論文 / 著書情報
Article / Book Information

論題(和文)	ニュース音声認識システムの検討
Title(English)	A study on continuous spech recognition system for broadcast news
著者(和文)	小林彰夫, 今井亨, 安藤彰男, 宮坂栄一, 赤松裕隆, 中川聖一, 小黒玲, 尾関和彦, 古井貞熙, 鈴木順子, 白井克彦
Authors(English)	Toru Imai, Akio Ando, SADAOKI FURUI
出典(和文)	日本音響学会 1997年秋季講演論文集, Vol. , No. 3-1-9, pp. 103-104
Citation(English)	, Vol. , No. 3-1-9, pp. 103-104
発行日 / Pub. date	1997, 9

ニュース音声認識システムの検討*

◎小林 彰夫 今井 亨 安藤 彰男 宮坂 栄一（NHK 技研） 赤松 裕隆 中川 聖一（豊橋技科大）
小黒 玲 尾関 和彦（電通大） 古井 貞熙（東工大） 鈴木 順子 白井 克彦（早大）

1.はじめに

ニュースなど放送番組への字幕付与が強く求められている．筆者らは，音声認識を利用して字幕作成の支援を行うことを目的に，ニュース音声認識の研究プロジェクトを発足させた．今回は，このプロジェクトの一環としてニュース音声認識システムを構築した．

近年，アメリカのHub4[1]など，ニュース音声を対象とした大語彙連続音声認識が行われている．日本でも今回のプロジェクトに関連した研究が報告済みである[2][3]．

本稿では

- これらの研究と同様の枠組で構成したベースラインシステムの構築
- ニュース原稿データベースから言語モデルを学習する際の学習期間の検討
- 実際のニュース番組における女性アナウンサーの発声を用いたシステムの評価

について報告する．

2.認識システムの概要

図1に認識システムの概要を示す．音響モデルには状態共有化した混合ガウス分布型トライフォンHMMを使った．言語モデルはbigramおよびtrigramモデルを用いた．

システムは認識を以下の2パス[4]で行う．

第1パス

bigramモデルを使い，beam-searchによるN-bestデコーディングを行う．デコードにはHTK (HMM Toolkit)[5]を用いた．

第2パス

trigramを用い，第1パスのN-best文候補のリスコアリングを行い，1位となった文を認識結果とした．

ニュース音声データベース[6]から，システム構築用のテストセットA（表1）を選択し，これを用いて言語モデルへの重み，ビーム幅等を決定した．テストセットは，3人のNHK女性アナウンサーが発声したものである．今回は特に背景に雑音や音楽のない音声データを選んだ．

3.言語モデル

言語モデルは，NHKニュース原稿データベースの，1991年4月1日から1996年6月3日までの5年間の原稿から作成した．ニュース原稿データベースは，放送後に書き起こしたニュース原稿である．

語彙の単位は形態素とし，形態素解析ツールJUMAN[7]を用いて形態素解析した．異なる形態素の中から，出現頻度にしたがって上位20kの形態素を選び，語彙とした．総形態素数，異なり形態素数，および学習データに対するカバー率を表2に示す．

言語モデルの作成にはCMU/Cambridge

表1 テストセットA（システム構築用）

放送日： 1996年6月4日	「おはよう日本」	(女性1)	20文
		(女性2)	10文
	「ニュース7」	(女性3)	53文
	計83文，総形態素数2974		

表2 学習データ・語彙

総形態素数	38.2M
異なり形態素数	147.8k
語彙数	20k
カバー率	97.9%

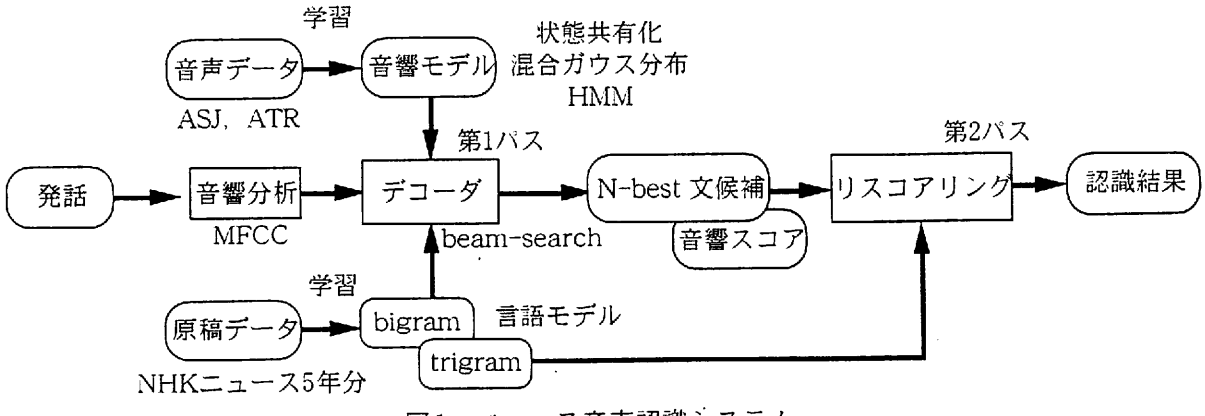

図1 ニュース音声認識システム

* "A study on continuous speech recognition system for broadcast news." by A.Kobayashi,T.Imai,A.Ando,E.Miyasaka(NHK Sci.&Tech.Res. Labs.),H.Akamatsu,S.Nakagawa(Toyohashi Univ. of Technology),R.Oguro,K.Ozeki(Univ. of Electro-Communications),S.Furui(Tokyo Institute of Technology),J.Suzuki,K.Shirai(Waseda Univ.)

SLM ToolKit Ver.2[8]を用いた．back offスムージングにはGood-Turingの推定を用いた．cut-offはbigram，trigramに対してそれぞれ，1，2とした．bigram，trigramの異なり形態素数は，それぞれ，879.9k，1.72Mとなった．

次に，この言語モデルを，学習期間を変えて作成した言語モデルと比較した．5年の学習期間はデータベースから取り得る最長の期間であり，これを評価時点（96年6月4日）の過去1年から4年までの学習期間と比較した．各言語モデルについて，語彙数を20kとしたときのテストセットAに対するパープレキシティの変化と総形態素数を調べた．図2に示すように，5年分の学習データから作成した言語モデルの，bigram，trigramでのパープレキシティはそれぞれ85.0，45.6となり，比較した中で最も低いことが確かめられた．

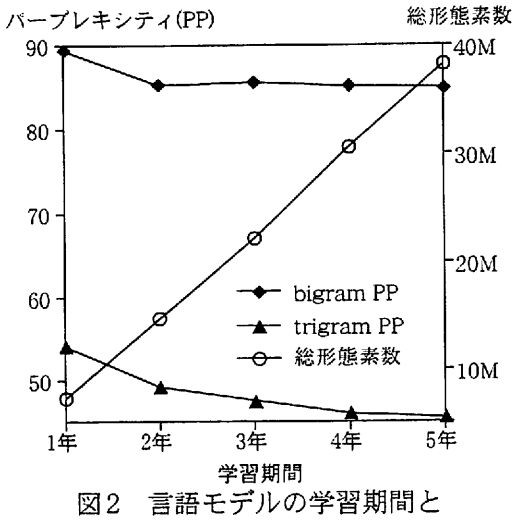

図2　言語モデルの学習期間とパープレキシティ

4.音響分析・音響モデル

音響モデルの学習には表3に示す計56名の女性話者を用いた．音響分析はフィルタバンク分析により，39次元の特徴パラメータ（12次元のメルケプストラム係数とパワー，およびそれぞれのΔ，ΔΔ係数）を得た．

音響モデルはトライフォンとし，tree-based clusteringによる状態共有化を行った．構築したHMMはモデル数1518，総状態数2362となった．また，ガウス分布の混合数は12とした．

表3　音響モデル学習データ

ATR連続音声データベース	女性計22名，4006文
ASJ連続音声データベース	女性計34名，5118文

5.トライグラムによるリスコアリング

図1のデコーダの出力はN-best文候補と対数化された音響スコアである．この音響スコアと，trigramによる対数スコアとの和をとり，文候補のリスコアリングを行った．和をとる際にはtrigramのスコアに重み付けをした．重み付けの値はテストセットAの認識実験で決定し，次節の評価実験ではこの値を用いた．150-bestでのテストセットAの単語正解精度はリスコアリング前で67.9％，リスコアリング後で71.3％となり，3.4％向上した．

6.認識実験

システムの評価は，ニュース音声データベースから得られた表4のテストセットBを用いた．テストセットBに対するパープレキシティはbigram，trigramそれぞれ78.7と38.1であった．また，未知語率は1.6％であった．150-bestまでの候補によるリスコアリングを行ったところ，第1パスでの単語正解精度74.6％，第2パスのリスコアリングで77.4％となり，2.8％認識率が向上した（表5）．

表4　テストセットB（システム評価用）

放送日：1996年6月5日	「おはよう日本」	（女性1）　14文
		（女性2）　19文
	「ニュース7」	（女性3）　33文
	計66文，総形態素数2636	

表5　認識結果（単語正解精度）

第1パス(bigram)	top-choice	74.6%
	150-best	81.1%
第2パス(trigram)	top-choice	77.4%

7.まとめ

ニュース音声に対する認識システムを構築し，単語正解精度77.4％を得た．今後はニュースの特質などを考慮しながらシステムの認識性能の向上を図っていく．

謝辞　認識システム構築にあたり，形態素解析ツールJUMANならびにCMU/Cambridge SLM Toolkitを利用させて頂いた．深く感謝する．

参考文献

[1]F.Kubala, H.Jin, S.Matsoukas, L.Nguyen, R.Schwartz, "Broadcast News Transcription", Proc. ICASSP-97 (1997)

[2]田口，大附，松岡，古井，白井，"ニュース音声を対象とした大語彙連続音声認識"，音講論集，pp.65-66 (1997.3)

[3]大附，松岡，松永，古井，"ニュース音声を対象とした大語彙連続音声認識と話題抽出"，信学技報，SP97-27，pp.67-74 (1997)

[4]R.Schwartz, Y.L.Chow, "The N-best Algorithm : An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," Proc. ICASSP-90 (1990)

[5]http://www.entropic.com/htk/htk.html

[6]安藤，宮坂，"ニュース音声データベースの構築"，音講論集，pp.157-158(1997.3)

[7]http://juman@pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[8]P.Clarkson, R.Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," Proc. EUROSPEECH-97 (1997)