論文 / 著書情報
Article / Book Information

論題	ニュース音声認識のための言語モデルと音響モデルの検討
Title	
著者	大附 克年, 古井貞熙, 桜井 直之, 岩崎 淳, 張 志鵬
Authors	SADAOKI FURUI
出典	社団法人 電子情報通信学会 研究報告, NCL98-44, SP98-108, pp. 1-7
Citation	, NCL98-44, SP98-108, pp. 1-7
発行日 / Pub. date	1998, 12
URL	http://search.ieice.org/
権利情報	本著作物の著作権は電子情報通信学会に帰属します。
Copyright	(c) 1998 Institute of Electronics, Information and Communication Engineers

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
NLC98-44, SP98-108(1998-12)

ニュース音声認識のための言語モデルと音響モデルの検討

大附克年[1] 古井貞熙[2] 桜井直之[2] 岩崎淳[2] 張志鵬[2]

[1]NTTヒューマンインタフェース研究所
〒239-0847神奈川県横須賀市光の丘1-1
[2]東京工業大学大学院情報理工学研究科
〒152-8552東京都目黒区大岡山2-12-1

あらまし 本稿では,放送ニュース音声認識における言語モデルおよび音響モデルについて検討した結果について報告する.我々はこれまで,単語(形態素)n-gram言語モデルと環境依存音素HMMを用いた大語彙連続音声認識システムによる放送ニュース音声の認識について検討を進めてきた.これまでの言語モデルでは,単語の読みが異なっていても表記や品詞が同じであれば同じエントリとして扱ってきたが,今回,表記が同じであっても読みの異なる単語は異なるエントリとして扱う読み依存言語モデルを構築した.放送ニュースでは,同じ話者が数文続けて発声することが多いため,入力音声の話者を識別しながら音響モデルを適応していくオンライン即時・逐次型話者適応について検討した.読み依存言語モデルを用いることにより単語誤り率が約10%改善され,オンライン話者適応を行うことにより単語誤り率が約15%改善されることが確認された.さらに,従来の音声認識で用いられている音響パラメータ系列に対して単語系列の事後確率を最大化する規準に対して,音響パラメータ系列に対して発声内容の事後確率を最大化する意図駆動音声認識を提案し,N-best候補の再評価に適用することでその効果を確認した.
キーワード 大語彙連続音声認識,放送ニュース音声,n-gram,オンライン話者適応,意図駆動音声認識

Language Modeling and Acoustic Modeling for Automatic Transcription of Japanese Broadcast-News Speech

Katsutoshi Ohtsuki[1], Sadaoki Furui[2], Naoyuki Sakurai[2], Atsushi Iwasaki[2], and Zhi-Peng Zhang[2]

[1]NTT Human Interface Laboratories
1-1 Hikari-no-Oka, Yokosuka-shi, Kanagawa 239-0847
[2]Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama Meguro-ku, Tokyo 152-8552

Abstract In this paper, we report on language modeling and acoustic modeling studies for broadcast-news speech recognition. We have been working on the development of a large-vocabulary continuous speech recognition (LVCSR) system for Japanese broadcast-news speech transcription. We constructed a language model that depended on the readings of words, whereas, usual language models depend on written words. In broadcast-news, each speaker utters several sentences in succussion, therefore we applied on-line speaker adaptation which is applied after identifying a speaker of the sentence. The reading-dependent language model reduced word error rate by about 10%, and the on-line speaker adaptation reduced word error rate by about 15%. We propose a new formulation for speech recognition, which maximizes the a posteriori probability of the speaker's intended message for a given observed acoustic sequence. We applied this formulation to rescoring N-best hypotheses and achieved better results with it.
Key words LVCSR, broadcast-news speech, n-gram, on-line speaker adaptation, message-driven speech recognition

1. はじめに

我々はこれまで,日本語放送ニュース音声の大語彙連続音声認識の研究を進めてきた[1][2][3]. 放送ニュース音声には,様々な話者による様々な環境での発声が含まれるため,新聞記事読み上げ音声などの人為的なデータに比べてより実際的な大語彙連続音声認識の評価を行うことができる. また,日本でも郵政省が放送番組への字幕付与を促す指針を策定するなど,放送音声の自動認識技術に対する需要の増大が見込まれる.

日本語の大語彙連続音声認識には,英語などにはないような日本語特有の問題がある. 日本語の漢字表記では,一つの表記に対して複数の読み方があることがあるため,認識辞書の一つのエントリに複数の読みが与えられる. それらの読みは実際には偏った頻度で出現するにもかかわらず認識時には等確率で扱われるために,実際には頻度の低い読みによって認識性能が劣化する問題がある. 今回,読みが異なる場合には認識辞書のエントリを別にする読み依存言語モデルを構築し,そのカバー率と性能について評価を行った.

放送ニュースでは,一つの番組の中で複数の話者が数発話ずつ発声することから,直前の話者と現在の話者とが同じであるかを識別しながら,話者ごとに音響モデルを話者適応していくことにより,効果的な適応が期待できる. そこで,現在の話者を識別しながら音響モデルを適応していくオンライン即時・逐次型(instantaneous/incremental)教師なし話者適応について検討した.

放送ニュース音声では,文頭や文中などに「え」や「えー」などといった余剰語が発声されることがある. それらの余剰語は,言語モデルの学習に用いるニュース原稿テキストには含まれないため,認識することができず認識誤りの原因となりやすい. そこで,それらの余剰語に対処した言語モデルについて検討を行った.

従来の音声認識では,入力音響パラメータベクトルの時系列に対する単語列の事後確率を最大化する規準が用いられている. しかし,音声認識を入力音声から単語列ではなく発声者の伝えようとした内容を抽出するプロセスとして考えると,入力ベクトル系列に対する発声者の意図の事後確率を最大化するような規準を用いることが必要となる. 本稿では,このための新しい定式化を提案する. この定式化は,これまでに試みられてきた種々の言語モデルに関する検討を包含し,かつ新しい方法を示唆する. 本稿では,発声者の意図は,発話中の単語の共起関係によって表されると考え,これによる定式化を示した上で放送ニュース音声認識に適用した結果を報告する.

2. 連続音声認識システム

2.1 言語モデル

放送ニュース原稿テキスト約5年分(1992年7月から1996年5月まで),約50万文を言語モデルの学習に用いた. 形態素解析システムJTAG[4]を用いて学習用テキストを形態素に分割し,その形態素を単語として単語bigramおよびtrigramを学習した. 単語出現頻度上位2万語の学習用テキストに対するカバー率は約98%であり,この上位2万語を認識語彙とした. 観測されなかったn-gramに対しては,Katzのback-off平滑化を適用した.

2.2 音響モデル

今回の実験で用いた音響モデルは,tree-based clustering[5]に基づいて設計された不特定話者文脈依存音素HMMである. 音響特徴量として,16次元LPCケプストラムと正規化対数パワーおよびそれらの一次微分の合計34次元を用いた. 学習用音声データと評価用音声データの収録系の違いに対処するために,一発話ごとにcepstral mean subtracitonによりケプストラムの正規化を行った. 学習用音声データは,ATR音声データベースBセット,日本音響学会連続音声データベース,および同模擬対話データベースから,男性53名による13270発話,女性56名による13367発話を用いた(男性音声,女性音声ともに約20時間分). 学習は男性音声と女性音声とで別々に行い,性別依存モデルを作成した. 男声モデル,女声モデルの総状態数は,それぞれ2106,2083であり,各状態のガウス分布の混合数はすべて4である.

2.3 評価データ

1996年7月に実際に放送されたニュース音声[6]から,スタジオで収録されたクリーンな発話(clean)とそれ以外の背景に雑音や音楽がのっている発話や記者レポートなどの発話(noisy)とをそれぞれについて評価用に抽出した. cleanおよびnoisyの各条件について,男性話者と女性話者の発声を各50発話ずつ抽出し評価セットとした. 各評価セットには5～6名の話者の音声が含まれている.

2.4 評価実験

2.1で述べた言語モデルと2.2で述べた音響モデルを用いて男性話者(male)によるcleanセットとnoisy

セットに対して音声認識実験を行った結果を表1に示す．trigram言語モデルを用いることによりbigram言語モデルの場合に比べ単語正解精度が絶対値で3%程度向上している．また，cleanセットに比べてnoisyセットに対する性能が低いのは，noisyセットに含まれる雑音とアナウンサー以外の話者による発話が影響していると考えられる．

表1: 音声認識結果(単語正解精度[%])

言語モデル (baseline)	評価セット	
	male.clean	male.noisy
bigram	78.5	58.4
trigram	81.3	60.8

3. 読みを考慮した言語モデル

3.1 読みの多様性

日本語の漢字表記では，同じ表記であっても複数の読み方をされる可能性がある．特に形態素解析によって一文字に区切られた漢字からなる単語は，前後の単語や意味の違いによって様々な読み方の可能性があることが多い．それらの様々な読み方の出現確率は，一様ではなくかなりの偏りを持っている[2][3]．各単語に対して可能性のある読みをすべて登録しておき，それらの読み方が等しい確率で出現するとして認識を行うと，実際には出現確率の低い読み方が認識誤りの原因となることがあった．本研究で用いた形態素解析システムJTAGは，テキストを形態素に切り分けるだけでなく高精度(形態素単位で99.6%)で読みを付与することができる．そこで，テキストの解析結果として出力される単語の読みを利用した言語モデルの構築について検討を行った．

3.2 読みの確率を用いた言語モデル[2][3]

学習用テキストの形態素解析結果から得られる各単語に対する読み方から，各単語に対する各読み方の確率を獲得し，認識時に各単語の確率に加えて読みの確率を用いるような言語モデルを構築した(LM1)．単語列中の k 番目の単語 w_k が読み $r_k \in \{r_{k1}, r_{k2}, \ldots, r_{kM}\}$ をもって現れる確率は次式のように近似して求めた．

$$P\left(w_{k=1}^{n}(r_k)\right) = \prod_{k=1}^{n} P\left(w_k(r_k) \middle| w_{i=1}^{k-1}(r_i)\right) \approx \prod_{k=1}^{n} P\left(w_k(r_k) \middle| w_k\right) P\left(w_k \middle| w_{i=1}^{k-1}\right) \quad (1)$$

3.3 読み依存言語モデル

学習用テキストの形態素解析結果から，表記と読みとの両方が一致するものだけが同じ言語モデルエントリとなり，表記が一致しても読みが異なる単語は別の言語モデルエントリとなるような言語モデルを学習した(LM2)．この読みに依存した言語モデルでは，従来の言語モデルで一つの言語モデルエントリだったものが同じ表記を持ち読みの異なる複数のエントリに分割されるため，語彙数が同じ場合にはカバー率が低下する．表2に従来の言語モデル(baseline/LM1)と読み依存言語モデル(LM2)の学習用テキストおよび評価セットに対する20k語彙の未知語率を示す．表2をみると言語モデルエントリを読み依存にすることによるカバー率の低下(未知語率の増加)はそれほど大きくないことがわかる．

表2: 20k語彙に対する未知語率[%]

言語モデル	学習用テキスト	評価セット	
		male.clean	male.noisy
baseline/LM1	2.27	0.81	2.88
LM2	2.39	0.86	3.02

3.4 評価実験

従来の読みを等確率に扱う言語モデル(baseline)，読みの確率を用いる言語モデル(LM1)，および読み依存言語モデル(LM2)をそれぞれ用いて評価実験を行った．実験結果を表3に示す．表3をみると読みの確率を導入することにより単語正解精度が改善されている．また，読み依存の言語モデルはカバー率が若干低いにもかかわらず読みの確率を用いたモデルよりも高い性能を示している．

表3: 読みを考慮した言語モデルの評価 (単語正解精度[%])

言語モデル		評価セット	
		male.clean	male.noisy
bigram	baseline	78.5	58.4
	LM1	79.0	60.5
	LM2	79.6	60.9
trigram	baseline	81.3	60.8
	LM1	82.4	63.7
	LM2	83.2	64.1

LM2は読みの種類数が言語モデルエントリ(認識語彙)の数と同じ20000であり，baselineおよびLM1の読みの種類数21541と比べて1割程度少なくなっている．また，認識処理に要する時間もLM2の場合にはLM1の場合に比べると2割以上削減されている．女声の評価セットを女声音響モデルとLM2を用いて認識した結果を表4に示す．

表4: 読み依存言語モデルの評価(female)
(単語正解精度[%])

言語モデル		評価セット	
		female.clean	female.noisy
bigram	LM2	82.7	57.3
trigram	LM2	86.4	60.6

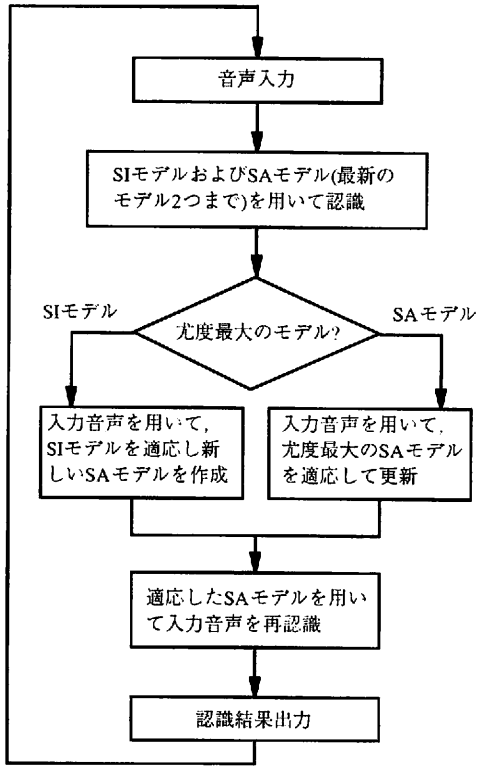
4. オンライン話者適応

4.1 放送ニュースにおける話者の変化

放送ニュース音声には，スタジオのアナウンサーによる発声だけでなく，中継先の記者の発声やVTR映像にあわせて原稿を読み上げた発声など様々な話者の発声が含まれている．しかし，一人の話者は少なくとも一つのニュースをまとめて発声するため，同じ話者が数発話に渡って続けて発声していることが多い．そこで，現在の話者が先行する発話の話者と同じ話者かどうかを判定しながら音響モデルを適応していくオンライン即時・逐次型(instantaneous/incremental)教師なし話者適応について検討した．

4.2 オンライン教師なし話者適応

話者の識別は，入力音声に対し不特定話者モデルと最大過去二人の適応話者モデルを用いて認識した際の音響尤度を比較することによって行った．不特定話者モデルが高い尤度を示した場合には，新しい話者が検出されたとして，その音声を用いて不特定話者モデルを適応して新しい適応話者モデルを作る．適応話者モデルの方が高い尤度を示した場合には，そのモデルの話者の音声だと判断してその適応話者モデルを適応して更新する．話者適応手法は，まず，MLLR[7]およびMAP[8]によりモデルパラメータの変換行列(平均のみ)を求め[9]，その後VFS[10]により移動ベクトルを平滑化する方法を用いた．オンライン話者適応処理の流れを図1に示す．



SI モデル: 不特定話者音響モデル
SA モデル: 話者適応音響モデル

図1: オンライン教師なし話者適応処理の流れ

4.3 評価実験

オンライン教師なし話者適応を行いながら男性話者によるcleanセット(male.clean)に対してbigram言語モデル(LM2)を用いて認識実験を行った結果を表5に示す．話者の識別は，評価セットの50発話中およそ3分の2の発話について正しく行われていた．また，評価セットを話者の変わる部分で区切ると9つの話者セグメントに分けることができるが，その話者セグメントごとの認識結果を図2に示す．図2をみると，9セグメント中8セグメントについて，オンライン話者適応により性能が改善していることが確認できる．

表5: オンライン話者適応の評価

音響モデル	単語正解精度[%]
baseline	79.6
on-line adaptation	82.7

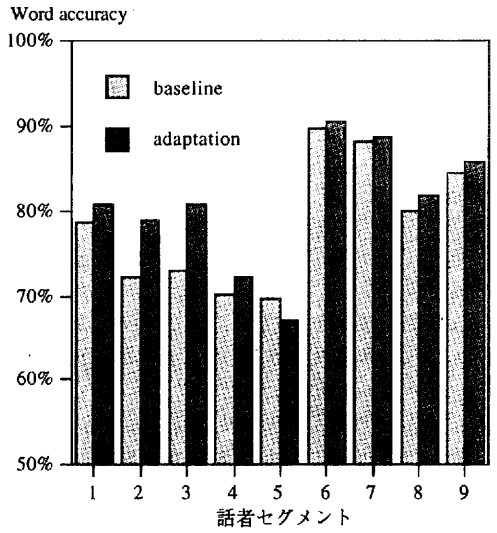


図2: 話者セグメントごとの評価(単語正解精度)

5. 余剰語対策

放送ニュース音声には,会話音声などに比べれば頻度が少ないが,「え」「えー」などの余剰語(不要語)が文頭あるいは文中に含まれることがある.これらの余剰語は,言語モデル学習用のニュース原稿テキストには含まれていないので,認識誤りの原因となりやすい.そこで,学習用テキストの文頭および読点「,」の部分に「え」または「えー」を一定の頻度でランダムに挿入し,そのテキストを用いて言語モデルを学習した(LM3).余剰語を挿入する頻度は,放送ニュース音声の書き起こしテキストに基づいて求めた.文頭の「え」12.1%,文頭の「えー」2.1%,読点の後の「え」3.4%,読点の後の「えー」1.0%の確率でそれぞれ学習用テキストに余剰語を挿入した.また,従来,読点は学習テキストから削除していたが,読点を削除せずに言語モデルを学習し,発音辞書の読点のエントリには,無音に加えて「え」および「えー」を追加することを行った(LM4).これらの言語モデルを認識実験により評価した.評価セットは,放送ニュース音声データベースから余剰語を含む男性話者の発話を30発話抽出したものを用いた.評価結果を表6に示す.評価セットの発話中の余剰語部分を波形編集ソフトを用いて取り除いた場合の結果(no interjections)もあわせて示す.言語モデルを余剰語に対処するように改良することで余剰語が含まれることによって劣化した認識性能のうち約20%を改善している.

表6: 余剰語対策言語モデルの評価

言語モデル	単語正解精度[%]
baseline	61.4
LM3	62.4
LM4	63.1
no interjections	68.9

6. 意図駆動音声認識

6.1 意図駆動音声認識

最近の音声認識では,本稿でここまで述べてきたものも含めて,音響パラメータベクトルの時系列 X に対する単語列 W の事後確率 $P(W|X)$ を最大化する規準が用いられている.読み上げ音声をディクテーションすることが目的の場合には,この規準で十分かもしれないが,音声認識を入力音声からその発声者が伝えようとした内容を抽出することであるとすると,図3に示すようなモデルで考えることが必要となる[11].図3のモデルに基づいて考えると,音声認識のプロセスは,次式のように音響パラメータベクトルの時系列 X に対して発声者の伝えようとした内容 M が与えられる確率 $P(M|X)$ を最大化するような M を選ぶことになる.

$$\arg\max_M P(M|X) = \arg\max_M \sum_W P(M|W)P(W|X) \quad (2)$$

式(2)はBayesの定理により,次式のように書き換えることができる.

$$\arg\max_M P(M|X) = \arg\max_M \sum_W P(X|W)P(W|M)P(M) \quad (3)$$

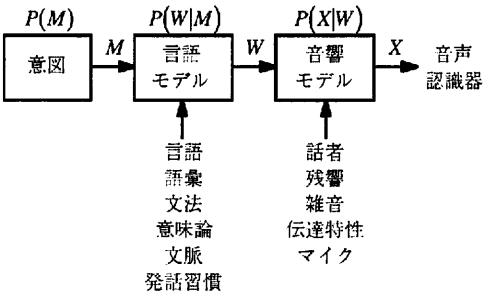


図3: A communication - theoretic view of speech generation & recognition

さらにこれを次式のように近似する．

$$\arg\max_{M} P(M|X) \approx \arg\max_{M,W} P(X|W)P(W|M)P(M) \quad (4)$$

ここで，$P(X|W)$は従来通りHMMなどの音響モデルで表されるとする．また，$P(M)$はMに無関係に等確率であるとする．すると問題は，$P(W|M)$を如何に表現するかということに帰着する．ここでは，$P(W|M)$を近似的に以下のように表すことができると考える．

$$P(W|M) \approx (1-\lambda)P(W) + \lambda P(W|M) \quad (5)$$

ここで，λは，$0 \le \lambda \le 1$の重みである．右辺の第1項$P(W)$は，$P(W|M)$の中でMに独立な部分を表し，実際には従来の統計的言語モデルによって表現される．右辺の第2項$P(W|M)$は，Mに依存する部分を表している．

$P(W|M)$の表現形式としては，Mに対して明示的(explicit)に定式化するか，暗示的(implicit)に定式化するかによって，種々の方法が考えられる．明示的に表現する場合には，通常Mを有限の数の話題（トピック）クラスで表すことが必要になり，話題クラスによって言語モデルを切り替える方法がこれに相当する[12]．一方，Mを暗示的に表現する方法の一つがcacheモデルである[13]．本稿では，Mは発話中の単語の共起によって表されると考える[14]．これに類する方法としては，シソーラスを用いる方法や単語をクラスタ化する方法も考えられる．これらの方法では，Mを明示的に表すことなく$P(W|M)$をあらわすことができるため，発声者が伝えようとした内容Mに関する情報を柔軟に用いることができる．

6.2 単語共起スコア

前述のLM3のtrigramによって得られた300-bestの単語列仮説に対して，単語共起の情報に基づいて再評価を行った．発声者の伝えようとした内容を単語の共起によって表す場合，発声の内容や話題を表す単語はほとんどが名詞であるので，単語仮説中の名詞のみを抽出した．単語w_iと単語w_jとの共起スコアは，相互情報量に基づいて次式のように求めた．

$$CoScore(w_i, w_j) = \log \frac{p(w_i, w_j)}{\left(p(w_i)p(w_j)\right)^{1/2}} \quad (6)$$

ここで，$p(w_i, w_j)$は単語w_iと単語w_jとが同じ発話の中に共起する確率，$p(w_i)$と$p(w_j)$はそれぞれ全学習データ中の単語w_iと単語w_jの出現確率である．式(6)の右辺の分母の平方根は，出現頻度の低い単語に対する値を補正するために適用している．認識語彙中の各名詞間の共起スコアを，言語モデルの学習に用いたものと同じテキストデータを用いて計算した．

6.3 評価実験

単語列仮説中のすべての名詞の組の共起スコアを，各単語列仮説の尤度に加算して仮説の再評価を行った．男性話者の評価セットに対する単語正解精度(message-driven)を表7に示す．共起スコアに対する重みは実験的に適切な値に設定した．表7からわかるように，$P(W|M)$を考慮することによって単語正解精度が向上することが確認された．

表7: 意図駆動音声認識結果(単語正解精度[%])

言語モデル	評価セット	
	male.clean	male.noisy
LM2	83.2	64.1
message-driven	83.9	64.3

7. まとめ

本稿では，日本語の放送ニュース音声認識のための言語モデルと音響モデルに関する検討について報告した．日本語の読みの多様性に対処するため，読み付与精度の高い形態素解析ツールを用いて得られた各単語に対する読みを利用して，読みに依存した言語モデルを構築した．その結果，同じ表記で読みの異なる単語を別の言語モデルエントリとしても，語彙サイズが同じ場合の未知語率はそれほど大きくならず，単語正解精度は従来の言語モデルに比べて向上した．放送ニュース番組の中での話者の変化を検出しながら音響モデルを各話者に対して適応していくオンライン即時・逐次型教師なし話者適応について検討した．bigram言語モデルを用いてmale.cleanセットに対して評価実験を行ったところ，評価セット中の3分の2の発話について正しく話者が検出され単語正解精度も79.6%から82.7%まで向上した．放送ニュース中の発話の文頭および文中に含まれる余剰語によって認識性能が低下するのを抑えるため，文頭および読点に余剰語の挿入を許す言語モデルを構築した．余剰語を含む発話30発話について評価を行ったところ，余剰語によって劣化した性能のうち約20%を回復することができた．

また，従来の音声認識で用いられていた音響パラメータベクトルの時系列に対する単語列の事後確率を最大化する規準に対して，発声内容（意図）の事後確

率を最大化する規準に基づく意図駆動音声認識を提案した．発話中の単語共起関係によって発声者が伝えようとした意図が表されると考え，単語共起スコアに基づく意図駆動音声認識の評価実験を行ったところ単語正解精度が改善された．

今後は，読み依存言語モデルに適した認識単位の検討，オンライン話者適応処理の効率化，余剰語用の音響モデル，などについて検討をすすめていく．また，意図駆動音声認識のための発声者の意図の表現方法についてもさらに検討していく．

謝辞

ニュース原稿とニュース音声を提供していただいた日本放送協会に感謝します．形態素解析ツールを提供していただいたNTTヒューマンインタフェース研究所情報通信研究所知的通信処理研究部の渕武志研究主任に感謝します．意図駆動音声認識に関して議論していただいたLucent Technology Bell LaboratoriesのB.-H. Juang博士に感謝します．本研究の一部は，財団法人国際コミュニケーション基金の助成を受けて行われたものである．

参考文献

[1] T. Matsuoka, Y. Taguchi, K. Ohtsuki, S. Furui, K. Shirai, "Toward Automatic Transcription of Japanese Broadcast News," Proc. EUROSPEECH, Vol. 2, pp. 915-918, 1997.

[2] S. Furui, K. Takagi, A. Iwasaki, K. Ohtsuki, T. Matsuoka, S. Matsunaga, "Japanese Broadcast News Transcription and Topic Detection," Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp. 144-149, 1998.

[3] 高木幸一，桜井直之，岩崎淳，古井貞熙，"ニュース音声を対象とした言語モデルと話題抽出の検討，" 信学技報，SP98-33, pp. 73-80, 1998.

[4] 渕武志，松岡浩司，高木伸一郎，" 保守性を考慮した日本語形態素解析システム，" 情報処理学会．自然言語処理研究会 NL97-4, pp. 59-66. 1997.

[5] S. J. Young, J. J. Odell, P. C. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modeling," Proc. DARPA Human Language Technology Workshop, pp. 307-312, 1994.

[6] 安藤彰男，宮坂栄一，"ニュース音声データベースの構築，" 音講論，2-Q-9, pp. 157-158, 1997-3.

[7] C. J. Leggetter, P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, pp. 171-185, 1995-9.

[8] J. -L. Gauvain, C. -H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994-4.

[9] 石井純，外村政啓，" 重回帰モデルに基づく話者適応方式の検討，" 音講論，3-3-17, pp. 119-120, 1996-9.

[10] 大倉計美，杉山雅英，嵯峨山茂樹，"混合連続分布HMM移動ベクトル場平滑化話者適応方式，" 信学論，D-II, Vol. J76-D-II, Nol. 12, pp. 2468-2476, 1993-12.

[11] B. -H. Juang, "Automatic Speech Recognition: Problems, Progress & Prospects," IEEE Workshop on Neural Networks for Signal Processing, 1996.

[12] S. F. Chen, K. Seymore, R. Rosenfeld, "Topic Adaptation for Language Modeling using Unnormalized Exponential Models," Proc. ICASSP' 98, pp. II-681-684, 1998.

[13] R. Kuhn, R. De Mori, "A Cache-based Natural Language Model for Speech Recognition," IEEE Trans. PAMI-12, 6, pp. 570-583, 1990.

[14] Z. S. Harris, "Co-occurrence and Transformation in Linguistic Structure," Language, 33, pp. 283-340, 1957.