論文 / 著書情報
Article / Book Information

論題(和文)	日本語話し言葉コーパスを用いた異なるタスクに対する音声認識
Title(English)	
著者(和文)	西井 俊介, 篠崎 隆宏, 古井 貞熙
Authors(English)	Shunsuke Nishii, Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会2010年春季講演論文集, , No. 1-6-10, pp. 27-28
Citation(English)	, , No. 1-6-10, pp. 27-28
発行日 / Pub. date	2010, 3

日本語話し言葉コーパスを用いた異なるタスクに対する音声認識*

☆西井 俊介，篠崎 隆宏，古井 貞熙 (東工大)

1　はじめに

音声認識のタスクは多様であるが，各タスクに対してどのようなコーパスに基づき学習した音響モデルを用いて音声認識をするのが適当であるかは，従来あまり調べられてこなかった．そこで本研究では代表的なタスクとして，話し言葉，読み上げ音声，対話音声に対して，日本語話し言葉コーパス (CSJ)[1] 及び新聞記事読み上げ音声コーパス (JNAS)[2] に基づき学習した音響モデルを用いて音声認識を行い，性能評価を行った．その結果より CSJ を用いて学習したモデルを用いることで，JNAS モデルと比較して多くのタスクにおいて同程度またはより高い認識性能が得られることを示す．

2　音響モデル

音響モデルの学習用データセットとして以下のものを用いた．

- CSJ : 日本語話し言葉コーパス (CSJ) に含まれる学会講演音声 232.58 時間.（うち，男性話者 189.47 時間，女性話者 43.11 時間.）
- JNAS : 新聞記事読み上げコーパス (JNAS) に含まれる音声 52.36 時間.（うち，男性話者 25.06 時間，女性話者 27.30 時間.）
- CSJ+JNAS : 上記二つを組み合わせたもの，合計 284.94 時間.

学習した音響モデルは，特徴量 38 次元（MFCC 12 次元，その一次微分，二次微分，対数パワーの一次微分，二次微分），総状態数 3000 の triphone HMM で表現される．各状態の混合数は，予備実験の結果により，CSJ，CSJ+JNAS モデルは 64，JNAS モデルは 32 とした．学習方法は最尤法であり，いずれのモデルも性別非依存の不特定話者モデルである．

3　実験条件

以下の 4 + 10 種類の異なるタスクのテストセットを用意した．Table. 1 に各テストセットの発話時間を記す．

Table 1　Test set

テストセット		発話時間 (分)
SC		138.25
RJ		20.11
RS		182.73
DT		57.03
NC	NormalCityFemale	203.79
	NormalCityMale	210.25
	NormalIdleFemale	204.64
	NormalIdleMale	209.23
	ProCityFemale	42.41
	ProCityMale	41.10
	ProHighFemale	42.46
	ProHighMale	41.18
	ProIdleFemale	43.04
	ProIdleMale	41.35

SC : 学会講演音声．CSJ に含まれる標準テストセット 1.

RJ : 新聞記事読み上げ音声．JNAS に含まれる IPA-98-TestSet.

RS : 新聞記事読み上げ音声（高齢者）．S-JNAS[3] より無作為に抽出.

DT : 生駒市北コミュニティセンターに設置されている対話型音声情報案内システム「たけまるくん」[4] の入力音声．テストに使用した音声は，2002 年 11 月～2004 年 10 月の入力音声のうち比較的明瞭なものから無作為に抽出．子供の話者 46.78 分，大人男性話者 6.08 分，大人女性話者 4.17 分.

NC : 自動車運転行動中発話コーパス [5] に収録されているカーナビコマンド発声．話者の類別，走行環境によって 10 種類のテストセットに分けた．City，High，Idle はそれぞれ市街地走行中，高速走行中，停車中の音声を指す.

認識に用いた認識器は，NC 以外には Julius 4.1.2[6] を，NC には T^3 speech decoder[7] を用いた．言語モデルは，SC，DT には CSJ 学会講演 + 模擬講演に基づき学習した 3 万語彙の 3-gram，RJ，RS には毎日新聞社の 1991 年～2002 年の新聞記事データに基づき学習した 6 万語彙の 3-gram，NC にはカーナビコマンド用文法定

*Evaluation of various tasks using a recognition system based on the Corpus of Spontaneous Japanese. By Shunsuke Nishii, Takahiro Shinozaki, Sadaoki Furui (Tokyo Institute of Technology)

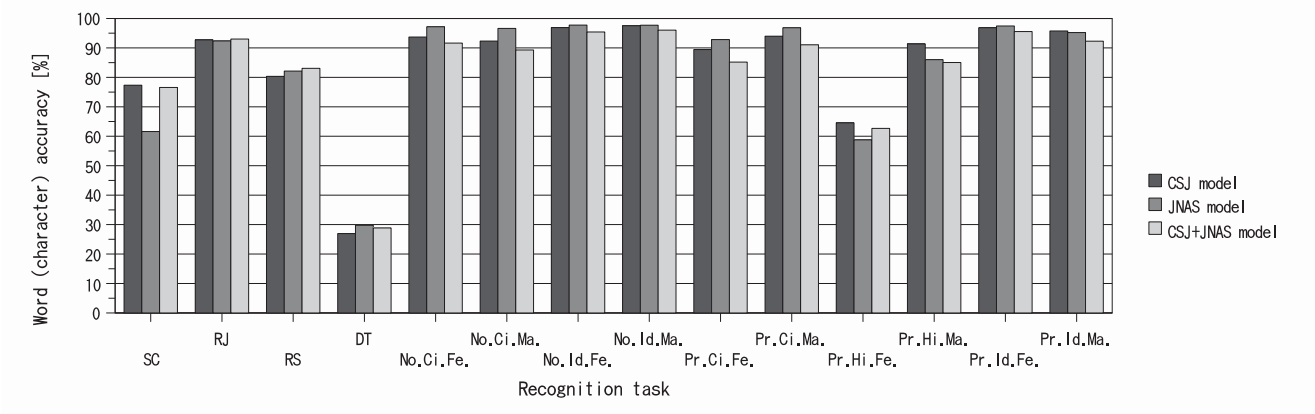

Fig. 1　Recognition results

義ファイルを用いた．認識時の言語重み，挿入ペナルティについては，予備実験より各条件ごとに定めた．

4　認識結果

各音響モデルを用いて，各テストセットに対する認識実験を行い，認識精度を評価した．Fig. 1 に結果を記す．値は単語認識精度 [%] である．ただし，RS，DT に対しては文字認識精度を用いた．

SC の認識結果より，話し言葉音声の認識では CSJ モデルが JNAS モデルよりも高い認識率を出している．これは先行研究の通りである [8]．RJ に対する認識率は，CSJ モデル，JNAS モデル間で同程度だった．RS，NC に対しては多くのテストセットで JNAS モデルが CSJ モデル よりも高い認識率を出しているが，話し言葉に対する認識率ほど大きな差はなかった．これは，読み上げ音声，ナビコマンド音声の認識では，話し言葉に基づいて学習した音響モデルを用いても十分な認識性能が得られることを示している．DT に対しては，どの音響モデルを用いても認識率が低かった．これには言語モデルにタスクに適したものを使っていないこと，及び，音響モデルの学習用音声の話者が大人のみで構成されているのに対し，評価用データセットには子供の話者が多く含まれることが理由として考えられる．CSJ モデル，JNAS モデルの比較としては，認識性能に大差はないと言える．なお，テストセットから子供の話者の音声を取り除いて認識すると，認識率は CSJ，JNAS，CSJ+JNAS の各モデルに対してそれぞれ 54.3，55.7，55.6 [%] となった．また，子供の話者の音声のみを CSJ モデルを用いて認識したところ，認識率は 20.1 [%] であったが，CSJ の女性話者のみからなるモデルを用いて認識したところ 27.1 [%] となり，認識性能が向上した．

今回用いたテストセットに対して，JNAS モデルを用いた認識性能が CSJ モデルを用いた場合に比べて大きく上回ることはなかった．また，CSJ+JNAS モデルを用いた認識について，CSJ モデルを用いた場合と大きな差は生じなかった．

5　まとめ

話し言葉コーパスに基づいて学習した音響モデルを用いて音声認識を行うことで，話し言葉以外のタスクとして読み上げ音声に対しても高い認識性能が得られることを示した．

謝辞　自動車運転行動中発話コーパスを使用させて頂いた旭化成（株）に感謝致します．

参考文献

[1] 篠崎 隆宏 他，情処学論 43(7)，pp. 2098-2107，2002-07.
[2] 板橋 秀一 他，音講論集，2-Q-36，pp. 187-188，1997-09.
[3] 馬場 朗 他，信学論 J85-D-II(3)，pp. 390-397，2002-03.
[4] 鹿野 清宏 他，情処学研報 SLP，pp. 33-38，2006-10.
[5] 加藤 智之 他，音講論集，3-Q-25，pp. 267-268，2007-09.
[6] 李 晃伸 他，信学技報 SP 107(406)，pp. 307-312，2007-12.
[7] Paul R. Dixon *et al.*, Proc. ASRU 2007, pp. 443-448, 2007-12.
[8] 篠崎 隆宏 他，音講論集，1-3-14，pp. 31-32，2001-03.