論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉音声認識における学習データ量と認識性能の関係
Title(English)	
著者(和文)	一場 知久, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2004年秋季講演論文集, Vol. , No. 2-1-9, pp. 53-54
Citation(English)	, Vol. , No. 2-1-9, pp. 53-54
発行日 / Pub. date	2004, 9

話し言葉音声認識における学習データ量と認識性能の関係 *

◎一場知久　岩野公司　古井貞熙 (東工大)

1　はじめに

話し言葉音声認識においては，未だ読み上げ音声認識ほどの認識率には達していないため，より高性能な言語，音響モデルの作成が求められている．また，音声認識において言語モデル，音響モデルの性能は学習データ量と密接な関連がある．本年度，話し言葉の大規模なデータベースである『日本語話し言葉コーパス』(CSJ) が完成し，公開された．そこで，本稿では CSJ を用いて話し言葉音声認識における学習データ量と言語モデル，音響モデルの性能の関係について報告し，CSJ の話し言葉音声認識に対する有効性を示す．また，使用する音響特徴量に MFCC の 2 次差分成分を含めることの効果についても述べる．

2　学習データと言語モデルの性能

まず，学習データ量と言語モデルの性能の関係を調べるため，使用する音響モデルを固定して学習データ量の異なる複数の言語モデルによる認識実験を行った．

2.1　言語モデルの作成

言語モデルの学習には CSJ の学会講演，模擬講演からテストセットを除いた 2671 講演，6.84M 形態素を用いた．CMU-Cambridge SLM Toolkit により，学習データ全体からその 1/8 までの 8 段階に学習データ量を変化させた．語彙は学習データに 4 回以上出現した形態素+発音形で構成し，back-off 平滑化に Witten-Bell 法を用いた．

各言語モデルの学習データに含まれる講演の種類，男女の比が異なると認識結果に影響する．そこで，学習データを学会講演男女，模擬講演男女の 4 つのグループに分け，各グループを 8 等分したものを組み合わせて学習データを構成した．また，言語モデルの場合，学習データ量が同じでも使用したテキストによって認識結果が変動することが考えられるので，学習データ量が 1/8 から 7/8 のグループに対しては同程度の学習データ量の複数の言語モデルを作成した．

学習データの内訳を表 1 に示す．データ量 (比) 1 から 7 までは複数の学習セットの平均である．

2.2　実験条件

デコーダには Julius-3.4 を使用した．評価データには CSJ のテストセット 30 講演を用いた．

音響モデルは音響特徴量として MFCC 12 次元，ΔMFCC 12 次元，Δpower 1 次元の計 25 次元を使用した，3000 状態 16 混合の triphone モデルである．音響モデルの学習データは CSJ の学会講演，模擬講演からテストセットを除いた 2670 講演，509 時間分を全て使用した．音響モデルの作成には HTK-3.2 を使用した．

表 1. 学習データの内訳

データ量 (比)	形態素数	講演数
1	0.86M	333.9
2	1.71M	667.8
3	2.56M	1000.7
4	3.42M	1380.5
5	4.28M	1670.0
6	5.13M	2003.5
7	5.99M	2336.0
8	6.84M	2671.0

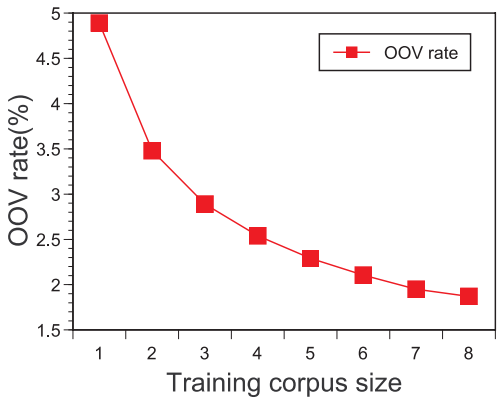

図 1. 学習データ量による未知語率の変化

2.3　実験結果

学習データ量に対する未知語率，単語誤り率およびパープレキシティの変化を図 1，2 に示す．

言語モデルにより語彙数が異なるので，パープレキシティの代わりに評価尺度として補正パープレキシティを用いた．学習データ量が 1/8 から 7/8 のグループの値は複数の言語モデルの平均値である．学習データ量を 8 倍に増やすことにより 5.6%単語誤り率が低下し，誤り削減率は 16.6%である．学習データ量の増加による単語誤り率の低下は，学習データ量が 7/8 から 1 でほぼ飽和している．また，未知語率，パープレキシティの低下と単語誤り率の低下がよく対応している．これらの結果から，これ以上学習データを増やしても認識性能の劇的な改善は望めないと考えられる．

すなわち，この結果は話し言葉講演音声認識のための言語モデルの学習データとして，CSJ がほぼ十分な量であることを示すものである．

∗ Relationships between trainig data size and recognition accuracy in spontaneous speech recognition
By Tomohisa Ichiba, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

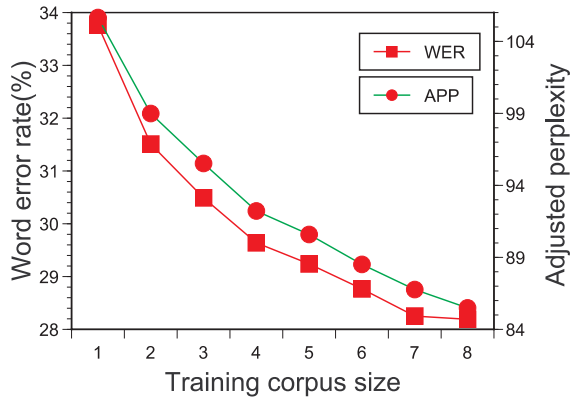

図 2. 学習データ量による単語誤り率，パープレキシティの変化 (言語モデル)

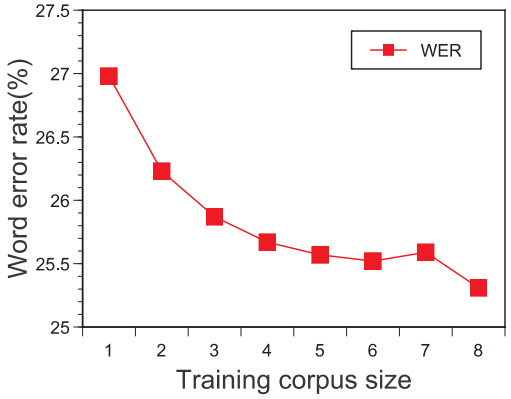

図 3. 学習データ量による単語誤り率の変化 (音響モデル)

3　学習データと音響モデルの性能

次に音響モデルについても同様の関係を調べるため，2 節と反対に言語モデルを固定して複数の音響モデルによる認識実験を行った．

3.1　音響モデルの作成

2 節と同様に，学習データ量を全体からその 1/8 まで 8 段階に変化させた複数の音響モデルを作成した．音響特徴量は MFCC 12 次元，ΔMFCC 12 次元，Δ power 1 次元，ΔΔMFCC 12 次元，ΔΔpower 1 次元の計 38 次元を使用した．使用した学習データ，状態数，混合数に関しては 2 節と同じである．また，言語モデルと違い，音響モデルにおいては学習に使用する音声の偏りによる影響は小さいと考えられるため，モデルの作成は各データ量に対し 1 つずつのみとした．

3.2　実験条件

言語モデルには，2 節で最も誤り率の低かった学習データ全体を使用した言語モデルを用いた．評価データ等は，2 節と同じ条件である．

3.3　実験結果

結果を図 3 に示す．学習データ量を 8 倍に増やすことにより 1.7%単語誤り率が低下し，誤り削減率は 6.3%である．言語モデルの場合と同様に，学習データ量の増加の効果が確認できるが，やはりデータ量が多くなるにつれて，誤り率の低下が小さくなっているのがわかる．音響モデルについても，2 節と同様に学習データとして CSJ がほぼ十分な量であると言える．また，誤り削減率の比較から言語モデルの方が学習データ量増加の効果は大きいことがわかる．

4　音響特徴量と認識性能の関係

25 次元の音響特徴量を使用した音響モデルと，ΔΔMFCC 12 次元，ΔΔpower 1 次元までを含めた 38 次元のモデルの比較を表 2 に示す．モデル 1，3 は学習データ全体，モデル 2 はその 1/8 のデータ量から作成した音響モデルである．言語モデルは共通で，学習データ全体から作成したモデルである．

モデル 1，3 の比較から，音響特徴量に 2 次差分成分を含めることにより 2.9%の認識性能の改善が得られたことがわかる．また，モデル 1，2 では学習データ量に 8 倍もの差があるにもかかわらず，モデル 2 の方が単語誤り率が低い．この結果から，音響モデルの学習に 2 次差分成分を含めることが認識性能の向上に非常に効果的であることがわかる．

表 2. 音響特徴量と認識性能の関係

モデル	音響特徴量	WER(%)
1	MFCC(25 次元，データ量 8)	28.2
2	MFCC(38 次元，データ量 1)	27.0
3	MFCC(38 次元，データ量 8)	25.3

5　まとめ

本稿では，日本語話し言葉コーパス (CSJ) を用いて，学習データ量と言語モデル，音響モデルの性能の関係について示し，CSJ の話し言葉講演音声認識の学習データとしての有効性を確認した．

また，使用する音響特徴量による音響モデルの性能についても比較を行い，MFCC の 2 次差分成分を含める効果を確認し，最終的に 25.3%の単語誤り率を得た．

2，3 節の結果から学習データの増加による認識性能の向上はほぼ限界に達していると考えられるため，今後，話し言葉音声認識においてはモデルの適応手法など他の方面からのアプローチが重要であると言える．

謝辞

2 次差分成分の有効性を指摘いただいた NTT の研究グループの方々に感謝致します．

参考文献

[1] 河原達也, “『日本語話し言葉コーパス』を用いた音声認識の進展,” 第 3 回話し言葉の科学と工学ワークショップ講演予稿集, pp.61-65, 2004.

[2] T.Kawahara, H.Nanjo, T.Shinozaki and S.Furui, “Benchmark test for speech recognition using the Corpus of Spontaneous Japanese,” Proc. SSPR2003, pp.135-138, 2003.