音声言語情報処理　34－25
(2000. 12. 22)

ニュース解説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

NHK 放送技術研究所

〒：157-8510 東京都世田谷区砧１－１０－１１

TEL:０３－５４９４－２３５０

E-mail：{ honma, akio, shoe, imai, ando }@strl.nhk.or.jp

あらまし　NHK は平成 12 年度より、音声認識を利用した「ニュース 7」の字幕放送を試行的に開始した。現在のところ、アナウンサーが原稿を読む部分に限定して字幕を作成しているが、「ニュース解説」に該当する音声は、そのほかの原稿が読まれる部分の音声とは性質が異なるために、認識精度が低下する傾向がみられる。そこで本稿では、「ニュース解説」の言語的な特徴と音響的な特徴を分析し、これらの特徴を考慮した認識率の改善を試みる。具体的には、記者原稿からルールにより生成した解説口調の単語列を用いた言語モデルの適応化、大量のニュース番組の書き起こしコーパスを併用した言語モデルの学習、話者適応、および、発声継続時間が短い音素に対処するための音響モデルの構造の変更を行った。これらの結果、単語正解精度が 5.1%向上した。

キーワード　ニュース音声 ニュース解説 音声認識 言語モデル 音響モデル

An Examination of Speech Recognition for News Commentary

Shinichi HOMMA　Akio KOBAYASHI　Shoei SATO　Toru IMAI　Akio ANDO

NHK Science & Technical Research Laboratories

1-10-11 Kinuta Setagaya-Ku Tokyo 157-8510 Japan

Tel:03-5494-2363

E-mail：{ honma, akio, shoe, imai, ando }@strl.nhk.or.jp

Abstract　The “speech in a news commentary” has some linguistic and acoustic features which are different from those of read speech. In this paper, we applied some rules, which represented the linguistic features of news commentary, to news manuscripts, and generated word sequences, which was then used for language model adaptation. We also added a large amount of transcriptions of the news programs to this training data. On the other hand, we found the speech rate of news commentary speech was sometimes much faster than that of read speech. Thus, we changed the structure of acoustic models with speaker adaptation so as to recognize phonemes whose durations were relatively short. As a result, we improved the word accuracy of 5.1%.

key words　broadcast news, news commentary, speech recognition, language model, acoustic model

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む部分に限定して認識結果を人手で確認・修正[1]することにより字幕を作成しているが、「ニュース解説」に該当する項目は、通常の原稿を読む音声と比べて性質に違いがあり、認識精度が低下する傾向がみられるため[2]、認識装置の性能改善が求められている。

本稿で扱う「ニュース解説」とは、「ニュース7」などの番組中で重要な項目や難解な項目を図表や模型などを用いてわかりやすく説明する箇所のことを指す。記者などが出演して対談形式となることもあるが、ここではアナウンサーが単独で発声する箇所だけを扱う。ニュース解説の放送直前に、記者が作成した原稿などを参考にしてアナウンサー自身が台本となる原稿(読み原稿)を作成するが、実際の放送においてこの読み原稿が一字一句忠実に読まれるケースは少ない。したがって、完全な朗読発話ではなく、一部において自発的(spontaneous)に近い傾向の発話が含まれるという特徴がある。これまでの研究においても示されている通り、朗読音声と対話・対談等の自発音声には性質の違いがあり、その認識性能には限界がある[3]～[7]。ニュース解説の音声についてもこれに類似した原因によって認識性能が劣化している可能性があるが、その性質を扱った研究はまだ十分になされてない。

本研究では、まず、アナウンサーが読む原稿と実際の発話との比較からニュース解説の言語的な特徴を調査し、その傾向の分類を行う。つづいて、音響的な特徴を調査した結果について述べる。そして、これらの特徴を考慮に入れた上で、新たな言語モデルと音響モデルを作成する。さらに、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　ニュース解説の言語的特徴

2000年3月27日から6月12日までの間に「ニュース7」で放送された12項目のニュース解説の読み原稿を入手した。そして、この読み原稿と実際の発話(書き起こし)を比較して、その相違からニュース解説の発話の言語的特徴を調査した。その結果、読み原稿の通り忠実に読まれた文の数の割合はわずか13%であり、他は同意の別表現に置き換えられたか、または、その場に応じて適時付加された表現(アドリブ)であることがわかった。

以下の①～⑨に、この比較によって得られたニュース解説の言語的な特徴の分類結果を示す。なお、今回得られた読み原稿は112文2,927単語、対応する書き起こしは139文3,140単語である。

① 不要語の頻出

原稿読み上げ部分の発話と比べて、不要語が多く出現する傾向がみられる。なお、本調査の結果では、32%の文でその文頭において間投詞がみられ、このうち出現頻度の91%を「で」「え」「えー」が占めていた。文頭以外にも文節の区切れにおいて間投詞が多く出現する傾向があり、その中で特徴的なものとして、助詞「が」の直後の間投詞「あー」、助詞「に」の直後の間投詞「いー」といったような直前単語がもつ最終の母音を引き伸ばす性質をもつ間投詞がみられた。このほか、言いよどみや言い直しに分類される不要語の頻度も多い傾向がみられた。

② 指示のための表現

図表や模型を指し示すための表現が多くみられる。この表現は読み原稿に書かれていない場合が多く、そのときはアドリブによる発声となる。これらはさらに、連体詞(例「この」「こういう」)として現れるものと、代名詞を伴う句あるいは文(例:「こちらで説明しましょう」)として現れるものに分類できる。

③ 「～と思います」「～てみます」「～てみたいと思います」

上記の言い回しが、事柄を説明する前置きの中で現れる。いくつかのバリエーションがみられ、この表現(例「～を振り返ってみます」)が読み原稿に書かれている場合であっても、同意の別表現(例「～を振り返ってみたいと思います」)に置き換えられて発話されることがある。

④ 口語特有の単語

「ずうっと」「ちょっと」など、一般には口語表現でしか用いられない単語が現れる。

⑤ ていねい表現

読み原稿の表現が、ていねいな表現に言い換えられて発話されることがある。例えば、読み原稿で「する」と書かれている表現が「します」に、「いる」と書かれている表現が「います」「おる」「いらっしゃる」などという表現に変化する。

⑥ 「～ですね」

主に文節の区切れや文末において「ですね」という表現が挿入される。

⑦ 「～ですが」「～ですけれども」

読み原稿で「～ですが」と書かれている表現が、「～ですけれども」という表現に変化する。また、主として強調することばの直後に上記の言い回しが挿入されることがある。(例「地割れについてですが」「地震の回数ですけれども」)

⑧ 「～んです」

例えば、読み原稿で「～する」「～しました」「～ですが」と書かれている表現が、それぞれ「～するんです」「～したんです」「～なんですが」などという表現に変化することがある。

⑨ 「～わけです」「～ということになります」

読み原稿に書かれていないにもかかわらず、文末に上記の言い回しが付加されることがある。

3 ニュース解説の音響的特徴

ニュース解説の発話と原稿読み上げの発話の音響的性質を比較するため、それぞれの同一話者が発声したテストセットを用意し、音響モデルにテストセットの正解文を与えて音素のアライメントをとった。表1と表2に、ここで使用したテストセット、および、音響モデルの緒元をそれぞれ示す。

表1 アライメントをとったテストセットの緒元

分類	放送日	文数	単語数
ニュース解説	'00.3.27～'00.4.28	121	2,512
原稿読み上げ	'00.3.28	45	1,718

表2 アライメントに用いた音響モデルの緒元

サンプリング周波数	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＋対数パワー それぞれの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 クラスタリング
triphone モデル数 状態数	5648 4016
学習データ	102,877文 630時間

なお、HMMには、図1に示す3状態のleft-to-rightモデルを用いた。

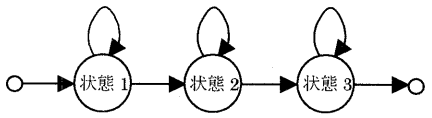
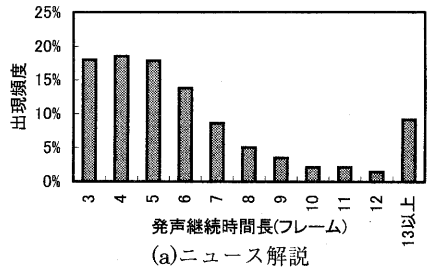


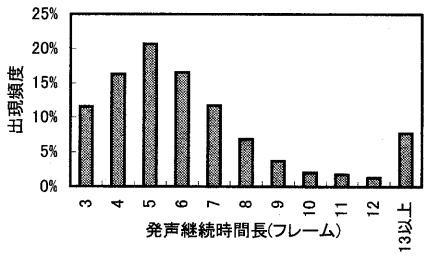
図1 HMMの構造

以下に、アライメントの分析結果を示す。

ニュース解説のアライメントを調査した結果、母音(/a/,/i/,/u/,/e/,/o/)、撥音(/N/)、半母音(/w/,/r/,/y/)、鼻子音(/m/,/n/)の発声継続時間長に、3～5フレーム(フレーム周期=10ms)程度の短いものの頻度が大きい傾向がみられた。そして、これらの音素のうち特に出現頻度の大きい母音と撥音に着目して、ニュース解説と原稿読み上げの発声継続時間長と出現頻度の分布をそれぞれ調べたところ、図2に示す結果が得られた。この図より、ニュース解説には、3～4フレームの発声継続時間が短いものの割合が多いことがわかる。さらに詳細を調べたところ、ニュース解説の母音/e/と/o/において、3フレームの出現頻度が最大となる特徴がみられた。一方、母音/u/に関しては、ニュース解説と原稿読み上げの間に明白な分布の違いはみられなかった。



(a)ニュース解説



(b)原稿読み上げ

図2 母音・撥音の発声継続時間長の分布

表3に母音・撥音の発声継続時間、および、音響尤度(フレーム平均)の比較を示す。両者の発声継続時間を平均値でみると差がみられないが、標準偏差の違いからニュース解説の発話の方がそのばらつきが大きいことがわかる。音響尤度については、ニュース解説の方が値が小さいことことから、音響モデルがあまりよく適合していないことがわかる。

表3 発声継続時間と音響尤度(フレーム平均)の比較

	ニュース解説	原稿読み上げ
発声継続時間の平均	6.7 フレーム	6.7 フレーム
同　標準偏差	5.5 フレーム	4.1 フレーム
音響対数尤度の平均	-68.2	-64.8
同　標準偏差	8.4	6.4

4 言語モデル

4.1 解説調の言い回しの自動生成

前述したニュース解説の 9 つの言語的特徴のうち、⑤~⑨は述部の変形であることに着目して、学習用の記者原稿に対して、ルールを適用することにより、ニュース解説特有の言い回しを含む単語列の自動生成を行った。なお、今回定義したルールは 30 種類である。trigram に反映するため、生成単語列には、ルールに当てはまる単語列(以下の下線部)に加えて、その前後の 2 単語も加える。

以下にこの生成ルールの例を示す。

【ルール例 1－特徴⑤の生成】
[連用形]、 → [連用形]まして
(記者原稿) ~ 開か れ 、 今後 月 ~
(生成単語列) 開か れ まして 今後 月
【ルール例 2－特徴⑥⑧の生成】
[連用形]ました → [過去形]んです(ね)
(記者原稿) ~ を 確認 し ました </s>
(生成単語列) を 確認 した んです </s>
を 確認 した んです ね </s>
【ルール例 3－特徴⑦⑧の生成】
[連用形]ました → [基本形]んです [が|けれども]
(記者原稿)
~ に 達して い ます が 、 店頭 で ~
(生成単語列)
に 達して いる んです が 店頭 で
に 達して いる んです けれども 店頭 で
【ルール例 4－特徴⑨の生成】
[任意]いました → [任意]いたわけです(ね)
(記者原稿) ~ が 乗って い ました </s>
(生成単語列) が 乗って いた わけ です </s>
が 乗って いた わけ です ね </s>

4.2 言語モデルの作成

言語モデルの学習データには、表 4 に示す「長期間ニュース原稿」、過去のニュース番組の「書き起こし」、および、「最新ニュース原稿」を用意した。なお、「書き起こし」は、ニュース原稿の読み上げ部分に加えて、ニュース解説等の原稿が忠実に読まれていない部分や、対談等の原稿が存在しない部分の表現を含んでいるため、ニュース解説の音声認識における学習効果が期待できる。

以下の(ⅰ)~(ⅳ)に言語モデル作成の流れを示す。

(ⅰ)「長期間ニュース原稿」と「書き起こし」を学習データとして N-gram 言語モデルを作成する。
(ⅱ)「最新ニュース原稿」に前述のルールを適用することにより、ニュース解説特有の言い回しを含む単語列を生成する。
(ⅲ)「最新ニュース原稿」に(ⅱ)で生成した単語列を加えたテキストを学習データとして N-gram 言語モデルを作成する。
(ⅳ)(ⅰ)の N-gram 言語モデルと(ⅲ)の N-gram 言語モデルを、表 4 に示す確率重みによって線形補間した言語モデルを作成する。ここで語彙は、両者の言語モデルの語彙の和集合を用いる。なお、確率重みの値は、異なる日付の学習データを利用して EM アルゴリズムにより推定した値を固定値として使用した[8]。

表 4 言語モデルの学習データの緒元

長期間ニュース原稿 (base)	'91.4.01~放送 4 時間前 [学習データサイズの例] '00.5.09：1.86M 文 '00.7.26：1.91M 文
書き起こし (base)	'97.6.01~'97.7.31. '98.4.01~'98.9.30 '99.4.01~放送前日まで [学習データサイズの例] '00.5.09：314k 文 '00.7.26：351k 文
最新ニュース原稿 (adpt)	放送 6 時間前~放送直前 [学習データサイズの例] '00.5.09：730 文 '00.7.26：665 文
ﾃｷｽﾄの確率重み	base：adpt = 0.85:0.15

本稿では、ルールにより生成した解説特有の単語列と書き起こしを学習データに加えたことによる効果を比較するため、学習データに表 5 に示す差異をつけた 4 種類の言語モデルを作成した。

表 5 各言語モデルの学習データの差異

言語モデル名	ﾙｰﾙから作成した解説特有単語列	書き起こし
LM-adpt	未使用	未使用
LM-rule	使用	未使用
LM-adpt+	未使用	使用
LM-rule+	使用	使用

4.3 言語モデルの評価

テストセットには、2000 年 5 月~7 月に「ニュース 7」で放送されたニュース解説の書き起こしすべて(149 文 3,399 単語)を使用した。そして、テストセットの各文の放送日に対応する言語モデルを放送日別に作成した。語彙サイズは 20k とし、語彙を統一するため LM-adpt には LM-rule と同一の

語彙を用い、LM-adpt+には LM-rule+と同一の語彙を用いることとした。

表 6 にテストセットパープレキシティー(PP)、trigram のヒット率(HIT)、および未知語率(OOV)を示す。なお、間投詞、言いよどみ、言い直し等の不要語は除外して評価した。表内の各値は、テストセットの各文の放送日別に異なる言語モデルで評価し、得られた最小値と最大値を記した。LM-adpt(+)のパープレキシティーの値が大きな値となっているのは、LM-rule(+)と語彙を統一したことにより、出現頻度がゼロに近い単語を語彙に含んでいるためである。よって、本表のパープレキシティーの値では各言語モデルの精度の公平な比較はできないが、trigram のヒット率に関しては、解説特有の単語列を加えること、および、学習データに書き起こしを加えることにより向上することがわかる。

表 6 言語モデルの評価

言語モデル名	PP	HIT(%)	OOV(%)
LM-adpt	33.9 ～14772.9	46.9 ～82.4	0.0 ～2.8
LM-rule	21.4 ～226.9	49.6 ～86.8	
LM-adpt+	14.9 ～253.0	50.4 ～86.2	0.0 ～2.6
LM-rule+	14.4 ～166.8	50.9 ～88.9	

5 音響モデル

以下の(ｉ)～(iii)に示す 3 種類の音響モデルを作成した。これらの音響モデルの違いを表 7 に示す。

(ｉ)AM-base － 表 2 に示した緒元による音響モデル。これを今回のベースラインのモデルとして扱う。

(ii)AM-adpt － AM-base に対し、MAP 推定[9]によって「ニュース 7」のアナウンサーの声(男性 1 名)で適応化を行った音響モデル。適応化のデータには、2000 年 3 月 27 日～4 月 28 日までの期間に「ニュース 7」で放送されたクリーン音声 9.2 時間分 3,262 文を使用した。これにより、全 5,648 モデル中 2,192 モデル(39%)が適応化された。

(iii)AM-skip －HMM を図 3 に示す状態のスキップを可能にした構造として(ii)と同様の適応化を行った音響モデル。なお、発声継続時間が短いものの出現頻度が大きい母音(/a/,/i/,/u/,/e/,/o/)と撥音(/N/)の HMM に限定して、状態のスキップを可能にした。

表 7 各音響モデルの差異

音響モデル名	話者適応	状態スキップ
AM-base	無	不可
AM-adpt	有	不可
AM-skip	有	可

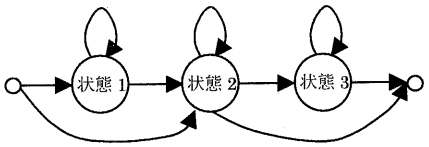


図 3 HMM の構造(スキップつき)

6 認識実験

文献[10]の 2 パスデコーダを用いて、認識実験を行った。第 1 パスは bigram による Viterbi ビームサーチから N-best(N=200)文候補を求め、第 2 パスでは trigram でリスコアリングして認識結果を得る。

テストセットには、前述の言語モデルの評価で使用したものと同じ、5 月 9 日から 7 月 26 日の期間に「ニュース 7」で放送されたニュース解説に対応するすべての音声を使用した。

認識実験の結果得られた単語正解精度(ACC)と認識処理時間の実時間比(RT×)を表 8 と表 9 に示す。なお、表 8、9 は、それぞれ言語モデルの学習データにニュース番組の書き起こしを追加する前と後の結果の比較である。

これによると、ベースラインと比較して、他のいずれの学習モデルの組み合わせにおいても認識率が向上しており、最大で AM-skip と LM-rule+の組み合わせにおいて 5.1%の認識率の向上がみられている。認識処理時間に関しては、音響モデルの話者適応を行うと高速になるが、HMM の構造でスキップを可能にすることにより遅くなり、結果としてベースラインとほぼ同じ認識処理時間となった。

詳細を調べると、言語モデルの学習に書き起こしを加えたことが認識率の改善に最も効果があった。また、ルールにより生成した解説特有の単語列を加えることによる効果は、書き起こしを加えることにより小さくなるが、逆に書き起こしが得られない場合には、特に有効であることがわかった。

表 8 認識実験結果(書き起こし追加前)

音響モデル	言語モデル	ACC(%)	RT×
AM-base	LM-adpt	**83.2**	**1.24**
AM-base	LM-rule	85.0	1.21
AM-adpt	LM-rule	86.9	0.96
AM-skip	LM-rule	87.5	1.24

表 9 認識実験結果(書き起こし追加後)

音響モデル	言語モデル	ACC(%)	RT×
AM-base	LM-adpt+	85.8	1.23
AM-base	LM-rule+	86.0	1.22
AM-adpt	LM-rule+	87.7	0.95
AM-skip	LM-rule+	**88.3**	**1.24**

次に、6月6日の「ニュース7」で放送されたニュース解説と同一話者による原稿読み上げ部分(39文 1,331 単語)をテストセットとして、表 7、表 8 に示した音響モデルと言語モデルの組み合わせにより認識実験を行った。この結果、ベースラインの単語正解精度は 97.3%であり、他の学習モデルの組み合わせによる単語正解精度は、97.2%～97.8%であった。これより、原稿読み上げ部分に関しては、音響モデル、言語モデルを変えることによる認識率の変化はほとんどみられないことがわかった。

7 まとめ

まずはじめに、ニュース解説にみられる言語的な特徴を分類した。このうち、ていねい表現と「～ですね」「～ですけれども」「～んです」「～わけです」「～ということになります」という言い回しに着目した。そして、ルールを定義して、記者原稿をもとに、ニュース解説における特徴的な言語表現を自動生成し、言語モデルの学習に加えた。さらに、大量のニュース番組の書き起こしを言語モデルの学習に加えた。

また、音響的な特徴を調べたところ、特に母音と撥音において、発声継続時間長が短いものの頻度が大きい傾向が見られた。これを考慮して、MAP 推定による話者適応に加えて、HMM の構造を状態スキップを可能にした音響モデルに変更した。

そして 4 種類の言語モデルと 3 種類の音響モデルを作成して認識実験を行い、認識率を比較した。その結果、ニュース解説の音声に対して、上記の学習モデルの改善の試みをすべて適用した場合において、最大で 5.1%の認識率の向上がみられた。一方、原稿読み上げ音声に対しては、新たに作成した学習モデルを用いても認識率に変化がみられなかった。これより、ニュース解説の音声認識においては、原稿読み上げ部分を対象とする場合に比べて、本稿で行ったような言語モデルや音響モデルの改善が重要であることが確認された。特に言語モデルの学習データに十分な書き起こしを入手できない場合には、ルールによる解説特有の表現の自動生成は有効であった。

今後は、不要語や指示表現など、ニュース解説にみられるその他の言語的な特徴にも着目した認識率改善のための検討を行っていきたい。また、誤認識箇所を詳細に調査し、音響的な問題点についてもさらに調査を進めていきたいと考えている。さらに、本手法が、講演、対談、インタビュー等のより自由発話に近いタスクの認識率改善に寄与できるかどうかについての調査検討も行っていく予定である。

参考文献

[1] 後藤 今井亨 清山 今井篤 都木 安藤 磯野, “ニュース音声認識結果のリアルタイム修正装置” 信学総大 A-15-15(2000.3) pp293

[2] 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－言語的特徴の利用の試み－” 音講論集 1-5-22 (2000.9) pp43-44

[3] 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) pp1741-1749

[4] 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) pp165-166

[5] 岩井 山本 中川, “朗読音声と自然発話音声の違いのスペクトルの分布、継続時間分布および認識率による検討” 音講論集 3-Q-3(1999.3)pp123-124

[6] 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)pp35-36

[7] 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)pp57-58

[8] 小林 今井 安藤 中林, “ニュース音声認識のための時期依存言語モデル” 情処論 Vol40,No.4(1999.4) pp1412-1429

[9] 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)

[10] 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