音響モデルの精度を考慮した雑音下音声認識の性能推定の検討*

☆高岡隆守，山田武志，牧野昭二，北脇信彦（筑波大）

1　はじめに

音声認識サービスを提供する際には，対象とする環境でどの程度の認識性能が得られるのかを事前に調査する必要がある．現時点で最も確実な方法は，サービスを運用する現場で認識実験を行うことである．しかし人的，時間的コストが極めて大きく，また専門的な知識や技術を要することが問題となっている．この問題を解決することにより音声認識サービスのさらなる普及が期待できるため，雑音下音声認識の性能を簡便に推定する手法が必要不可欠である．

我々はこれまでに，音声のひずみの大きさから認識性能を推定する手法を開発し，その有効性を示した[1]．本手法では，ひずみの大きさと認識性能の関係を表す推定式を実験的に求める．しかし，認識タスクや音響モデルが異なると同じ雑音環境であっても認識性能は変動するため，対象とする認識タスクや音響モデルに専用の推定式をその都度求める必要があった．我々はこの問題に対処するため，認識タスクの複雑さをパラメータに持つ推定式，音響モデルの精度をパラメータに持つ推定式をそれぞれ提案し，有効性を示した[2, 3]．

本稿では，さらに認識タスクの複雑さと音響モデルの精度の両方をパラメータに持つ推定式を提案し，その有効性を検証する．このような推定式を一度求めておけば，以降は各パラメータを指定することにより，対象とする認識タスク・音響モデルに専用の推定式を容易に得ることができる．

2　雑音下音声認識の性能推定法

認識性能推定の基本的な流れをFig. 1を用いて説明する．まず，原音声（雑音が重畳していない音声）と劣化音声（雑音が重畳している音声）を入力とし，劣化音声のひずみの大きさを計算する．そしてそのひずみの大きさを以下に示す推定式に代入することで認識性能を推定する[1]．

$$y = f(x) = \frac{a}{1+e^{-b(x-c)}} \tag{1}$$

ここで，y は認識性能の推定値，x はひずみの大きさである．また，a, b, c は定数であり，a は原音声に対する認識性能，b はひずみが大きくなったときの認識性能の低下の急峻さ，c はひずみに対する頑健性に相当する．これらの値は，劣化音声に対するひずみの大きさと認識性能を実験的に求め，両者の関係を最適近似することにより決定する．ただし，上述したように，認識タスクや音響モデルが異なるとひずみの大きさが同じであっても認識性能が変動するため，認識タスクや音響モデルに専用の推定式を求める必要がある．

我々はこの問題に対処するため，まず認識タスクの複雑さをパラメータとして持つ推定式，音響モデルの精度をパラメータに持つ推定式をそれぞれ提案した[2, 3]．

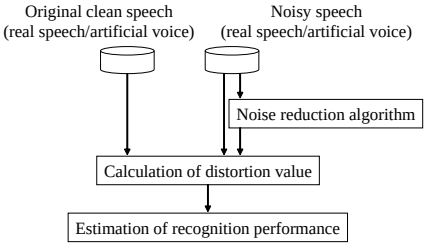

Fig. 1 Overview of the recognition performance estimation

$$y = f(x,\alpha) = \frac{p_1\alpha^{q_1} + r_1}{1+e^{-(p_2\alpha^{q_2}+r_2)(x-(p_3\alpha^{q_3}+r_3))}} \tag{2}$$

$$y = f(x,\beta) = \frac{s_1\beta + t_1}{1+e^{(-s_2\beta-t_2)(x-s_3\beta-t_3))}} \tag{3}$$

ここで α は認識タスクの複雑さを表すSMRパープレキシティ（各単語の出現確率の逆数の相加平均），β は音響モデルの精度を表す原音声に対する誤認識率である．式(2)と式(3)は，式(1)の a, b, c を αの関数，βの関数にそれぞれ置き換えたものである．

3　提案法

認識タスクの複雑さと音響モデルの精度の両方をパラメータに持つ推定式を提案する．提案法は，式(1)の a, b, c を，式(2)と式(3)におけるαの関数とβの関数の一次結合に置き換えたものである．具体的な推定式は以下である．

$$y = f(x,\alpha,\beta) = \frac{p_1\alpha^{q_1} + r_1\beta + s_1}{1+e^{-(p_2\alpha^{q_2}+r_2\beta+s_2)(x-(p_3\alpha^{q_3}+r_3\beta+s_3))}} \tag{4}$$

ここで α はSMRパープレキシティ，β は原音声に対する誤認識率である．p_1～p_3, q_1～q_3, r_1～r_3, s_1～s_3 は定数であり，様々な認識タスクや音響モデルを対象として劣化音声に対するひずみの大きさと認識性能を実験的に求め，両者の関係を最適近似することで決定する．

4　提案法の有効性の検証

本章では，様々な認識タスクと音響モデルの組み合わせを対象とし，提案法による認識性能の推定精度を検証する．ここで，推定式の定数の最適化および認識性能推定には，計算機上で加法性雑音のみを重畳した音声データを用いる．なお，本実験ではひずみ尺度としてPESQ[4]を用いた．

4.1 推定式の定数の決定

認識実験のために用いた認識タスクを以下に述べる．

- 孤立単語認識タスク：音声データとして，東北大－松下単語音声データベース[5]に収録されている鉄道駅名3285語を読み上げたものを用いた．本実験では語彙サイズを100, 200, 400, 800, 1600, 3285とすることにより，6種類

* Performance estimation of noisy speech recognition considering the accuracy of acoustic models, by Takashi Takaoka, Takeshi Yamada, Shoji Makino, and Nobuhiko Kitawaki (University of Tsukuba).

の認識タスクを設定した．

- 記述文法認識タスク：音声データとして，AURORA-2J[6]と同じ1～7桁の数字列を読み上げたものを用いた．単語（数字）の数は読みの違いを含めて11であり，これらを任意回数繰り返すという記述文法を作成した．認識タスクは1種類である．
- 大語彙連続認識タスク：音声データは，新聞記事読み上げ音声コーパス（JNAS）のテストセット100文（男性話者）である．語彙サイズは2種類用意した．言語モデルとしては，IPAの「日本語ディクテーション基本ソフトウェア1999年度版」[7] に含まれている3-gramモデルのうち，語彙サイズ5k，20k，60kの3種類を用いた．語彙サイズと言語モデルを組合せることにより5種類の認識タスクを設定した．

各認識タスクの音声データには，電子協騒音データベース[8]の雑音（car1, hall1, train2, lift2）を20, 15, 10, 5, 0, -5dB のSNRで重畳した．サンプリング周波数は16kHzである．これらの雑音重畳音声データを認識するための音響モデルとしては，IPAの「日本語ディクテーション基本ソフトウェア1999年度版」に含まれているモノフォン性別非依存モデル（4, 8, 16, 64 混合分布）およびトライフォン性別非依存モデルの5種類を用いた．

上記の12種類の認識タスクと5種類の音響モデルの組み合わせの60種類に対し，SMRパープレキシティ，原音声に対する誤認識率，単語正解精度，PESQ Scoreを求め，これらの関係を最適近似するように式(4)の定数を求めた結果，次式が得られた．

$$y = f(x,\alpha,\beta) = \frac{a(\alpha,\beta)}{1+e^{-b(\alpha,\beta)(x-c(\alpha,\beta))}} \quad (5)$$

$$a(\alpha,\beta) = -0.66\alpha^{0.05} - 0.96\beta + 99.9$$
$$b(\alpha,\beta) = -8.3\times10^{-9}\alpha^{1.67} - 0.0091\beta + 4.23$$
$$c(\alpha,\beta) = -6.97\alpha^{-0.011} + 0.0062\beta + 8.51$$

式(5)に x と α, β の値を代入することにより，単語正解精度の推定値を得ることができる．また，式(5)の α, β にのみ値を代入することにより，対象とする認識タスク・音響モデルの推定式（式(1)に相当）を得ることもできる．

4.2 実験結果

4.1 節で述べた60種類の認識タスクと音響モデルの組み合わせ各々の一部に対する推定式をFig. 2に示す．ここで，横軸はPESQ Scoreであり，値が小さいほど音声が劣化していることを表す（値が4.5のときは原音声に一致する）．また，図中の曲線は認識タスクと音響モデルの各組合せに対する推定式を表しており，式(5) に α, β の値を代入することにより求めた．Fig. 2より，各推定式の上下の位置関係は概ね α, β の値によって定まることが分かる．

次に4.1 節で述べた60種類の認識タスクと音響モデルの組み合わせ各々に対して単語正解精度の推定を行った．推定のために用いた雑音重畳音声データは推定式の最適化のために用いたものと同じである．Fig. 3に真の単語正解精度と推定した単語正解精度の関係を示す．真の単語正解精度と推定した単語正解精度の決定係数は0.98，RMSEは2.5であり，高い精度で推定できていることが示された．しかしFig. 3から推定精度が他と比べて低い点（Fig. 3の赤点）も存在することが分かる．調査の結果，PESQ Scoreが推定式の変曲点付近にある場合に推定精度が低いことが分かった．これは，定数 b, c の表現方法に改善の必要があることを示唆している．

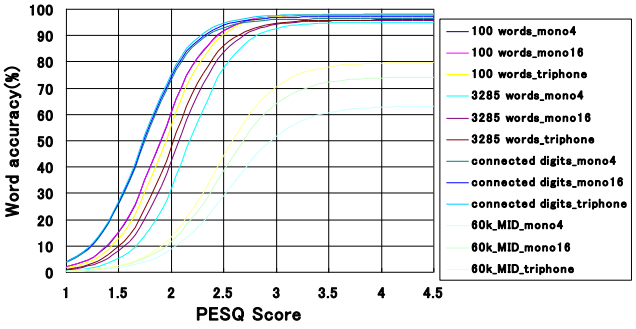

Fig. 2 The estimators for each pair of the recognition task and the acoustic model

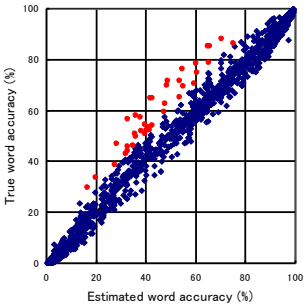

Fig. 3 Relationship between the true word accuracy and the estimated word accuracy

5 おわりに

本稿では，認識タスクの複雑さと音響モデルの精度の両方をパラメータに持つ推定式を提案した．また，その有効性を検証するために，推定式の最適化および，認識性能の推定実験を行った．その結果，提案法により高い精度で認識性能を推定できることが示されたものの，推定式の変曲点付近のPESQ Scoreがある場合に推定精度が低いことが分かった．今後はこの問題の解決を図る予定である．

参考文献

[1] T. Yamada *et al.*, "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.

[2] 中島智弘ら.,〝雑音下音声認識の性能推定に用いるタスクの複雑さを表す尺度の検討,〟日本音響学会春季研究発表会, pp. 147-150, Mar. 2009.

[3] T. Takaoka *et al.*, "Performance estimation of noisy speech recognition considering the accuracy of acoustic models," TJASSST11, Sat-No331, Nov. 2011.

[4] ITU-T Rec. P.862., "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[5] S. Makino *et al.*, "Tohoku University and Matsushita isolated spoken word database," Journal of the Acoustical Society of Japan, Vol. 48, No. 12, pp. 899-905, 1992.

[6] S. Nakamura *et al.*, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp. 535-544, Mar. 2005.

[7] 河原達也ら.,〝日本語ディクテーション基本ソフトウェア（99 年度版）,〟日本音響学会誌, Vol. 57, No. 3, pp. 210-214, Mar. 2001. 2005.

[8] http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.