ノンリファレンス特徴量を用いた自然発話音声認識の性能推定の検討*

☆郭翎，山田武志，牧野昭二（筑波大学）

1　はじめに

近年，スマートホンやタブレットの普及と共に，音声認識サービスが広く使われるようになってきた．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．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する際には自然発話音声に対する認識性能を効率的に調査する手法が必要である．この手法を用いることで，音声認識サービスを提供中に認識性能をモニタリングすることが可能となる．また，この手法は音声対話における信頼度尺度として使うことができる．

そのためのアプローチとして音声のひずみの大きさから認識性能を推定する手法がある [1]〜[4]．例えば，我々はこれまでにケプストラム距離と SMR パープレキシティを用いて認識性能を推定する手法を提案した [4]．これらの手法は，実際に認識を実行することなく認識性能を推定することができるが，ひずみの大きさを求めるためにリファレンス音声を必要とするため，自然発話に適用することはできない．

そこで本稿では，自然発話に適用できる認識性能推定手法を提案する．提案手法では，まず openSMILE（open-Source Media Interpretation by Large feature-space Extraction）[5] を用いて，リファレンス音声を必要としない特徴量（ノンリファレンス音響特徴量）を抽出する．そして，SVR（Support Vector Regression）[6] を用いて認識性能を推定する．オンラインゲーム音声チャットコーパス（OGVC）[7] に含まれている音声データを用いた実験により，提案手法の有効性を検証する．

2　提案手法

2.1　概要

提案手法の処理フローを図 1 に示す．提案手法では，まずユーザの発話音声のみからノンリファレンス音響特徴量を抽出する．そして SVR により認識性能を推定する．提案手法においては認識性能の定義とノンリファレンス音響特徴量が重要となるので，以下で詳しく述べる．

2.2　認識性能の定義

上述したように，提案手法を音声認識サービスにおける性能モニタリングや音声対話における信頼度尺度に適用するためには，個々の発話に対する認識性能を推定することが有用であると考えられる．よって，本稿では一発話における単語正解率を用いることにする．

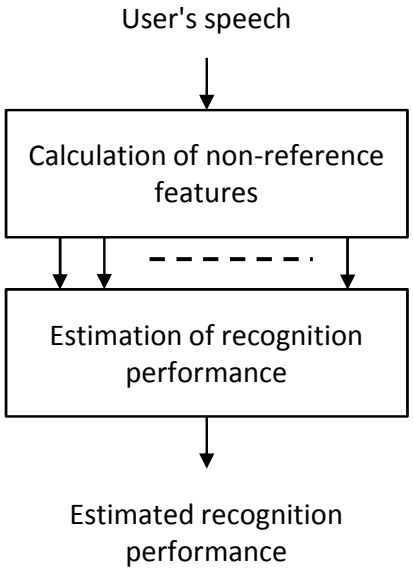


Fig. 1　Estimation of the recognition performance from the non-reference features.

$$\%Corr = \frac{H}{N} \tag{1}$$

ここで，H は正しく認識した形態素 (単語) の数，N は発話に含まれる形態素の総数である．しかし，この定義には N が小さいとき，単語正解率の分解能が低いという欠点がある．例えば，一つの形態素しか含んでいない発話の場合，単語正解率は 0 か 100 のどちらかとなる．このことは認識性能の推定に影響を及ぼすと考えられる．

2.3　ノンリファレンス音響特徴量

1 章で述べたように，自然発話音声にはリファレンス音声が存在しないという問題を解決するために，ノンリファレンス音響特徴量を使用する．本稿では，ノンリファレンス音響特徴量として，openSMILE における INTERSPEECH 2009 Emotion Challenge feature set[8] を用いる．この特徴量セットは感情認識に広く使われているものである．感情音声は典型的な自然発話音声であると考えられるため，この特徴量は自然発話音声の性質を表すと期待できる．

*Performance estimation of spontaneous speech recognition with non-reference feature sets, by Ling GUO, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba).

Table 1　Low-level descriptors for the feature set.

Low-level descriptor	Description
RMSenergy	Root-mean-square signal frame energy
MFCC	Mel-frequency cepstral coefficients 1-12
pcm_zcr	Zero-crossing rate of time signal (frame-based)
voiceProb	Voicing probability computed from the ACF
F0	Fundamental frequency computed from the cepstrum
(a first-order delta coefficient of the above features)	

Table 2　Functionals for the feature set.

Functional	Description
max	Maximum value of the contour
min	Minimum value of the contour
range	max-min
maxPos	Absolute position of the maximum value (in frames)
minPos	Absolute position of the minimum value (in frames)
amean	Arithmetic mean of the contour
linregc1	Slope (m) of a linear approximation of the contour
linregc2	Offset (t) of a linear approximation of the contour
linregerrQ	Quadratic error computed as the difference between the linear approximation and the actual contour
stddev	Standard deviation of the values in the contour
skewness	Skewness (third-order moment)
kurtosis	Kurtosis (fourth-order moment)

この特徴量セットには 384 個の音響特徴量が含まれている．各特徴量は，表 1 と表 2 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，音声品質などの特徴を表している [8].

3　提案手法の有効性の評価

3.1　実験条件

提案手法の有効性を検証するために，自然発話音声の認識性能を推定する実験を行う．

本実験には大語彙連続音声認識エンジン Julius[9]（Rev. 4.3.1）を用いた．音声認識のための特徴量としては，12 次元の MFCC（Mel-Frequency Cepstral Coefficients）とその Δ 係数，及び 1 次元の Δ 対数パワーの計 25 次元を用いる．音響モデルは，ASJ-JNAS データベース [10] のクリーン音声で学習された，性別非依存の DNN-HMM（Deep Neural Network Hidden Markov Model）である．ここで，HMM は Triphone であり，状態確率が DNN によって与えられる．言語モデルは現代日本語書き言葉均衡コーパス [10] で学習された単語 Trigram モデルである．

提案手法では，SVR（Radial basis kernel function）を用いて認識性能を推定する．SVR の学習は，個々の音声データに対するノンリファレンス音響特徴量と単語正解率のペアデータを用いて行う．

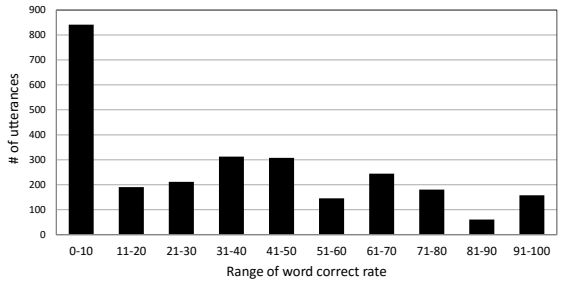


Fig. 2　Histogram of the word correct rate for the speech data used.

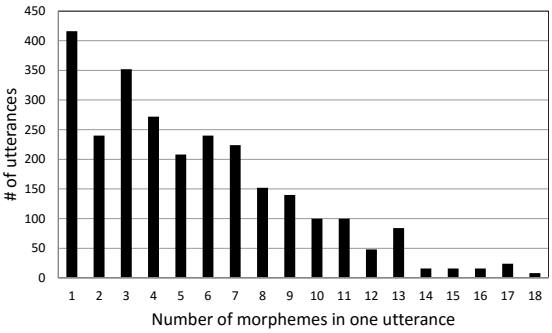


Fig. 3　Histogram of the number of morphemes for the speech data used.

3.2　自然発話音声データ

本実験では，オンラインゲーム音声チャットコーパス [7] に収録された演技音声データを用いた．この音声データは女性 2 名と男性 2 名による計 2656 発話からなる．音声データのサンプリングレートは 16 kHz である．各音声データは 8 種類の感情（恐れ、驚き、悲しみ、嫌悪、怒り、期待、喜び、受容）と 4 種類の強度（普通、弱、中、強）により特徴付けられており，多様な発話スタイルが含まれていると言える．各音声データに対して発話内容の書き起こしテキストが用意されている．なお，未知語率は 0.03%であり，本稿では未知語が認識性能推定に及ぼす影響を考慮しない．

まず，各音声データに対する単語正解率のヒストグラムを図 2 に示す．ここで，横軸は各音声データに対する単語正解率，縦軸は音声データ数を示している．この図から，各音声データに対する単語正解率が全体的に低いことが分かる．平均単語正解率は 35.42%であった．次に，各音声データに含まれる形態素数のヒストグラムを図 3 に示す．横軸は各音声データに含まれる形態素数，縦軸は音声データ数を示している．この図から，各音声データに含まれる形態素数は総じて少ないことが分かる．平均形態素数は 5.49 であった．

Table 3 Correlation coefficient R and RMSE for each speech data set.

Minimum number of morphemes in speech data set (n)	R	RMSE
n=1	0.59	24.54
n=2	0.52	24.56
n=3	0.54	23.38
n=4	0.59	21.11
n=5	0.58	20.19
n=6	0.60	19.12
n=7	0.62	18.04
n=8	0.71	15.76
n=9	0.77	13.77
n=10	0.84	11.55

3.3 提案手法の有効性の検証

提案手法の有効性を検証するために，10 セットの音声データを用意した．各セットは n 個以上の形態素を含む音声データからなる（$n = 1, 2, ..., 10$）．これは 2 章で述べたように，一発話に含まれる形態素数が認識性能推定にどの程度影響を及ぼすかを調べるためである．本実験では，各音声データセットを 4 つのグループに分け，話者オープンのクロスバリデーションテストを行う．各グループには一人の話者の音声データが含まれている．3 つのグループの音声データを用いて学習し，それ以外の一つのグループの音声データを用いて評価する．

各音声データセットに対する推定単語正解率と真の単語正解率の相関係数 R と RMSE（Root Mean Square Error）を表 3 に示す．ここで，表中の R と RMSE はクロスバリデーションによって得られた 4 つの結果の平均である．表より，n が大きいときほど，推定誤差が小さいことが分かる．$n = 10$ のとき，RMSE は 11.55 であった．n が小さいときに，RMSE が大きくなるのは，単語正解率の分解能が低いことが原因であると考えられる．

$n = 10$ のときの推定単語正解率と真の単語正解率の関係を図 4 に示す．ここで，図 4 の (a)〜(d) はクロスバリデーションによって得られた各話者の結果である．図から，R と RMSE はそれぞれ 0.83〜0.85，10.90〜11.88 であり，提案手法は話者の違いにさほど依存せずに良好な推定ができることが分かる．

Table 4 Correlation coefficient R and RMSE for each feature set.

Feature set	R	RMSE
INTERSPEECH 2009 Emotion Challenge	0.84	11.55
openSMILE/openEAR ‘emobase’ set	0.86	11.77
INTERSPEECH 2010 Paralinguistic Challenge	0.84	12.10
Large openSMILE emotion feature set	0.82	12.59

3.4 ノンリファレンス音響特徴量の比較

より高次元で表現力が豊かな特徴量セットを用いることにより，さらに高い推定精度が得られると期待できる．そこで，openSMILE/openEAR ‘emobase’ set[5](988 次元)，INTERSPEECH 2010 Paralinguistic Challenge[11](1582 次元)，Large openSMILE emotion feature set[5](6552 次元) を提案手法の特徴量セットとして用いて，前節と同様の実験を行った．

$n = 10$ のときの，各特徴量セットを用いたときの推定単語正解率と真の単語正解率の相関係数 R と RMSE を表 4 に示す．ここで，表中の R と RMSE はクロスバリデーションによって得られた 4 つの結果の平均である．表より，特徴量の次元数が増えても，RMSE は改善していないことが分かる．これは学習に用いる音声データ量の不足が原因の一つであると考えられる．今後，音声データ量を増やして，再度検証する必要がある．

4 まとめ

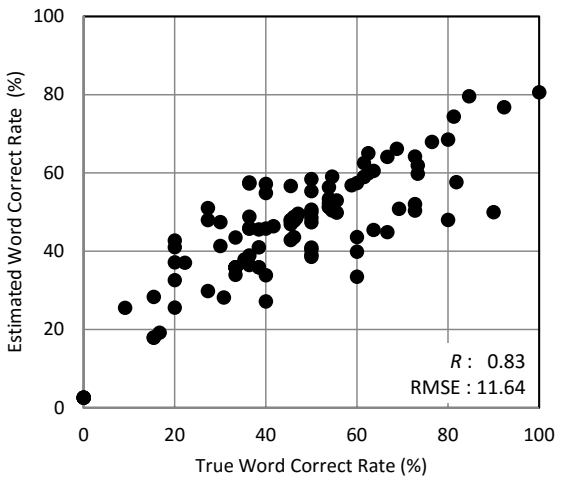
本稿では，ノンリファレンス音響特徴量を用いた自然発話音声の認識性能推定手法を提案した．ノンリファレンス音響特徴量として INTERSPEECH 2009 Emotion Challenge feature set を用いて推定実験を行った結果，一発話に含まれる形態素数が十分多い発話に対しては良好な精度で単語正解率を推定ができることが分かった．また，提案手法は話者の違いにロバストであることを確認した．

参考文献

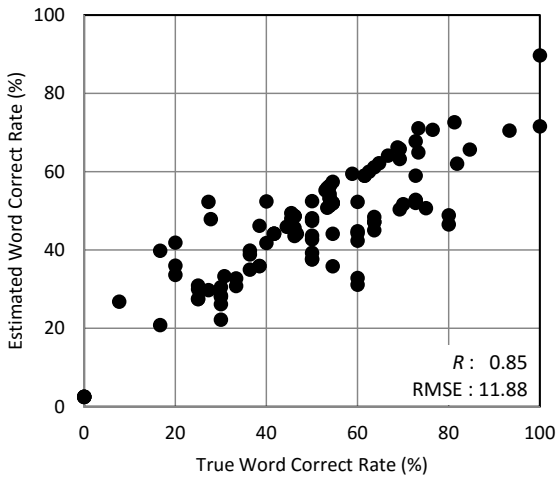
[1] H. Sun, L. Shue and J. Chen, “Investigations into the relationship between measurable speech quality and speech recognition rate for telephone speech,” Proc. ICASSP 2004, Vol. 1, pp. 865–868, May. 2004.

[2] T. Yamada, M. Kumakura and N. Kitawaki, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 6, pp. 2006–2013, Nov. 2006.

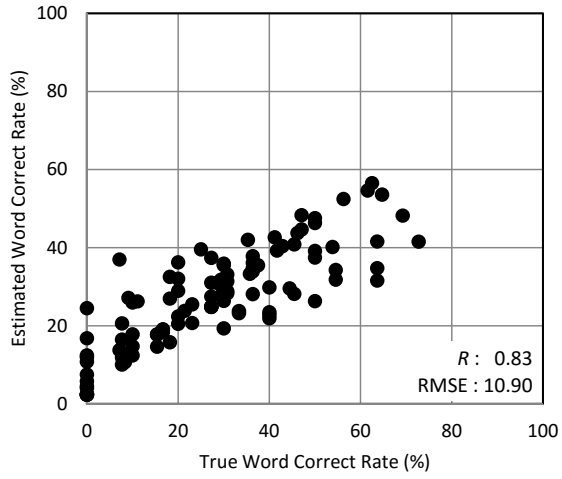
[3] T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, “Estimation of speech recognition



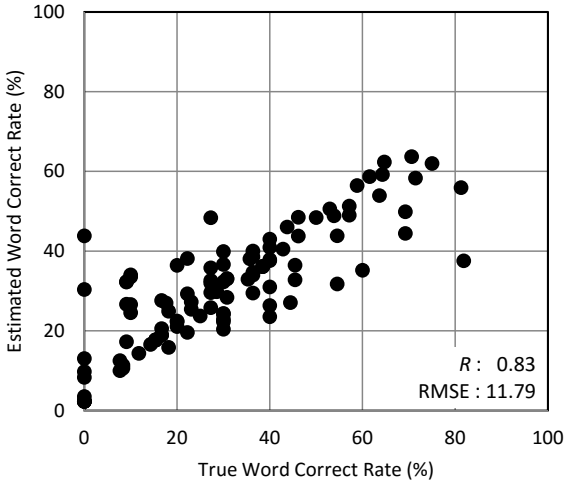
(a) Female 1



(b) Female 2



(c) Male 1



(d) Male 2

Fig. 4 Relationship between the true word correct rate and the estimated word correct rate for each speaker.

performance in noisy and reverberant environments using PESQ score and acoustic parameters," Proc. APSIPA ASC 2013, Paper ID: 144, Oct. 2013.

[4] 郭翎，山田武志，宮部滋樹，牧野昭二，北脇信彦 "ケプストラム距離と SMR-パープレキシティを用いた雑音下音声認識の性能推定の検討," 日本音響学会春季研究発表会, pp. 145–148, Sep. 2015.

[5] F. Eyben, M. Wollmer and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010. pp. 1459–1462.

[6] V.N. Vapnik, *Statistical Learning Theory*, A Wiley-Interscience Publication, 1998.

[7] http://research.nii.ac.jp/src/en/OGVC.html

[8] B. Schuller, S. Steidl and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," Proc. INTERSPEECH 2009, pp. 312–315, 2009.

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

[10] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano, "Free software tool kit for Japanese large vocabulary continuous speech recognition," Proc. ICSLP 2000, pp. 476–479, Oct. 2000.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C.A. Muller and S.S. Narayanan "The INTERSPEECH 2010 paralinguistic challenge," Proc. INTERSPEECH 2010, pp. 2795–2798, 2010.