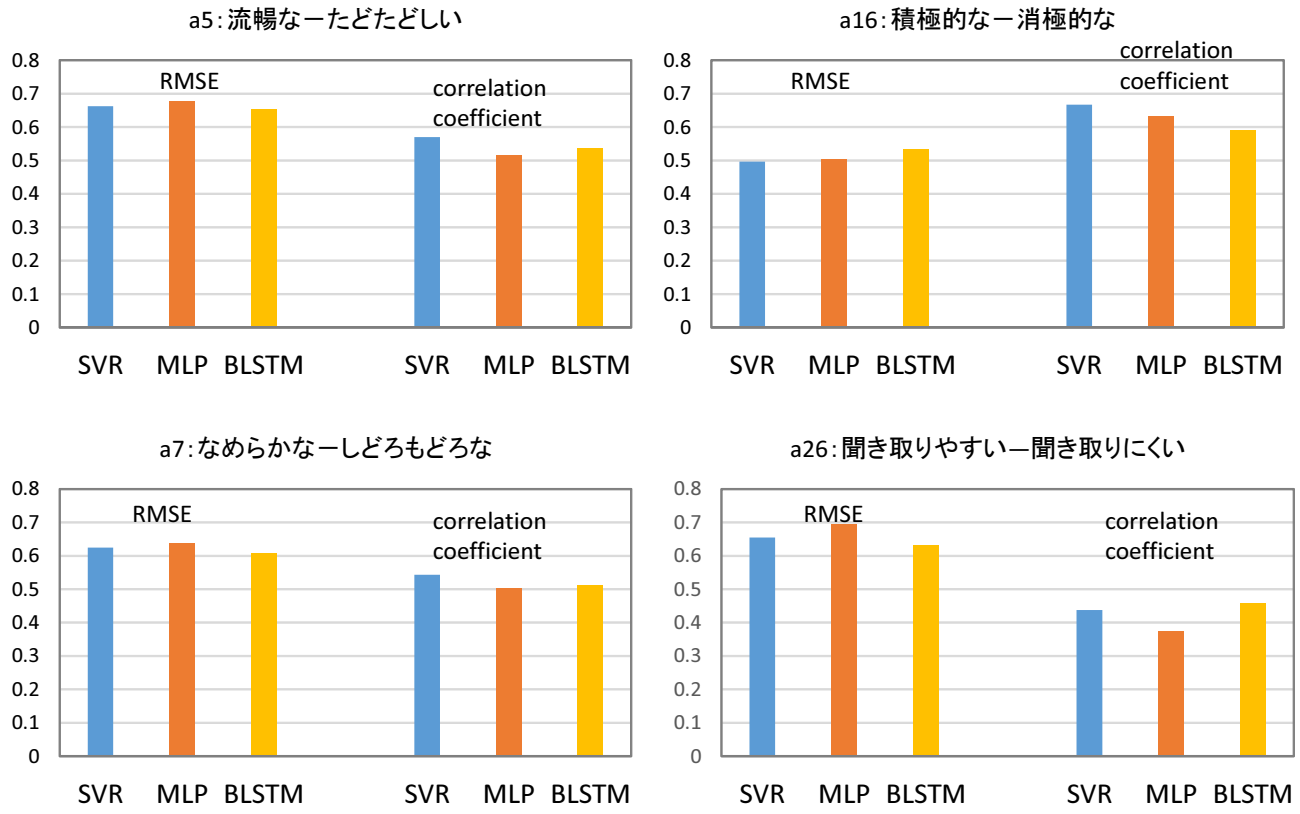

Fig. 4: Result of the experiment.

5 おわりに

本稿では、誤認識の原因となり得る発話特徴として CSJ の印象評定項目を選定し、音響特徴量からその値を推定する手法を提案した．実験により手法の性能の比較検討を行った結果，BLSTM を用いた手法が他の手法と比べてわずかながら良い結果となった．

謝辞 本研究は JPSP 科研費 17K00224 の助成を受けた．

参考文献

[1] J. Hui, “Confidence measures for speech recognition: A survey,” Speech Communication, Vol. 45, No. 4, pp. 455–470, 2005.

[2] 青木智充, 山田武志, 宮部磁樹, 牧野昭二, 北脇信彦, “認識性能予測に基づく雑音環境下音声認識のユーザビリティ改善の検討,” 音講論，pp. 133–136，Mar. 2015.

[3] 後藤孝宏, 山田武志, 牧野昭二, “雑音下音声認識における必要発話音量提示機能の実装と評価,” 音講論，pp. 117–120，Sep. 2016.

[4] 前川喜久雄, “「日本語話し言葉コーパス」の概要,” 日本語科学，Vol. 15,pp. 111–133, 2004.

[5] 山住賢司, 籠宮隆之, 槙洋一, 前川喜久雄, “講演音声の印象評定尺度,” 日本音響学会誌, Vol. 61, No. 6 ,pp. 303–311, 2005.

[6] AJ. Smola and B. Scholkopf, “A tutorial on support vector regression,” Statistics and computing, Vol. 14, No. 3 ,pp. 199–222, 2004.

[7] 久保陽太郎, “ディープラーニングによるパターン認識,” 情報処理, Vol. 54, No. 5 ,pp. 500–508, 2013.

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

[9] http://julius.osdn.jp/

[10] B. Schuller, S. Steidl and A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” INTERSPEECH 2009, pp. 312–315, 2009.