自然言語処理 143－10
音声言語情報処理 36－5
（２００１．６．１）

講演スタイルの解説番組を対象にした音声認識の検討

本間真一† 小林彰夫† 佐藤庄衛† 今井亨† 安藤彰男†
宇津呂武仁‡ 中川聖一‡

†NHK 放送技術研究所
‡豊橋技術科学大学情報工学系

我々は、ニュース解説を対象にした音声認識の研究を行っている。これまでの研究では、解説音声は原稿読み上げ音声と異なる音響的特徴および言語的特徴をもつことや、学習データ量も不足していることから、まだ十分な認識精度は得られていない。そこで本稿では、比較的多くのデータ量が得られる講演スタイルの解説番組「あすを読む」を対象にした音声認識について検討を行う。ニュース原稿と「あすを読む」の書き起こしの混合による言語モデルの適応化、言語モデルの学習テキストと発音辞書におけるフィラーの扱いの見直し、音響モデルの話者適応などを行った結果、単語正解精度が 67.4%から 84.9 %まで改善した。

An Examination of Speech Recognition for Broadcast Commentary of Lecture Style

Shinichi HOMMA† Akio KOBAYASHI† Shoei SATO† Toru IMAI† Akio ANDO†
Takehito UTSURO‡ Seiich NAKAGAWA‡

†NHK Science and Technical Research Laboratories
‡Department of Information and Computer Sciences, Toyohashi University of Technology

We are studying speech recognition for news commentary. So far we haven't achieved satisfied accuracy for it, because speech of news commentary has different linguistic and acoustic features from read speech and supplies insufficient training data. Therefore, this paper treats speech recognition of a broadcast commentary program called “Asu wo Yomu (Reading Tomorrow)”, which has rather more training data. We adapted language models by mixing the news manuscripts and transcriptions of “Asu wo Yomu” in their training texts, changed how to treat pause fillers in the training texts and word lexicon, and carried out speaker adaptation of acoustic models and so on. As a result, we improved the word accuracy from 67.4% to 84.9%.

1．はじめに

近年、聴覚障害者や高齢者を中心に、生番組、特にニュース番組の字幕サービスの拡充を求める声が高まっている。NHK はこうした要望を受けて、2000 年 3 月 27 日より NHK 総合テレビ「ニュース 7」で字幕放送を試行的に開始した。現在のところ、この番組では、アナウンサーが原稿を読み上げる部分に限定して認識結果を人手で確認･修正することにより字幕を作成しているが[1]、「ニュース解説」に該当する項目において、認識精度が低下する傾向がみられており、いまその改善が求められている。

朗読音声と比較して、対話･対談等の自発音声には音響面、言語面の双方に性質の違いがあり、一

般にその認識性能は低い傾向にある[2]~[6]。文献[7]では、ニュース解説の音声においても、完全な朗読発話ではなく、一部において自発音声(spontaneous speech)に近い発話がみられるという特徴を捉え、これを考慮した音響モデルと言語モデルの改善を試みた。しかし、ニュース解説独自の学習データが不十分であることもあり、まだ十分な認識精度を得るまでには至っていない。そこで本稿では、比較的大量の解説スタイルの発話データが得られるテレビ番組「あすを読む」を取り上げ、これを対象にした音声認識について検討を行う。

具体的には、ニュース原稿と「あすを読む」の書き起こしの混合による言語モデルの適応化、言語モデルの学習テキストと発音辞書におけるフィラーの扱いの見直し、および、音響モデルの話者適応化などを行い、新たな言語モデルと音響モデルを作成する。そして、これらのモデルを用いて認識実験を行い、その効果の検証を行う。

2．解説番組「あすを読む」の特徴

「あすを読む」は、現在 NHK 総合テレビで放送されている 10 分間の解説番組である。番組の基本パターンは、1 名の解説委員が出演し、講演的なスタイルで、あるひとつの時事的なトピックを掘り下げる形式である。以下の①~④に、本番組の発話をニュース番組の原稿読み上げ発話と比較した場合にみられる顕著な相違点を示す。

① 解説スタイルの表現

台本となる原稿は用意されるが、完全な読み上げではない。口調はニュース原稿を読み上げる場合とは異なり、文献[7]で示したニュース解説の言語的特徴を含む。具体的には、フィラー(間投詞、言いよどみ、言い誤り)が頻出すること、図表を指示する表現が頻出すること、述部がていねいな表現や口語調の表現に変化すること、「~と思います」「~みます」などといった思考や意図を表す表現が頻出することなどが挙げられる。

② 話者がアナウンサーとは限らない

必ずしも十分に発声の訓練を受けた話者による発話であるとは限らない。話者によっては発声が不明瞭であったり、言いよどみや言い直しが多くみられたりする場合がある。

③ 放送日毎に話者が異なる

25 人の解説委員が在籍しており、放送日毎に話者が入れ替わる。

④ 放送日毎に話題が異なる

放送日毎に話題をひとつに特定できるが、その話題について一般のニュースよりも深く掘り下げて解説するため、専門的な用語や言い回しが多く見られる。

3． 言語モデルの作成と評価実験

3.1 学習コーパス

言語モデルの学習用として、ニュースの原稿と書き起こしより作成したコーパスと、「あすを読む」の書き起こしより作成したコーパスを用意した。それぞれのデータサイズを表 1 と表 2 に示す。

表 1 ニュースコーパス

総文章数	ニュース原稿　：1.6M ニュース書起し：31.8K
総単語数	ニュース原稿　：69.9M ニュース書起し：968.6K
ﾌｨﾗｰ単語数	ニュース原稿　：なし ニュース書起し：15.3K(1.6 %)

表 2 「あすを読む」コーパス

総文章数	16.7K
総単語数	461.1K
ﾌｨﾗｰ単語数	31.8K (6.9 %)

3.2 テストセット

テストセットの緒元を表 3 に示す。各話者のデータはすべて男声であり、それぞれ「あすを読む」一番組(10 分間)で発声されたすべての発話内容を使用した。なお本データは、表 2 の学習用コーパスに含まれておらず、学習コーパスに含まれる全データよりも後の日付に放送されたものを用いた。

表 3 テストセット

話者	文章数	単語数	ﾌｨﾗｰ単語数
A	86	1,952	63 (3.2 %)
B	84	1,984	120 (6.1 %)
C	55	1,709	183 (10.7 %)
D	73	1,635	48 (2.9 %)
E	50	1,942	265 (13.7 %)
合計	348	9,222	679 (7.4 %)

3.3 初期モデルの作成と評価

まずはじめに、表 1 のニュースコーパスより n-gram 言語モデル **LM-n** を作成し、表 2 の「あすを読む」コーパスより言語モデル **LM-a** を作成した。**LM-n** における cut-off 値は、bigram=1、trigram=2 とし、語彙サイズは 20K とした。**LM-a** においては、学習データの量が少ないことを考慮して cut-off は行わず、すべての語彙 17K を採用した。

表 3 のテストセット用いて **LM-n** と **LM-a** を評価した結果を表 4 に示す。これより、**LM-a** の方がパープレキシティー(PP)の値は小さいが、trigram のヒット率(HIT)が小さく、未知語率(OOV)にも改善の余地があることがわかった。

表 4 初期モデルの評価

言語モデル	PP	HIT	OOV
LM-n	179.2	54.3 %	3.9 %
LM-a	93.4	38.1 %	2.5 %

3.4 語彙の最適化

未知語率を小さくするために、表 5 の①～④に示す 4 通りの語彙の選定方法を比較した。その結果、④のニュースと「あすを読む」の単語とフィラーを組み合わせる方法が最も未知語率が小さくなることがわかった。この方法を定めるにあたり、「あすを読む」コーパスのフィラーの出現分布を調べたところ、言い直しや言いよどみを含めると、フィラーには数多くのバリエーションがあるが、その頻度の上位 12 種類で全フィラーの 9 割を占めることがわかった。そして、フィラーの種類を限定した語彙ファイルを作ったところ、「あすを読む」コーパスだけでは 20K の語彙に満たなかったため、残りの語彙にニュースコーパスの上位頻度の単語を加えて 20K のサイズにした。なお、これ以降に記述する言語モデルの作成には、すべてこの語彙データを用いる。

④の語彙を用いて、言語モデルを再構築して評価を行った結果を表 6 に示す。表 4 に対応する言語モデルの名称をそれぞれ **LM-N**、**LM-A** とし、ニュースと「あすを読む」の両コーパスを足し合わせて作った言語モデルの名称を **LM-N+A** とする。なお、**LM-N+A** における cut-off 値は、bigram=1、trigram=2 とした。いずれの言語モデルを用いた場合も、表 4 の **LM-a** と比べて未知語率と trigram のヒット率は改善されるが、若干パープレキシティーは大きくなる結果となった。これは、未知語が減った分、出現頻度が少ない単語がパープレキシティーの算出に加えられるようになったためと考えられる。

表 5 語彙の選定方法と未知語率

語彙の選定方法	OOV
①ﾆｭｰｽｺｰﾊﾟｽの単語 頻度上位 20K	3.9 %
②「あすを読む」ｺｰﾊﾟｽの単語すべて 17K	2.5 %
③ﾆｭｰｽｺｰﾊﾟｽ+「あすを読む」ｺｰﾊﾟｽの単語 頻度上位 20K	2.4 %
④フィラー頻度上位 12 単語 +「あすを読む」ｺｰﾊﾟｽの単語すべて 17K +ﾆｭｰｽｺｰﾊﾟｽの頻度上位の単語=計 20K	1.8 %

表 6 言語モデルの評価 ～ 語彙の最適化

言語モデル	PP	HIT	OOV
LM-N	226.3	51.9 %	1.8 %
LM-A	100.1	43.6 %	
LM-N+A	134.1	57.2 %	

3.5 「あすを読む」重みづけモデル

「あすを読む」コーパスの方が解説の特徴を多く含んでいることに着目して、ニュースコーパスに対し、「あすを読む」コーパスのテキストを *n* 倍に重みづけをして足し合わせた言語モデルを作成した。その言語モデルの名称を **LM-N+*n*A** とし、*n=10, 100, 1000* としたときの評価結果を表 7 に示す。**LM-N+A** と比較して、*n=10* のときにパープレキシティーが若干小さくなるが、さらに *n* を大きくするとパープレキシティーは大きくなった。

表 7 言語モデルの評価～「あすを読む」重み付け

言語モデル	PP	HIT	OOV
LM-N+*10*A	129.3	62.2 %	1.8 %
LM-N+*100*A	153.8	62.2 %	
LM-N+*1000*A	303.9	62.2 %	

3.6 フィラーの扱いの見直し

今回使用したコーパスは、フィラーの表記が人手で行われたものであるため、母音と長母音(例:「あ」と「あー」)の区別の仕方があいまいとなっていた。また、人手でフィラーであると分類された「この」に着目すると、フィラーとも指示語とも解釈できるケースがみられた。以上の二点を考

慮に入れて、フィラーの末尾の長母音化した音素を母音に修正し、フィラーと分類された「この」をフィラーとして扱わないことにした。このように変換したニュースコーパスと「あすを読む」のテキストを足し合わせて作成した言語モデル **LM-N'+A'**、および、「あすを読む」のテキストに 10 倍、100 倍の重みをつけて足し合わせて作成した言語モデル **LM-N'+*10*A'**、**LM-N'+*100*A'**の評価結果を表 8 に示す。表 6、7 の結果と比較して、パープレキシティーが小さくなり、trigram のヒット率が向上した。

表 8 言語モデルの評価 ～ フィラーの見直し

言語モデル	PP	HIT	OOV
LM-N'+A'	118.7	59.3 %	1.8%
LM-N'+*10*A'	111.3	64.0 %	
LM-N'+*100*A'	127.1	64.0 %	

3.7 認識実験

前述の語彙を用いた発音辞書と各種言語モデル、表 9 に示す音響モデル、および、文献[8]に示すニュース音声認識システムのデコーダを用いて認識実験を行った。なお、**LM-N'+A'**、**LM-N'+*10*A'**、および、**LM-N'+*100*A'**による認識の際には、発音辞書において、フィラーが含む母音表記の発音に長母音の発音を併記することにより、フィラーの母音と長母音の識別のあいまいさに対処した(例:「あ」の発音 = /a/ or /a:/)。また、テストセットは、表 3 のものから雑音を含んでいた 7 文を除外したものを用いた。

表 9 音響モデルの緒元

ｻﾝﾌﾟﾘﾝｸﾞ周波数	16kHz
分析窓	ハミング窓 25ms
フレーム周期	10ms
分析パラメータ	12 次元 MFCC＋対数ﾊﾟﾜｰ 各々の 1 次,2 次回帰係数 計 39 次元
HMM	状態共有化 8 混合分布 triphone
状態共有化	tree-based ｸﾗｽﾀﾘﾝｸﾞ
triphone ﾓﾃﾞﾙ数 状態数	6.0 K 4.7 K
学習データ	ニュース音声(「あすを読む」の話者とは異なる) 228K 文 605 時間

各言語モデルと単語正解精度(ACC)の関係を表 10 に示す。**LM-N'+*100*A'**を用いたときにもっともよい認識率が得られた。

なお、本稿の単語正解精度は、すべてフィラーを除外して算出した。こうした理由は、放送においては音声認識出力のフィラーを削除し、字幕化しないことを考慮に入れたためである。

表 10 言語モデルと単語正解精度

言語モデル	ACC
LM-N	67.4 %
LM-A	68.0 %
LM-N+A	68.9 %
LM-N+*10*A	71.0 %
LM-N+*100*A	72.5 %
LM-N+*1000*A	71.2 %
LM-N'+A'	70.0 %
LM-N'+*10*A'	71.6 %
LM-N'+*100*A'	73.5 %

4. 音響モデルの作成と評価実験

4.1 話者適応

「あすを読む」は、25 名の解説委員が持ち回りで各日の番組に出演するため、過去の放送に、テストセットの話者と同一話者が発話した音声データが存在する。これを利用して、MLLR[9]と MAP 推定[10]により話者適応を行った。

4.2 認識実験

デコーダとテストセットは、3.7 と同じものを使用した。本テストセットは、5 日分の番組を利用しており、表 3 に示す 5 名の話者を含んでいる。なお、言語モデルは、**LM-N'+*100*A'**を使用した。

実験結果を表 11 に示す。適応化用の音声データに、テストセットと同一の話者が過去に出演した番組の音声データを 1 番組分(10 分間)ずつ増やしていった。その結果、認識率は徐々に改善され、適応前は 73.5%であった単語正解精度が、3 番組分の音声で適応化を行った時点で 84.4%まで上昇した。

表 11 話者適応と単語正解精度

適応化データの量	ACC
1 番組 (10 分)	82.7%
2 番組 (20 分)	84.0%
3 番組 (30 分)	84.4%

4.3 話者毎のばらつきに関する考察

話者適応の効果を話者別にみた場合の比較を図 1 に示す。話者毎に認識率とその改善効果を比較してみると、かなりばらつきがあることがわかった。この理由について以下に考察する。

表 12 に、話者適応前、および、3 番組分のデータで適応後の単語正解精度(ACC)と、単語誤り削減率(改善率: Improvement)を示す。また表 13 に、話者毎のパープレキシティー(PP)と、trigram のヒット率(HIT)、および、未知語率(OOV)を示す。

話者 A と話者 B において認識率が低いのは、パープレキシティーが大きく、言語的に複雑であったためと考えられる。

話者 C において改善率が大きいのは、話者適応前の音響モデルがミスマッチであったために適応の効果が大きく得られ、また、パープレキシティーが比較的小さく、言語的に容易であるために高い認識率が得られたものと考えられる。

話者 D は、発声が明瞭であり、かつ、表 3 からもわかるようにフィラーも少ない。このため、話者適応を行わなくても高い認識率が得られていたものと考えられる。改善率が小さい理由としては、未知語率が大きい点が挙げられる。

話者 E の認識率が最もよい理由には、trigram のヒット率が最大であることが挙げられる。trigram のヒット率と単語正解精度の関係をプロットしたところ、話者適応後は図 2 に示すように両者の相関が高いことがわかった。

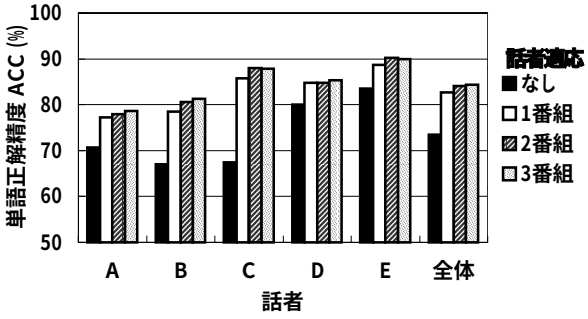

図 1 話者別の単語正解精度

表 12 話者別の単語正解精度と単語誤り削減率

話者	ACC		Improvement
	適応前	適応後	
A	70.7 %	78.7 %	27.3 %
B	67.1 %	81.2 %	43.1 %
C	67.5 %	87.9 %	62.7 %
D	80.1 %	85.4 %	26.4 %
E	83.5 %	89.9 %	39.0 %
全体	73.5 %	84.4 %	41.0 %

表 13 話者別にみた言語モデルの評価

話者	PP	HIT	OOV
A	144.2	60.1 %	2.3 %
B	144.5	61.5 %	1.7 %
C	112.1	66.2 %	1.2 %
D	108.0	64.8 %	2.6 %
E	125.9	67.6 %	1.2 %
全体	127.1	64.0 %	1.8 %

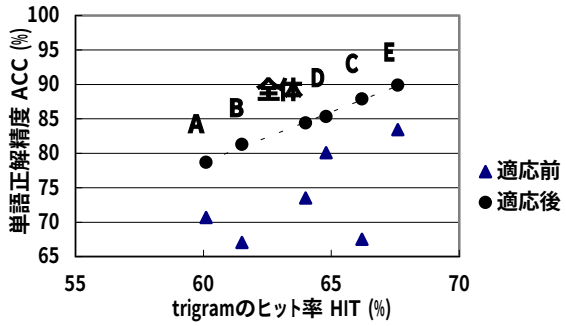

図 2 trigram のヒット率と単語正解精度

5．フィラーの影響と透過単語化

表 1 と表 2 のフィラー数を比較してわかる通り、「あすを読む」の発話は、ニュースに比べてフィラーの出現頻度が高い。このフィラーによる誤認識への影響を調べるため、フィラーの前後 2 単語を抽出して、その認識率の調査を行った。なお、言語モデルは **LM-N'+*100*A'**、音響モデルは話者適応後のものを使用した。その結果、フィラーの前後 2 単語の単語正解精度は 83.4%であった。表 11 によると、テストセット全体の単語正解精度は 84.4%である。よって、フィラーの前後では 1%劣化しているに過ぎなかった。これより、フィラー以外の単語と比較して、フィラーの出現が認識率に大きく悪影響を及ぼしているとは言えないことがわかった。

次に、フィラーによって言語制約が弱められないようにするために、フィラーの予測はするが単語履歴には含めない、いわゆる透過単語化の処理

を行った[11]〜[12]。単語列 $w_1\ w_2\ w_{trans}\ w_3$ が与えられると、通常 w_3 の生起確率は $P(w_3|\ w_2 w_{trans})$ と推定するのに対して、w_{trans} を透過単語とする場合は、$P(w_3|w_1 w_2)$ と推定することで、透過単語 w_{trans} による w_2 と w_3 の単語連鎖の分断を回避する。

まず、言語モデルに **LM-N'+*100*A'** を用い、デコーダのみで透過処理語の考慮して認識実験を行った結果、テストセット全体の単語正解精度は 84.9%となった。このときフィラーの前後 2 単語の単語正解精度は 85.7%となり、若干認識率が向上した。つづいて、フィラーを透過単語として言語モデルを再構築し、認識実験を行ったところ、テストセット全体の単語正解精度は 84.5%となった。このときのフィラーの前後 2 単語の単語正解精度は 86.1%となり、フィラー周辺の認識率は向上したと言えるが、全体としては透過語処理を行わない場合と比べてほとんど違いがなかった。

6．まとめ

講演スタイルの解説番組「あすを読む」の音声認識の検討を行った。

言語モデルの学習テキストとして、ニュース原稿と書き起こしからなるコーパスと、「あすを読む」の書き起こしからなるコーパスを用意した。語彙として、出現頻度が高いフィラーの上位 12 単語と、「あすを読む」コーパスに含まれる単語すべて、および、ニュースコーパスの上位頻度の単語により 20K のサイズとして、未知語率を削減した。また、ニュースのテキストに対して、「あすを読む」のテキストに重みをつけて足し合わせた言語モデルを作成し、パープレキシティーと trigram のヒット率を改善した。また、学習テキストにおけるフィラーの表記のゆらぎに対処するため、フィラーに含まれる長母音の表記を母音に統一し、発音辞書で母音と長母音の両方の発音を許容するようにしたことにより、認識率を改善した。

つづいて、音響モデルの話者適応を行い、適応前と比較して、単語誤り率を 41%改善した。また、話者毎の認識率やその改善率の違いについて考察した。

最後に、フィラーを透過単語として扱うデコーダと言語モデルを用いて認識実験を行ったところ、さらに若干の認識率を改善した。

今後は、話題選択したニュース原稿を用いた言語モデルの適応化などを行いたい。そして、本研究の成果をニュース解説の音声認識の改善に役立てていきたい。

謝辞

本研究に協力していただいた豊橋技術科学大学の学生、小玉康弘君、田中敬志君に感謝いたします。

参考文献

[1] 安藤, “ニュース音声自動字幕化システム”信学技報 SP2000-102 (2000.12) pp.43-48
[2] 村上 嵯峨山, “自由発話音声における音響的な特徴の検討”信学論 Vol.J78-D-Ⅱ No.12 (1995.12) pp.1741-1749
[3] 本間 今井 安藤, “対談番組を対象にした音声認識の検討”音講論集 3-Q-24 (1999.3) pp.165-166
[4] 山本 中川, “発話スタイルよる話速・音韻間距離・ゆう度の違いと音声認識性能の関係”信学論 D-Ⅱ Vol. J 83-DⅡ No.11 (2000.11) pp.2438-2447
[5] 三村 河原, “ディクテーションと対話音声における音響モデルの差異”音講論集 2-8-4 (2000.3) pp.35-36
[6] 尾上 世木 佐藤 今井 田中 安藤, “ニュース番組における認識率変動要因の検討”音講論集 2-8-15 (2000.3) pp.57-58
[7] 本間 小林 今井 田中 安藤, “ニュース解説を対象にした音声認識の検討”信学技報 SP2000-99 (2000.12) pp.25-30
[8] 今井 小林 尾上 安藤,“ニュース番組自動字幕化のための音声認識システム”音声言語情報処理研究会 23-11(1998.10) pp59-64
[9] C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, Vol.9, (1995.9) pp.171-185
[10] J.L Gauvian, C.H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, IEEE Trans. S.A.P. Vol.2,No.2 pp291-298 (1994)
[11] 西村 伊東, “講義コーパスを用いた自由発話の大語彙音声認識”, 信学論 Vol.J83-DⅡ No.11 (2000-11) pp.2473-2480
[12] 加藤 河原, “講演音声認識のための言語モデルの検討”, 音講論集 1-3-2 (2001.3) pp.27-28