複数の音声認識システムの出力の共通部分を用いた認識誤り検出*

小玉 康広　宇津呂 武仁　西崎 博光　中川 聖一
豊橋技術科学大学 工学部 情報工学系
{kodama,utsuro}@cl.ics.tut.ac.jp, {nisizaki,nakagawa}@slp.ics.tut.ac.jp

1 はじめに

近年，音声認識結果の正解部分と誤り部分を分離することを目的として信頼度 (確信度，Confidence Measure) の研究が行なわれている (例えば，連続音声認識では [Kemp97, Wessel99, 緒方 00, 堀部 01] など)．ここで，これまで提案されてきた信頼度尺度の多くは，いずれも，単一の認識エンジン・認識モデルが出力する認識結果を用いて，その正解部分と誤り部分を分離するというものであった．一方，連続音声認識の認識率そのものの向上を目的とする研究においては，複数の認識システムの出力を統合する方式も提案され，一定の効果が報告されている ([Fiscus97, Schwenk00] など)．本論文では，音声認識結果の正解部分と誤り部分を分離するための信頼度尺度として，複数の音声認識システムの出力の共通部分を用いる方法を提案し，その有効性を示す．

2 実験条件

大語彙連続音声認識システムとしては，
SPOJUS [赤松 98] および Julius [河原 00] を使用した．それぞれのシステムの特性を表 1 に示す．評価データとしては新聞記事読み上げ音声コーパス (JNAS) の 100 文 (男性話者 10 人,1565 単語) の音声，および，NHK のニュース「ニュース 7」と「おはよう日本」(1996 年 6 月 1 日) の 175 文 (男性話者,6813 単語) の二種類を使用した．これらの評価データに対する各認識システムの単語認識率を表 1 に示す．

3 認識正解単語の推定

3.1 複数システムの認識結果 (1-best) を用いる方法

3.1.1 複数システムの認識結果の共通部分

複数の認識システムの認識結果 (1-best) の共通部分の単語を正解とする方法により認識正解単語の推定を行ない，再現率・適合率 (精度) を評価した．なお，本論文のいずれの方法においても，複数の認識結果の対応付けは，DP マッチングにより行なった．

評価データとして新聞記事読み上げ音声を用いた場合には，以下について評価を行なった．

- 以下の三種類の単一システム
  - SPOJUS(新聞 tri-gram) — 図 1 中 “SM2”
  - Julius(新聞 {bi,tri}-gram)
    — 図 1 中 “JM{1,2}” の二種類
- SM2，JM{1,2} の任意の二つの組合わせの出力の共通部分
- SM2，JM{1,2} の三システムの出力の共通部分
  — 図 1 中 “all”

これらの評価結果を図 1 に示す．再現率・適合率の両方から総合的に判断すると，SPOJUS(新聞 tri-gram) と Julius(新聞 tri-gram) の組合わせが，再現率 80%以上，適合率 99%以上という最も良い結果が得られた．

評価データとしてニュース音声を用いた場合には，以下について評価を行なった．

- 以下の五種類の単一システム
  - SPOJUS(ニュース tri-gram)
    — 図 2 中 “SN2”
  - Julius({ ニュース，新聞 }{bi,tri}-gram)
    — 図 2 中 “J{N,M}{1,2}” の四種類
- 五種類のシステムの任意の組合わせの出力の共通部分

これらの評価結果を図 2 に示す (複数システムの組合わせについては主なものを抜粋)．こちらも，再現率・適合率の両方から総合的に判断すると，SPOJUS(ニュース tri-gram) と Julius({ ニュース，新聞 }tri-gram) の組合わせが，再現率 40%程度，適合率 94%程度という最も良い結果が得られた．

3.1.2 複数システムの認識結果の過半数一致部分

複数の認識システムの認識結果の過半数が一致する単語を正解とする方法により認識正解単語の推定を行な

*Recognition Error Detection based on Agreement among Multiple Speech Recognizers' Outputs

表 1: 大語彙連続音声認識システムの特性

		SPOJUS	Julius
音響モデル等		音節モデル, 5 状態, 4 混合, 性別依存 (男性), 12KHz サンプリング, フレーム周期 8ms	triphone モデル, 3 状態, 16 混合, 性別依存 (男性), 16KHz サンプリング, フレーム周期 10ms
特徴ベクトル		LPC·MEL·CEP (10 次元 × 4 フレームを KL 展開で 20 次元に圧縮) + ΔCEP + ΔΔCEP + ΔPOW + ΔΔPOW (計 42 次元)	MFCC(12 次元) + ΔMFCC + ΔPOW (計 25 次元)
言語モデル		NHK 汎用ニュース原稿 (5 年分) または 毎日新聞 (45ヵ月分) から作成した tri-gram モデル (語彙数 2 万)	NHK 汎用ニュース原稿 (5 年分) または 毎日新聞 (75ヵ月分) から作成した bi-gram モデル/tri-gram モデル (語彙数 2 万)
単語認識率	新聞	正解率 86.8%, 正解精度 83.1% (新聞 tri-gram)	正解率 91.3%, 正解精度 87.7% (新聞 tri-gram)
	ニュース	正解率 51.1%, 正解精度 45.9% (ニュース tri-gram)	正解率 63.3%, 正解精度 53.3% (ニュース tri-gram) 正解率 62.3%, 正解精度 56.1% (新聞 tri-gram)

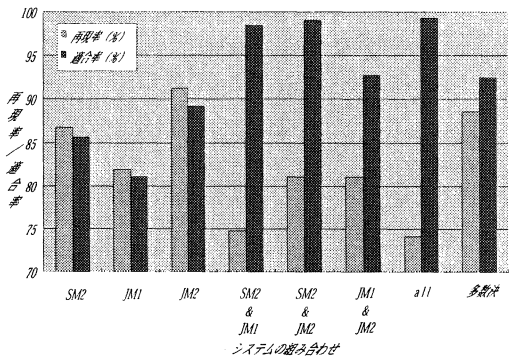

図 1: 認識正解単語推定の性能 (新聞記事読上げ音声)

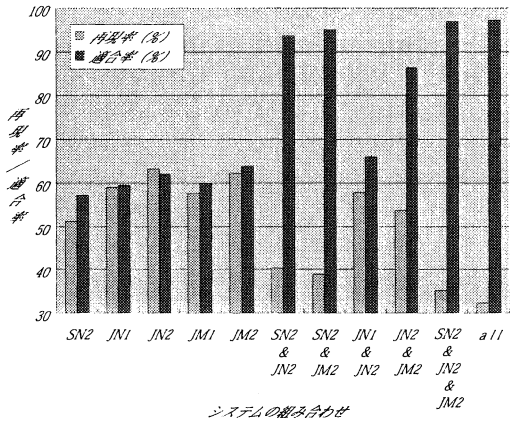

図 2: 認識正解単語推定の性能 (複数システム共通部分, ニュース音声)

い, 再現率・適合率を評価した. 評価データと複数認識システムの組合わせは以下の通りである.

- 評価データとして新聞記事読み上げ音声を用いた場合— SM2, JM{1,2} の三システム (図 1 中 "多数決")
- 評価データとしてニュース音声を用いた場合 — SN2, J{N,M}{1,2} の五システムのうち, 任意の三システム以上の組合わせ (図 3 に抜粋)

さらに, 表 2 では, 再現率・適合率の両方から総合的に判断した最も良い結果と, 共通部分を正解とする前節の方法による最も良い結果を比較した. ある程度の再現率でかつ単一のシステムを越える高い適合率という条件で評価すると, 新聞記事読み上げ音声・ニュース音声のいずれにおいても, SPOJUS(tri-gram) と Julius(tri-gram) という特定の組合わせの共通部分の単語を正解とする前節の方法が最も良い結果となった.

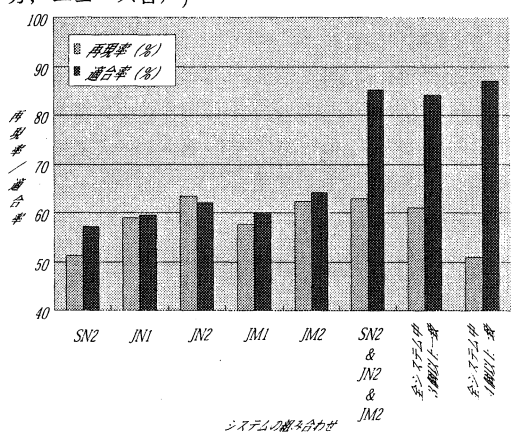

図 3: 認識正解単語推定の性能 (複数システム過半数一致部分, ニュース音声)

表 2: 認識正解単語の推定の性能 (%)

評価データ	方式		再現率	適合率
新聞記事読み上げ音声	複数システム	**SPOJUS-M-tri,Julius-M-tri 共通部分**	**81.0**	**99.1**
		S-M-Tri,J-M-{bi,tri} の過半数一致部分	88.6	92.5
		S-M-Tri,J-M-{bi,tri} の認識結果の和 (適合率 92.7%以上)	88.6	93.0
	単一システム (N-best 中の出現割合で足切り)	SPOJUS-M-tri, 出現割合 0.55 以上	80.4	90.6
		SPOJUS-M-tri, 出現割合 1.00	39.9	98.3
		Julius-M-tri, 出現割合 0.80 以上	82.4	93.1
		Julius-M-tri, 出現割合 1.00	66.0	94.5
ニュース音声	複数システム	**SPOJUS-N-tri,Julius-N-tri 共通部分**	**40.3**	**93.8**
		SPOJUS-N-tri,Julius-M-tri 共通部分	**38.9**	**95.0**
		S-N-tri,J-N-{bi,tri},J-M-tri の過半数一致部分	56.7	86.6
		S-N-tri,J-{N,M}-{bi,tri} の 3 つ以上一致部分	61.0	84.0
		S-N-tri,J-{N,M}-{bi,tri} の 4 つ以上一致部分	50.8	87.1
		S-N-tri,J-{N,M}-{bi,tri} の認識結果の和 (適合率 93.8%以上)	44.5	92.1
	単一システム (N-best 中の出現割合で足切り)	SPOJUS-N-tri, 出現割合 0.95 以上	42.4	74.4
		Julius-N-tri, 出現割合 1.00	64.0	82.0
		Julius-M-tri, 出現割合 1.00	63.3	83.7

3.1.3 複数システムの認識結果の和 (適合率に下限)

まず，認識結果中の各単語 w に，3.1.1 節で測定した複数システムの出力の共通部分の適合率をスコアとして割当てる．ただし，単語 w を共通部分に含む複数システムの組合わせが何通りかあり，それぞれの組合わせの適合率が異なる場合には，最も高い値を割当てる．例えば，単語 w を共通部分に含む複数システムの組合わせとその適合率が以下のように与えられていて，各適合率の間の大小関係が $p_{abc} > p_{ab}, p_a$ の場合は，最大値 p_{abc} を単語 w のスコアとする．

システムの組合わせ	A	$A\&B$	$A\&B\&C$
適合率	p_a	p_{ab}	p_{abc}

そして，このスコアが下限値を上回る単語を正解とする方法により認識正解単語の推定を行ない，再現率・適合率を評価した．表 2 にこの方法と他の方法の比較結果を示すが，前節と同様，特定のシステムの組合わせの共通部分の単語を正解とする 3.1.1 節の方法の結果は上回らなかった．

3.2 単一システムの認識結果 (N-best) を用いる方法

次に，比較のために，単一システムの認識結果 (200-best) から算出した信頼度についても同様の評価を行なった．ここでは，単語グラフ中のエッジ接続数 [緒方 00] や仮説密度 [Kemp97] を用いる信頼度を参考にして，単一システムの認識結果 (200-best) 中での各単語の出現割合を信頼度として用い，出現割合が下限値以上の単語を正解とする方法により認識正解単語の推定を行ない，再現率・適合率を評価した．表 2 で比較するように，前節と同様，特定のシステムの組合わせの共通部分の単語を正解とする 3.1.1 節の方法の結果は上回らなかった．

また，その他に，SPOJUS(新聞 bi-gram) による新聞記事読み上げ音声の認識結果に対する認識正解単語の推定において，音響尤度を用いた推定結果と言語尤度を用いた推定結果の論理和をとる方法で，再現率 45%程度，適合率 90%以上という結果が得られている [堀部 01]．ここでも，特定のシステムの組合わせの共通部分の単語を正解とする 3.1.1 節の方法の結果は，この結果を上回っている．

4 確信度 (信頼度) を用いた N-best 認識結果の絞り込み

3.1.3 節の方法で求めた各単語のスコアを用いて，N-best 認識結果の絞り込みを行なった．N-best 認識結果の各解に対して，その解に含まれる単語のスコアの総和を認識結果全体のスコアとし，N-best 認識結果中でスコア最大の解の単語認識率を評価した．ここで，各

表 3: 確信度 (信頼度) を用いて N-best 認識結果を絞り込んだ場合の単語認識率 (%)

新聞記事読み上げ音声					
システム (適合率の閾値)	尤度最大		出力結果中で精度最大		出力
	正解率	正解精度	正解率	正解精度	結果数
SPOJUS-M-tri (200-best)	86.8	83.1	92.1	89.3	200.0
SPOJUS-M-tri (0.891)	87.9	83.9	90.8	87.6	32.3
Julius-M-tri (200-best)	91.3	87.7	93.9	90.4	83.6
Julius-M-tri (0.891)	91.4	87.7	92.4	88.6	30.9

ニュース音声					
システム (適合率の閾値)	尤度最大		出力結果中で精度最大		出力
	正解率	正解精度	正解率	正解精度	結果数
SPOJUS-N-tri (200-best)	51.1	45.9	55.3	50.5	200.0
SPOJUS-N-tri (0.938)	51.1	45.9	53.9	48.9	53.7
Julius-N-tri (200-best)	63.3	53.3	64.2	54.2	23.4
Julius-N-tri (0.938)	63.5	53.3	63.8	53.7	9.8
Julius-M-tri (200-best)	62.3	56.1	63.4	57.4	46.5
Julius-M-tri (0.938)	62.6	56.4	63.0	57.0	19.9

単語のスコアに閾値を設定し，N-best 認識結果の各解のスコア算出の際には，スコアが閾値以上の単語のみを考慮するようにした[1]．閾値を変化させて最も性能がよかった場合と，200-best 全体の場合について，尤度最大の解，および，出力結果中で精度最大の解の，単語正解率・単語正解精度を表 3 に示す．表中の出力結果数と尤度最大の解の認識率から分かるように，出力される認識結果数が 1/2~1/6 程度に絞り込まれ，その中で尤度最大となる認識結果の単語認識率も若干向上している．ただし，出力結果中で精度最大の解の認識率はやや下がることから，200-best 全体での最適解は選択できていない．

5 おわりに

音声認識結果の正解部分と誤り部分を分離するための信頼度尺度として，複数の音声認識システムの出力の共通部分を用いる方法についてその有効性を示した．今後は，正解部分・誤り部分の分離誤り率 (Confidence Error Rate) の評価，他の様々な信頼度尺度との比較，音響モデルや音響・言語重みなどが異なる複数システムの共通部分の評価などを行なう予定である．また，機械学習の枠組みにより，信頼度に関する複数の手がかりを統合する方式 [Kemp97] についても検討を行なう．

謝辞

ニュース音声データベース，ニューステキストデータベースを提供して頂いた NHK 放送技術研究所の関係諸氏に深く感謝する．

[1] スコアが閾値以上の単語数で正規化する方法も試したが，正規化をしない場合の方が高い性能が得られた．

参考文献

[赤松 98] 赤松裕隆, 花井建豪, 甲斐充彦, 峯松信明, 中川聖一：新聞・ニュース文をタスクとした大語彙連続音声認識システムの評価, 情報処理学会第 57 回全国大会講演論文集, pp. 35–36 (1998).

[Fiscus97] Fiscus, J. G.: A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–354 (1997).

[堀部 01] 堀部千寿, 中川聖一：音響尤度と言語尤度を用いた単語正解率の算出, 日本音響学会 2001 年春季研究発表会講演論文集 (2001).

[河原 00] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏：日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価, 情報処理学会研究報告, Vol. 2000, No. (2000–SLP–31), pp. 9–16 (2000).

[Kemp97] Kemp, T. and Schaaf, T.: Estimating Confidence using Word Lattices, *Proceedings of the 5th Eurospeech*, pp. 827–830 (1997).

[緒方 00] 緒方淳, 有木康雄：信頼度を組み込んだデコーディングによる 音声認識の検討, 情報処理学会研究報告, Vol. 2000, No. (2000–SLP–32), pp. 1–6 (2000).

[Schwenk00] Schwenk, H. and Gauvain, J.-L.: Combining Multiple Speech Recognizers using Voting and Language Model Information, *Proceedings of the 6th ICSLP*, Vol. II, pp. 915–918 (2000).

[Wessel99] Wessel, F., Macherey, K. and Ney, H.: A Comparison of Word Graph and N-Best List based Confidence Measures, *Proceedings of the 6th Eurospeech*, pp. 315–318 (1999).