自由発話音声における音素の揺らぎについて

A Discussion of Acoustic Problems in Sponteneous Speech Recognition

村上仁一

Jin'ichi Murakami

ATR 自動翻訳電話研究所

ATR Interpreting Telephony Research Labs.

概 要

本論文では、音響的な特徴について、自由発話と朗読発声を比較した結果を述べる。自由発話の音声認識は今後の大きな課題であるが、従来の朗読発声の認識と比較すると、さまざまな困難が予想される。したがって、まず始めに自由発話の持つ特徴を把握するために、音響的な揺らぎの調査を行なった。その結果、発話速度は自由発話と朗読発声に差は見られなかったが、視察によるラベリング作業において決定することができなかった音素境界 (融合ラベル) は、朗読発話と比較すると、自由発話では、約 2 割増加した。また、音素認識実験を行なったところ、誤り率は約 2 倍に増加した。ただし、これらの値は話者によって差がある。

abstract

In this paper, we present a preliminary study of spontaneous speech recognition, describing both the acoustic characteristics of spontaneous speech.
Recognition of spontaneous speech is one of the hardest problems in the speech recognition area, it seems for example to be much more difficult than read speech recognition. To get a better understanding of spontaneous speech, a preliminary study was done to compare spontaneous and read speech. In hand-labeled spontaneous speech, the labeling uncertainty increased by about 20% in number. A phoneme recognition experiment resulted in a factor of two increase in the error rate.

1 まえがき

近年、連続音声認識の研究が盛んに行なわれ、いくつかの研究機関で文音声システムが構築されている [1],[2]。しかし、これらのシステムの多くは、朗読音声のような丁寧に発声された音声を入力対象にしている。しかし、人間同士のコミュニケーションでは、「あのー」、「えーと」などに代表される冗長語や、言い淀みや言い誤りおよび言い直しなどが頻繁に見受けられる。このような音声でも認識が可能な、いわゆる自由発話の音声認識が、今後の重要な研究課題になると思われる。

現在、自由発話の認識に関する報告としては、日本では自由発話に対してスポッティングの方法を適用して認識率を報告した例がある [3]。また、海外では、自由発話のデータベースを作成して自由発話の特徴を報告し、従来の音声認識で使用されたアルゴリズムを利用して、認識率を報告した例がある [4],[5]。

ここでは、自由発話の音声認識の第一歩として、自由発話の音声の特徴を調べた。そこで自由発話と従来の朗読発話の差を見るために、まず発話速度と融合ラベルの付与率と従来の認識手段である HMM の音素認識性能を調査した。本論文では、これらの結果について報告する。

表 1: 調査に用いた自由発話の音声データの収録条件

発声者	アナウンサー 2 名 (通称 MTK,FKN)
収録環境	遮音室
発話内容	国際会議の申し込みに関する参加者と事務局の対話 「トピック」(質問項目と、その背景に関する情報) や「バックグラウンド」(会話の前提になる背景) を詳細に設定して対話したもの。
入力系	マイクロフォン、ＤＡＴ録音
データ量	26 対話 437 文 (MTK) 26 対話 569 文 (FKN)
音素数	約 8919 音素 (MTK) 約 14867 音素 (FKN)
発話様式	自由発話

2 調査に用いたデータベース

現在ＡＴＲでは、各種言語現象を調査するために対話文を中心とする言語データベースの作成を進めているが [6]、この目的のために録音された音声は、同時に、音声データベースとして利用できる。本来は、自由発話音声の収録は、話者に録音していることを気づかれずに録音することが好ましいが、通信の守秘義務などの問題の他にも、話題が次々に移行するため、使われる単語は膨大な種類になるという問題も発生する。このため、会話にある程度の制約を入れた模擬会話で収録することが現実的である。ATR では、役割を与えられた 2 名の話者により行なわれた会話音声を録音収録し、後に文字化して言語データベースを作成している [7]。現在、発話内容で 5 種類、収録環境で 2 種類、話者で 2 種類、発話様式で 2 種類の variety を含むデータベースを収集中である [6]。

今回の調査には、このデータベースのなかから、電話による国際会議の問い合わせに関する音声データを使用した。これの収録条件を表 1 に示す。音声データは、遮音室でアナウンサーが発声したもので、ドアの開閉音などの日常雑音や話者の舌打ちの音などは含まれていない。また両話者は、完全に分離されて録音されているため、音声区間の重畳はない。この意味で、この音声データは、自由発話音声としてはかなり clean な音声であると言ってよい。調査は個人差も調べるため、アナウンサー 2 名について行なった。ただし、両者の発話内容は異なっている。

3 自由発話の音響的な特徴

ここでは、自由発話の音響的な特徴を調べるために、発話速度と融合ラベルの付与率と HMM による音素認識率における、自由発話と朗読発声の差を調べた。

3.1 ラベリング作業からみた自由発話音声の崩れ

表 2 に、音声データのラベリングの作業において見受けられた、自由発話の音素の定性的な特徴を挙げる。なお、ラベリングの基準は文献 [8] にしたがった。これらから、自由発話においては音素の境界はかなり曖昧になっていることや、朗読発声には見られない音が現れることがわかる。

表 2: 自由発話の音素の定性的特徴

1	文の語尾の音素が不明瞭になる。 (例: 「なんですか」の「か」がほとんど聞こえない。)
2	母音/a,i,u,e,o/全てに無声化が存在する。 (朗読発声では/a,e,o/は、あまり無声化しない。)
3	2 重に解釈できる音素が存在する。 (例:「んー」(考え込むとき発声している音) は /N/あるいは/uN/の 両者に解釈できる。)
4	子音/r/をともなう音節の発音が全体的に弱い。 (例:「そうす る と」の「る」がほとんど聞こえない。)
5	母音 (特に、文末の母音『ａ』) の第 1 フォルマントがあらわれないことがある。 しかし、第 1 フォルマント以外の構造は明確で、 波形も母音の特徴を備えている。(話者 MTK においてのみ)

3.2 発話速度からみた自由発話音声

自由発話における発話速度の違いを調べるためにモーラ速度を調べた。同一話者における 4 種類の発話様式 (単語発声、文節単位の朗読発声、文単位の朗読発声、自由発話) における発話速度の差を調査した。これを図 2 に示す。なお、調査の際、自由発話では、文中に息つぎなどのために長いポーズ区間が存在した。これらの区間は、発話速度を計算するときに除去した、また発話内容は、朗読音声の文節単位と文単位は同一であるが、単語発声と自由発話の発声内容とは各々異なっている。この結果からわかることを示す。

1. 発話速度は単語発声・文節単位の朗読発声・文

単位の朗読発声の順に早くなる。

2. アナウンサ (MTK) は自由発話は文単位の朗読発声より早く発話されているが、 アナウンサ (FKN) では自由発話は文単位の朗読発声より遅くなっている。

3. アナウンサー 2 名とも自由発話における発話速度は、朗読発声における文発声と大きな差がない。

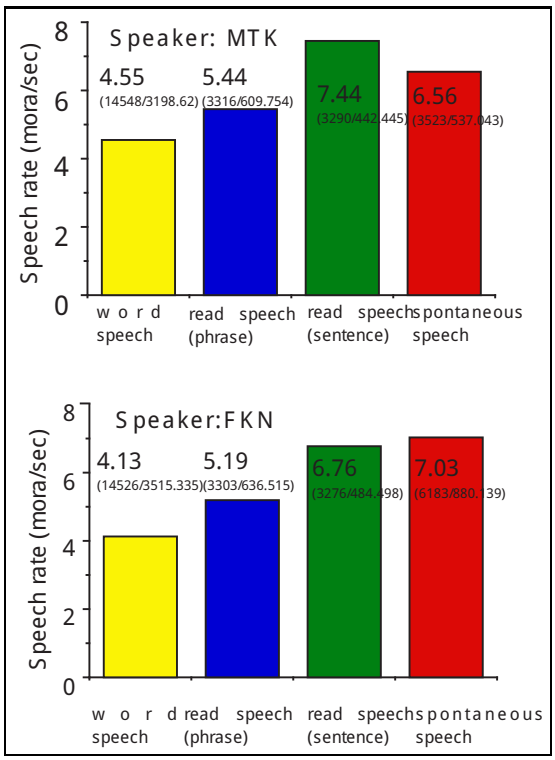

図 1: 発話様式の違いによる発話速度の変化

3.3 融合ラベルの付与率から見た自由発話の音声の崩れ

音声のラベリング作業において、発話テキストを参照しながら人手で音素境界を決定していくときに、音素境界が不明瞭な音素区間に対して付与するラベルのことを、融合ラベルと呼んでいる。

同一話者の 4 種類の発話様式における融合ラベルの付与率を調査した。これを図 2 に示す。この結果からわかることを示す。

1. 単語発声・文節単位の朗読発声・文単位の朗読発声・自由発話の順に融合ラベルの付与率が増加する。

2. 単語発声・文節単位の朗読発声・文単位の朗読発声では 2 人のアナウンサーの融合ラベルの付与率に差がない。

3. 自由発話では 2 人のアナウンサにおいて融合ラベルの付与率に差が見られる。つまり個人差が大きい。

4. 自由発話では、全音素の約 $1/5$ 以上が融合ラベルになる。

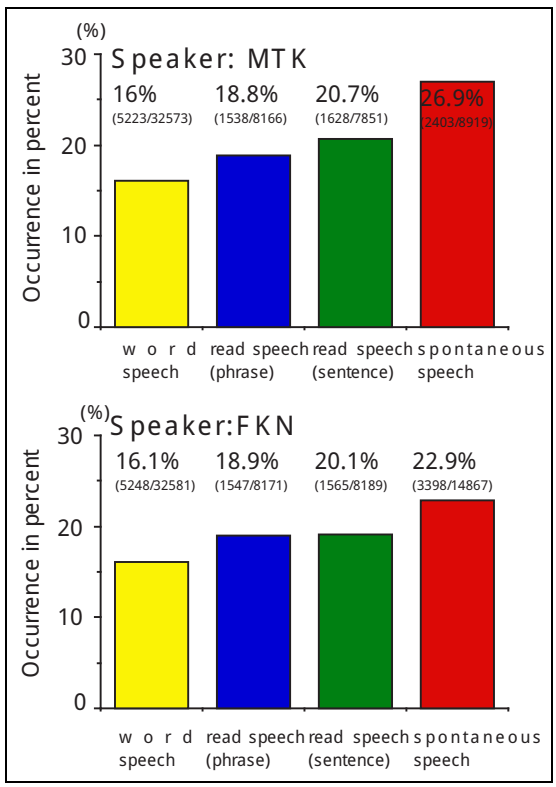

図 2: 発話様式の違いによる融合ラベルの付与率の変化

3.4 HMM による音素認識率から見た自由発話の音声の崩れ

従来の音声認識アルゴリズムにおける認識率の各発話様式の差を調べるために音素の認識実験を行なった。ただし、学習データとテストデータにおいて融合ラベルを付与された音素は実験から除いた。認識

アルゴリズムには、混合連続分布型 HMM を用い、学習データには、単語発声から視察によって切り出した音素を使用した。その他の実験条件を表 3 に示す。実験は、同一話者の単語発声と、文節単位の朗読発声と、文単位の朗読発声、および自由発話の、４種類のデータで行なった。

各発声様式における音素認識誤り率を図 3 に、各音素毎の音素認識率を表 4 に示す。この結果から次のような傾向が見られる。

1. 単語発声、文節単位の朗読発声、文単位の朗読発声、自由発話の順で音素認識率が低下している。
2. 自由発話と朗読発話の認識率の個人差がかなりあると思われる。アナウンサー MTK は自由発話の音素認識率が朗読発話の音素認識と比較して約 60%程度しか増加しないのに対し、アナウンサー FKN では約 140%も増加している。
3. 自由発話では、/i/や/g/の音素の認識率の低下が著しい。
4. /k/,/s/の認識率は、発話様式に、あまり差がない。

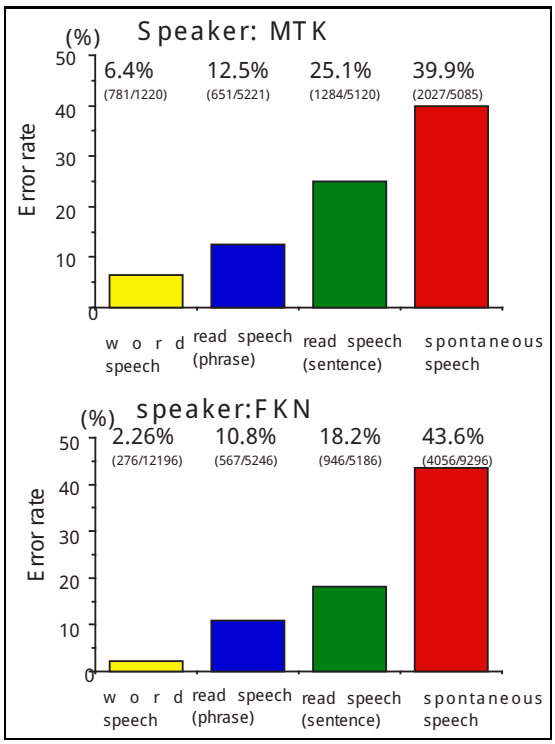

図 3: 発話様式の違いによる音素認識誤り率の変化

表 3: 音素認識の実験条件

認識対象	32 音素
サンプリング周波数	12kHz
話者	男性アナウンサー
学習データ	単語音声
音響パラメータ	log power + 16 次 LPCcepstrum + Δlog power + 16 次 Δcepstrum
フレーム窓長	20ms
フレーム周期	5ms
音素モデル	3state 10 mixture Gaussian continuous HMM

3.5 自由発話の音響的な特徴の考察

ここでは、自由発話と朗読発声の音響的な差を調べるために、主に発話速度および融合ラベルの付与率および HMM における音素認識率について調査した。その結果、自由発話は文単位の朗読発声と比較すると、発話速度に大きな差がないが、融合ラベルの出現頻度は約 2 割近く増加し、音素認識誤り率は平均で約 2 倍に増加した。

一方、文節単位の朗読発声と文単位の朗読発声を比較すると、融合ラベルの出現頻度に大きな差はないが、融合ラベルを除いた音素認識誤り率は約 2 倍近くにまで増えた。このような認識誤りの比率は、文節単位の朗読発声と文単位の朗読発声の比率と同等程度であると見ることもできる。また、約 40%という音素認識率は、見方によってはさほど低くないとも考えられる。したがって少なくとも音素モデルに関しては、自由発話音声認識と朗読認識において大きな差はないように思われる。

ただし、これらの値は話者によって差がある。特に文単位の朗読発声に対する自由発話の音素認識誤り率の増加率は、話者 MTK と FKN では大きな差

がある。したがって自由発話は、話者によって認識率に大きな差があることが予想される。

また、自由発話には言語的には言い直しや言い淀みなどが存在する。この解決方法として、精密な文法を作成することによる HMM-LR 法や、統計的な言語モデルや、認識アルゴリズムとしてスポッティングの採用も考えられるが、多くの問題点を含んでいると思われる。

4 まとめ

ここでは、自由発話の認識にむけて、自由発話と朗読発話の音響的な差を調べた。この結果、朗読音声と比較すると、自由発話の音声は、発話速度ではあまり差がないが、融合ラベルの付与率が約 1 割から 2 割に増加すること、また音素認識誤り率が約 2 倍ないし 3 倍に増加することがわかった。ただし、これらの値は話者によって差がある。しかし、融合ラベルを除いた音素認識率誤り率は 39.9%もしくは 43.6%と、著しく低い値ではない。また、融合ラベルの増加の問題は、連結学習により解消されることが期待できる。したがって少なくとも音素モデルに関しては、自由発話音声認識と朗読認識において大きな差はないように思われる。

ただし、ここで扱った自由発話は、言葉の対応に慣れた人たちが限定した条件の下で発話したデータであるため、これは clean な音声と言うべきである。したがって、一般の話者が、雑音下で制約の少ない状態で話した音声では、この論文で調査した結果と若干異なる可能性がある。

謝辞

研究で使用した自由発話の言語のデータベースの解析は、当時ＡＴＲ、現ＮＨＫの江原氏の指示のもとになされたものです。また、自由発話の音素ラベリングには、木田嬢を始めとする多くのラベラーの方々の協力によってなされました。また、ATR 自動翻訳電話研究所の森元室長や飯田室長の他、音声とデータと言語の各研究員に多くの協力をいただきました。これらの方々に厚くお礼を申しあげます。

参考文献

[1] Kai-Fu Lee, “Large-Vovabulary Speaker Independent Continuous Speech Recognitton: The SPHINX System,” Computer Science Department Carnegie Mellon University, Pittsburgh, Pennsylvania 15213 CMU-CS-88-148 (April 18, 1988).

[2] 永井 明人, 他, “HMM-LR 連続音声認識装置の開発と性能評価,” 日本音響学会平成 3 年度秋季研究発表会, 1-5-23 pp.45-46, (Oct. 1991).

[3] 坪井 宏之, 橋本 秀樹, 竹林 洋一, “連続音声理解のためのキーワードラティスの解析,” 日本音響学会平成 3 年度秋季研究発表会, 1-5-11, pp.21-22, (Oct. 1991).

[4] Victor Zue, James Glass, David Goodine, et al. “The MIT ATIS System: Preliminary Development, Spontaneous Speech Data Collection,And Performance Evaluation,” Eurospeech 91, pp.537-540, Genova, Italy, (Sep. 1991)

[5] Victor Zue, Nancy Daly, et al, “The Collection and Preliminay Analysis of a Spontaneous Speech Database,” DARPA Workshop 1989, pp.126-134 (1989).

[6] 江原 暉将, 小倉 健太郎, 森元 逞, “電話対話データベースの構築,” 第 40 回情報処理全国大会, pp.486-491 (1990).

[7] 篠崎 直子, 小倉 健太郎, 森元 逞, “言語データベース作成のためのシュミレーション会話,” 第 37 回情報処理全国大会, pp.1000-1001 (1988).

[8] 武田 一哉, 匂坂 芳典, 片桐 滋, 桑原 尚夫,“音韻ラベルの持つ日本語音声データベースの構築”SP87-19(1987).

[9] 有田 英一, 小暮 潔, 野垣内 出, 飯田 仁, “メディアに依存する会話の様式,” NL61-5 (1987).

[10] 山本 幹雄, 小林 聡, 中川 聖一, “音声対話文における助詞落ち・倒置の分析と解析手法,” 情報処理学会論文誌 Vol.11, No11, pp.1322-1330 (1992).

付録

表 4: 音素認識率 (MTK)
認識率% (正解音素数/調査音素数)

認識音素	単語発声	朗読発話 (文節)	朗読発話 (文)	自由発話
a	98.4%(1729/1758)	97.4%(855/878)	89.9%(779/867)	70.8%(625/883)
i	91.6%(1171/1278)	85.1%(470/552)	56.4%(295/523)	40.8%(208/510)
u	90.3%(1289/1428)	89.2%(222/249)	69.0%(165/239)	57.8%(114/198)
e	93.0%(625/672)	94.1%(380/404)	73.8%(296/401)	59.5%(308/518)
o	95.0%(818/861)	89.7%(496/553)	67.5%(366/542)	54.1%(282/521)
b	81.2%(168/207)	85.4%(35/41)	68.3%(28/41)	70.7%(29/41)
d	92.9%(158/170)	83.3%(175/210)	76.7%(161/210)	56.0%(140/250)
g	90.9%(230/253)	78.0%(92/118)	53.0%(62/117)	19.3%(21/109)
m	84.2%(406/482)	78.4%(149/190)	70.0%(133/190)	39.8%(72/181)
n	93.2%(247/265)	74.4%(206/277)	70.6%(192/272)	49.1%(121/247)
N	94.0%(409/435)	63.1%(53/84)	35.8%(29/81)	53.7%(29/54)
p	66.7%(10/15)	0.0%(0/2)	0.0%(0/2)	0.0%(0/4)
t	92.5%(407/440)	90.8%(228/251)	84.0%(210/250)	66.1%(164/248)
k	91.5%(1042/1139)	94.8%(440/464)	93.7%(432/461)	83.1%(280/337)
s	74.0%(399/539)	89.0%(121/136)	91.1%(123/135)	90.4%(226/250)
sh	79.1%(254/321)	95.2%(100/105)	99.0%(100/101)	74.4%(67/90)
h	95.8%(206/215)	48.2%(68/141)	36.4%(51/140)	38.3%(67/175)
ts	83.0%(151/182)	75.5%(37/49)	71.7%(33/46)	31.0%(13/42)
by	33.3%(1/3)	0.0%(0/0)	0.0%(0/0)	0.0%(0/0)
ch	73.6%(89/121)	81.1%(30/37)	48.6%(18/37)	36.4%(28/77)
f	86.7%(39/45)	33.3%(1/3)	66.7%(2/3)	25.0%(1/4)
gy	15.4%(2/13)	0.0%(0/1)	0.0%(0/1)	0.0%(0/0)
hy	20.0%(2/10)	0.0%(0/8)	0.0%(0/8)	0.0%(0/0)
j	86.9%(146/168)	96.6%(56/58)	93.1%(54/58)	94.0%(63/67)
ky	94.2%(49/52)	97.1%(33/34)	100.0%(32/32)	0.0%(0/16)
ny	0.0%(0/8)	0.0%(0/2)	0.0%(0/2)	0.0%(0/2)
r	97.8%(711/727)	91.1%(214/235)	90.6%(211/233)	92.4%(171/185)
ry	82.9%(34/41)	78.6%(11/14)	85.7%(12/14)	33.3%(2/6)
w	82.3%(65/79)	56.3%(9/16)	16.7%(2/12)	5.3%(1/19)
y	93.3%(166/178)	80.9%(72/89)	39.5%(32/81)	36.1%(13/36)
z	92.8%(103/111)	85.0%(17/20)	85.7%(18/21)	86.7%(13/15)
平均	91.1%(11126 /12216)	87.5% (4570/5221)	78.4%(4012/5120)	61.4% (3122/5085)

表 5: 音素認識率 (FKN)
認識率% (正解音素数/調査音素数)

認識音素	単語発声	朗読発話 (文節)	朗読発話 (文)	自由発話
a	99.8%(1756/1759)	95.7%(844/882)	90.8%(787/867)	59.7%(1032/1729)
i	99.4%(1269/1277)	85.8%(467/554)	69.3%(372/537)	38.4%(366/954)
u	99.1%(1412/1425)	89.9%(232/258)	79.4%(204/257)	85.5%(383/448)
e	99.7%(670/672)	96.0%(388/404)	87.9%(355/404)	29.9%(249/832)
o	100.0%(860/860)	85.7%(474/553)	75.8%(403/532)	50.9%(444/873)
b	96.6%(200/207)	85.4%(35/41)	87.8%(36/41)	70.6%(36/51)
d	93.5%(159/170)	83.8%(176/210)	72.9%(153/210)	45.5%(177/389)
g	90.1%(228/253)	89.8%(106/118)	72.6%(85/117)	45.5%(95/209)
m	95.6%(461/482)	93.2%(177/190)	94.2%(178/189)	83.7%(343/410)
n	93.6%(248/265)	79.1%(219/277)	64.2%(176/274)	25.8%(116/450)
N	98.4%(432/439)	75.9%(63/83)	71.1%(59/83)	71.6%(68/95)
p	46.7%(7/15)	0.0%(0/2)	0.0%(0/2)	12.5%(1/8)
t	97.7%(430/440)	98.4%(247/251)	97.2%(244/251)	73.3%(296/404)
k	95.9%(1090/1137)	95.7%(450/470)	96.4%(451/468)	94.6%(651/688)
s	99.8%(523/524)	99.3%(146/147)	91.7%(143/156)	75.0%(231/308)
sh	99.7%(314/315)	93.6%(102/109)	89.9%(98/109)	26.3%(54/205)
h	98.6%(212/215)	46.0%(64/139)	37.9%(53/140)	44.0%(125/284)
ts	98.9%(179/181)	98.0%(48/49)	95.7%(45/47)	69.5%(57/82)
by	0.0%(0/3)	0.0%(0/0)	0.0%(0/0)	0.0%(0/0)
ch	100.0%(121/121)	100.0%(37/37)	91.9%(34/37)	55.6%(55/99)
f	85.1%(40/47)	0.0%(0/3)	33.3%(1/3)	100.0%(11/11)
gy	38.5%(5/13)	0.0%(0/1)	0.0%(0/1)	0.0%(0/6)
hy	50.0%(5/10)	0.0%(0/6)	0.0%(0/7)	0.0%(0/0)
j	98.2%(165/168)	94.9%(56/59)	82.8%(48/58)	48.2%(55/114)
ky	98.1%(51/52)	100.0%(34/34)	97.1%(33/34)	90.4%(47/52)
my	0.0%(0/0)	0.0%(0/0)	0.0%(0/0)	0.0%(0/0)
ny	12.5%(1/8)	0.0%(0/2)	0.0%(0/2)	0.0%(0/4)
r	97.8%(711/727)	93.2%(219/235)	86.0%(202/235)	69.3%(287/414)
ry	85.4%(35/41)	71.4%(10/14)	57.1%(8/14)	0.0%(0/4)
w	81.0%(64/79)	33.3%(5/15)	8.3%(11/12)	3.2%(1/31)
y	93.3%(166/178)	72.3%(60/83)	56.8%(46/81)	32.0%(32/100)
z	93.8%(106/113)	100.0%(20/20)	83.3%(15/18)	66.7%(28/42)
平均	97.74% (11920/12196)	% 89.2% (4679/5246)	81.8%(4240/5186)	56.4%(5240/9296)