論文 / 著書情報
Article / Book Information

論題(和文)	対談音声を対象とした音声認識の検討
Title(English)	Investigation into Recognizing Interview Speech
著者(和文)	斎藤洋平, 古井 貞熙
Authors(English)	Yihei Saito, SADAOKI FURUI
出典(和文)	日本音響学会2001年春季講演論文集, Vol. , No. 1-3-17, pp. 37-38
Citation(English)	, Vol. , No. 1-3-17, pp. 37-38
発行日 / Pub. date	2001, 3

対談音声を対象とした音声認識の検討*

◎ 斎藤 洋平　古井 貞熙（東工大）

１．はじめに

NHK の TV 番組「クローズアップ現代」を用いて、対談音声の認識を試みた。97 年 7 月 29 日および 30 日に放送された音声をテストセットとして用いた[1,2]。内容はそれぞれ「新人官僚の研修」と「ヨットレース事故」で、いずれも、男女の対談部分および女性キャスターの単独発話部分を含んでいる。認識タスクとしては、対談部分の女性発話を用いた。また、比較対照用として、女性キャスターの単独部分の認識も行った。テストセットの大きさを表１に示す。

表１　テストセットのサイズ

Test Set	文数	形態素数
対談	19	494
キャスター	22	450

２．言語モデル

以下の言語モデルを作成した。各学習データの形態素解析には JTAG を用いた。

LM1：放送ニュース原稿５年分（92 年 7 月から 96 年 5 月まで）から作成。総学習文数 0.4M。語彙サイズ 20K。

LM2：「クローズアップ現代」書き起こしテキスト（97 年 6 月 2 日から 99 年 6 月 30 日放送分まで合計 159 日分（テストセットの 2 日間を除く）。総形態素数 0.8M。語彙サイズ 20K。

LM3：LM2 に品詞 N-gram を用いて未知語の追加を行った[3]。追加する未知語の選択は、ニュース原稿の書き起こしから、テストセットの内容に近いものを自動的に選び出し、その書き起こし中に含まれる単語の中で、頻度の上位 M 単語を選択した。その単語の品詞（主品詞および細品詞）情報に基づくクラス N-gram を用いて、未知語と既知語間の単語 N-gram を計算した。なお、本実験では M=50 とした。

LM4：テストセットと同日に放送された番組中の VTR の書き起こし（約 6.5K 形態素）を、学習テキストに重み付けして足し合わせたモデル。

３．音響モデル

音響モデルは、以下の３種類のモデルを用いた。

AM1：IPA の男女別不特定話者モデル。状態数 2000、混合数 16。

AM2：話者に適応化したモデル。対談の女声部分は、女性キャスターと同一人物であるため、女性キャスターの単独発話部分を用いて、AM1 をもとに ML 推定を行った。

また、対談音声認識の重要な問題点であるクロストーク部分に関しては、音響的な back-off（以下、AB）を行った[4]。すなわち、複数話者の発声が同時に行われている場合、その部分は音響スコアの計算を行わず、その話者に対するフレームごとの平均尤度を割り当てるという手法を用いた（クロストーク区間は、何らかの方法で検出できていると仮定した）。

４．実験結果

表２と３に、各言語モデルのパープレキシティを示す。ニュース原稿を用いて学習した LM1 に比べ、テストセットと同じ番組の書き起こしを用いることにより、パープレキシティを大幅に削減することができる。LM4 に関しては、AM1,2 それぞれにおいて、認識率が最も高くなる重みで VTR 書き起こしを加えたときの平均値を示す。

表２　Test Set Perplexity (Bigram)

Test Set	LM1	LM2	LM3	LM4
対談	539.99	69.16	82.24	94.61
キャスター	322.48	113.72	128.17	162.9

表３　Test Set Perplexity (Trigram)

Test Set	LM1	LM2	LM3	LM4
対談	510.79	56.56	66.42	85.53
キャスター	306.70	101.49	115.24	198.7

図 1 に、各言語モデルにおける未知語率を示す。LM1 と LM2 を比較すると、キャスター部分ではそれほどの違いは見られないが、対談部分に関しては、LM2 で未知語率が大幅に削減されている。これは、話し言葉独特の言い回しが LM2 で学習されることによる。また、関連ニュース原稿を用いて未知語を追加することにより、未知語率を大幅に削減できることが分かる（LM3）。また、放送された VTR を用いることにより、LM3 よりも容易に未知語率を削減できることが分かった（LM4）。

---

*Investigation into recognizing interview speech. By Yohei Saito and Sadaoki Furui(Tokyo Institute of Technology)

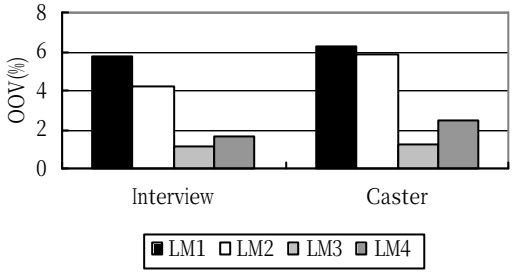

図１　未知語率

各言語モデル別の認識率を図 2 に示す。認識実験には Julius2.1 を用いた。音響モデルは AM1 を用いている。また、各音響モデルと音響 back-off を使用した場合の認識率を図 3 に示す。言語モデルは LM2 を用いている。この図中の「Interview」では、クロストーク区間だけでなく、対談の相手話者の(重なっていない)相槌などに対しても適用している。そこで、クロストーク部分のみに AB を適用した結果を「Crosstalk Only」に示す。クロストークを含むテストセットは対談 19 文中 9 文、クロストーク区間長は全体の約５%であった。これらの結果から、ここで検討した種々の方法の有効性が分かった。

また、LM4 においてVTR書き起こしに対する重みを変えたときの認識率の変化を図４に示す。言語モデル重みや挿入ペナルティは一定値を用いたが、音響モデルの違いによって認識率の飽和点が大きく異なることが分かった。

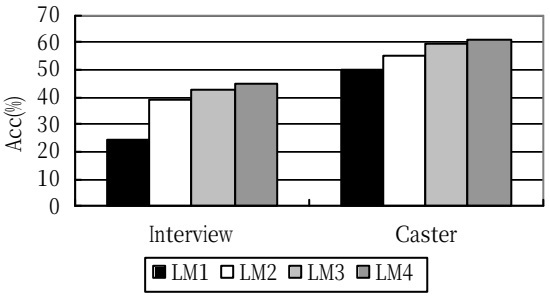

図２　言語モデルの性能評価

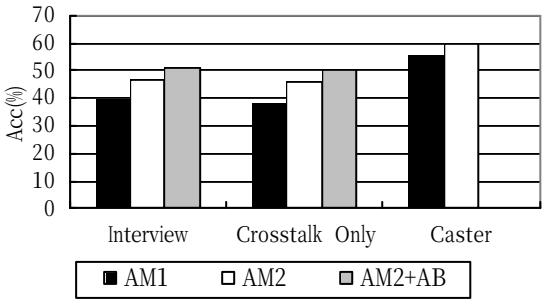

図３　音響モデルの性能評価

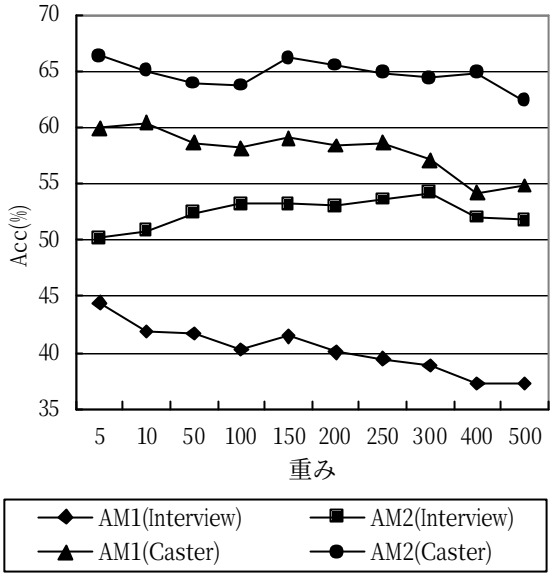

図４　重みによる認識率の変化

５．考察

話し言葉を忠実に書き起こしたデータから学習した言語モデルの有効性が確認できた。また、対談の話題に沿った単語を言語モデルに追加することにより、未知語率を削減することができた。また、少量のテキストでも、対談の話題に近いものを学習コーパスに足し合わせることにより、認識率を向上させることが分かった。

音響的には、音響モデルの話者適応や再学習により認識率を向上させることができた。また、対談の大きな問題点であるクロストークに関しても、音響的な back-off を適用することが有効であることがわかった。

謝辞

放送音声や書き起こしコーパスを提供いただいたNHK放送技術研究所、また、古井研究室の皆様の協力に感謝する。

参考文献

[1] 本間、今井、安藤：“対談番組を対象にした音声認識の検討”、春季音学講論、3-Q-24 (1999)

[2] 篠崎、斎藤、堀、古井：“話し言葉音声の認識を目指して”、信学技報、SP2000-96 (2000)

[3] 松井、加藤、小林、今井、田中、安藤：“ニュース音声認識における直前原稿を利用した認識性能の改善”、信学技報、SP99-128 (1999)

[4] J.de Veth,B.Cranen & L.Boves,”Acoustic backing-off in the local distance computation for robust automatic speech recognition”, Proc.ICSLP-98, Sydney, pp. 1427-1430 (1998)