音声言語情報処理 34－22
(2000. 12. 22)

話し言葉音声の認識を目指して

篠崎 隆宏、斎藤 洋平、堀 智織、古井 貞熙
東京工業大学大学院情報理工学研究科計算工学専攻

〒152-8552 東京都目黒区大岡山 2-12-1
Tel/Fax : 03-5734-3480

{staka, yohei, chiori, furui}@furui.cs.titech.ac.jp

あらまし　話し言葉音声の認識を目指して平成11年度に開始したプロジェクトに関連して、講演音声、対談音声、討論音声などを対象として進めている種々の検討状況を報告する。実際の話し言葉コーパスから作成した音素モデルや言語モデル、話題の分野に対応した過去のニュースや教科書を用いた未知語の登録、対談のクロストークの区間における音響 back-off などが有用であることが確認された。認識性能に個人差が大きく、発話速度、フィラー数、言い直し数などに関連していることなどが確認された。会議などの議事録を、音声認識システムとユーザとが対話を行いながら効率的に作成する方法についても検討した。話し言葉の音声認識性能はまだ低く、認識対象としての文単位の抽出法、発音辞書、コーパス作成における書き起こし法など、今後解決しなければならない研究課題が多い。

キーワード　話し言葉音声認識、話し言葉プロジェクト、講演、対談、討論、未知語、音響 back-off

Toward Spontaneous Speech Recognition

Takahiro Shinozaki, Yohei Saito, Chiori Hori and Sadaoki Furui
Tokyo Institute of Technology, Department of Computer Science

2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Tel/Fax : 03-5734-3480

{staka, yohei, chiori, furui}@furui.cs.titech.ac.jp

Abstract　This paper reports various investigations on recognizing spontaneous speech such as lectures, interviews and discussions conducted in relation with our national project started in 1999. Usefulness of acoustic and linguistic modeling based on actual spontaneous speech corpora, registration of new words using past broadcast news or a textbook related to the areas of topics, and an acoustic backing-off method for the periods of cross talk in interviews have been confirmed. Recognition accuracy has a wide speaker-to-speaker variability according to the speaking rate, number of fillers, number of repairs, etc. This paper also reports a method for efficiently making minutes of meetings based on interaction between a speech recognition system and a user. The recognition accuracy for spontaneous speech is still very low, and there exist a large number of research issues including how to extract pseudo-sentence unit speech for recognition, how to build pronunciation dictionaries, and how to transcribe spontaneous speech in corpora.

key words　spontaneous speech recognition, national project, lectures, interviews, discussion, OOV words, acoustic backing-off

1．はじめに

新聞記事などの書き言葉の読み上げ音声や、アナウンサーが発声している部分のニュース音声のように書き言葉に近い音声であれば、高い精度で音声認識ができるようになってきたが、自由に発声した話し言葉になると認識率が急激に低下するのが、現在の音声認識技術の実態である。その主たる原因は、話し言葉と書き言葉（書き言葉を読み上げた音声）は、音響的にも言語的にも大きく異なるにもかかわらず、これまでの音響モデルや言語モデルが、書き言葉あるいはそれを読み上げた音声に基づいて作られているためである。これから音声認識の用途を広げていくためには、話し言葉の認識性能を向上させることが必須である。

このような背景から、話し言葉の構造を明らかにし、話し言葉の音声認識理解技術を高めることを目標として、昨年新しいプロジェクト[1]が開始された。５年間で、大規模な話し言葉コーパスの構築と、話し言葉の音声認識理解技術の確立を目指している。このプロジェクトを遂行する主たる研究グループは、国立国語研究所、郵政省通信総合研究所、東京工業大学である。このプロジェクトは、次の３つの目標のもとに５年間継続される予定である。

(1) 約 7M 語からなる長さ約 800 時間分の大量の話し言葉コーパスを構築する。主として講演などのモノローグを収録し、書き起こし、形態素解析する。全体の 1/10（コア）に関しては、パラ言語情報など、詳細情報をマニュアルで付与する[2]。
(2) 話し言葉音声の言語モデルと音響モデルを構築し、音声認識理解および要約技術を構築する。
(3) 話し言葉の自動要約システムのプロトタイプを作成する。

これまでに、種々の学会の講演音声や、模擬講演を収録し、予備的な実験を始めている。本論文では、プロジェクトに関連して東工大で進めている、話し言葉の音声認識に関する研究状況を報告する。

2．講演音声の認識

2.1 認識タスクと実験条件

上記プロジェクトで録音した日本音響学会と日本音声学会の音声に関する講演の音声（４名分）を、認識対象として用いた。ヘッドセット型の接話型指向性マイクロホンで収音し、DAT に録音した後、書き起こしてある。各講演音声の正式データ名称は下記の左側であるが、簡単のため、ここでは右側の表記を用いる。

AS99SEP022 → “022”　AS99SEP023 → “023”
AS99SEP097 → “097”　PS99SEP025 → “025”

音声は 16kHz で標本化、16bit で量子化した。音響パラメタは MFCC12 次元、Δケプストラム 12 次元、対数エネルギーの１次差分の 25 次元で、切り出した発話区間ごとに平均ケプストラムによる正規化(CMS)を行った。形態素を単位とする統計的言語モデルを用い、正解文や言語モデルの作成には NTT で開発された形態素解析ツール JTAG を使用した。

各講演の長さを図１に、講演の発話速度を形態素数と音素数で計った結果を図２に示す。発話速度の計算には、講演時間中の実際に発話されている区間を用いた。フィラー数と言い直し数を図３に示す。

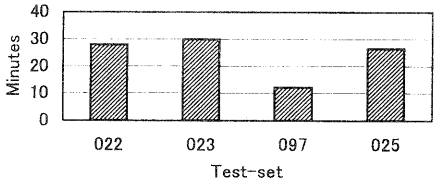

図１　各講演の長さ

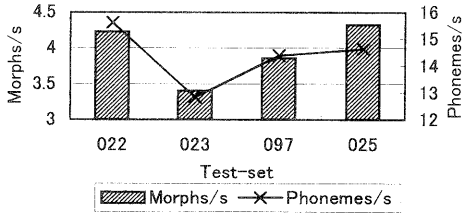

図２　発話速度

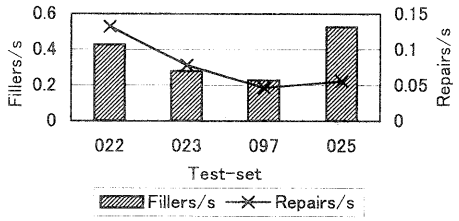

図３　発話中のフィラーと言い直し回数

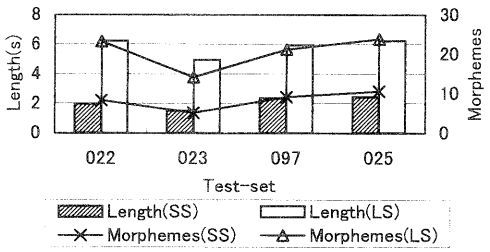

図４　実験に使用した文長

音声認識に用いる文単位としては、プロジェクトで 200ms の無音を基準に切り出した転記単位(SS: Short Sentence)を、人手で文らしくまとめて作った単位(LS: Long Sentence)を用いた。話し言葉では文境界があいまいなため、まとめ方には任意性がある。まとめる際、間に挟まれる雑音区間はテストセットに含めた。LS の文長を SS と比較して図４に示す。

2.2 言語モデル

形態素を単位とする以下の統計的言語モデルを使用した。

LM2-1：World Wide Web 上で公開されている講演書き起こしテキストを収集しコーパスを作成した（総形態素数：約 2M）[3]。講演音声では、文頭および読点の前後に間投詞が出現することが多いので、この 3 箇所に間投詞（22 種類）がどのような頻度で出現しているかを、プロジェクトの書き起こしデータを用いて調べた。この頻度に基づいて、上記の講演書き起こしコーパスに間投詞を加えた後、言語モデルの学習を行った。言語モデルの語彙数は 20k である。

LM2-2：LM2-1 のコーパスにテキストの段階で、教科書「音声情報処理」[4]テキスト（総形態素数：63k）を加えてモデルを作成した。語彙数は 20k である。

LM2-3：プロジェクトで収録した講演の書き起こし（テストセットと異なる男女計 97 名、18.7 時間、23 万形態素）について、転記単位を文末に来やすい語彙を手がかりに文らしい長さにまとめて用いた。転記単位の接続部には句点を追加した。語彙数は 8k である。

2.3 音響モデル

音素モデルとして、次の状態共有環境依存不特定話者 HMM を使用した。

AM2-1：IPA の男性モデル（読み上げ音声から作られた 2k 状態 16 混合のモデル）

AM2-2：プロジェクトでの収録講演（テストセットと異なる男性 66 名、12.4 時間）から作成したモデル。学習時の音素ラベルは発音表記から作成した。無音は母音と/N/の後に forced alignment を用いて入るかどうか決定した。状態数は 1k、1.5k、2k の中では、予備実験で 1.5k が全体として最も良かった。以下では 1.5k の場合の結果を示す。

2.4 実験結果

作成した言語モデル(3-gram)のパープレキシティと未知語率を図 5 に示す。語彙数が違うので LM2-1、LM2-2 と LM2-3 のパープレキシティはそのまま比較できないが、各講演に対する相対的な比較のため同一の図に示す。LM2-1 では未知語率が高いが、音声の教科書を加えることで LM2-2 では LM2-3 並となっている。

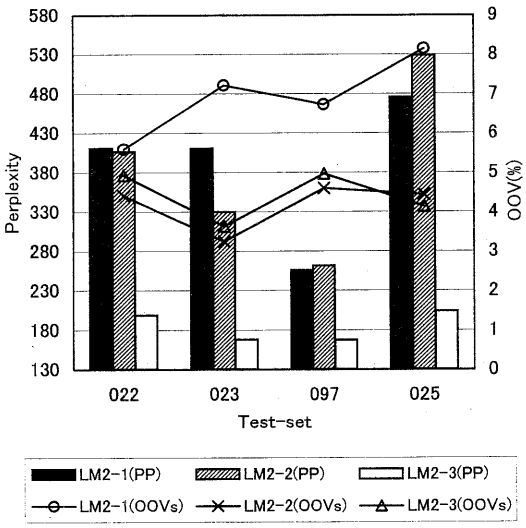

図 5　パープレキシティと未知語率

認識実験には Julius3.1 を使用した。実験にはそれぞれの講演全体から、LS を単位として、3 文に 1 つの割合で抜き出した文集合を用いた。結果のスコアリングは形態素の単位で行った。フィラーもスコアに含め、「え」「えー」などは異なる形態素として計算した。

言語モデルとして LM2-3 を用いたときの、音響モデルと認識率の関係を図 6 に示す。読み上げ音声から作られた AM2-1 よりも話し言葉コーパスから作られた AM2-2 の方が高い認識率を示した。

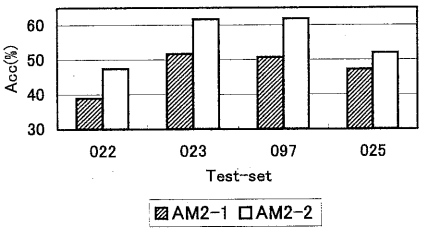

図 6　音響モデルと認識率

言語モデルと認識率の関係を図 7 に示す。LM2-3 は LM2-1、LM2-2 に比べ 1/10 程度のデータ量しかないにもかかわらず、比較的高い性能を示した。

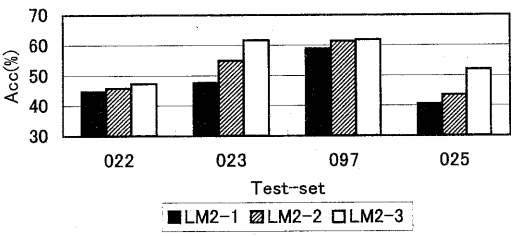

図 7　言語モデルと認識率

2.5 考察

得られた講演音声の認識率はまだ低いが、少量でも話し言葉コーパスを用いることの有効性が確認できた。発話速度が速く、フィラーや言い直しが多い話者の認識率が低い。認識誤りを見てみると、「音声」が/oNse:/でなく/oNsei/になっているなど、発音辞書に振られた発音と実際の発音の不整合が原因と思われるものがある。なお、文単位の問題として、上記言語モデルを用いて、プロジェクトの転記単位(SS)をそのまま認識すると認識率がやや下がる傾向がある。

3．対談音声の認識

3.1 認識タスクと実験条件

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[5]。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、男女の対談部分を用いた。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

テストセット	文数	形態素数
男女対談部分	40	1396
女性単独発話	22	450

3.2 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。

LM3-1：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。

LM3-2:「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。

LM3-3:LM3-2に品詞N-gramを用いて未知語の追加を行った。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。

3.3 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。

AM3-1：IPAの男女別モデル。

AM3-2: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、AM3-1をもとにML推定を行った。対談の男声部分は4章の話者と同じなので、その単独発話音声で話者適応したモデルを使った。

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[6]。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）。

3.4 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LM3-1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。

表2　Bigramのテストセットパープレキシティ

テストセット	LM3-1	LM3-2	LM3-3
対談	922.52	78.60	85.62
キャスター	322.48	113.72	128.17

表3　Trigramのテストセットパープレキシティ

テストセット	LM3-1	LM3-2	LM3-3
対談	887.42	64.97	70.21
キャスター	306.70	101.49	115.24

図8に、各言語モデルにおける未知語率を示す。LM3-1とLM3-2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、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LM3-1では学習されていないことによる。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。

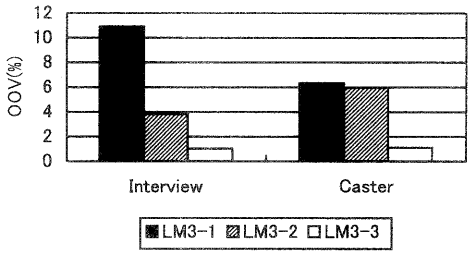

図8　未知語率

各言語モデル別の認識率を図9に示す。認識実験にはJulius2.1を用いた。音響モデルはAM3-1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10に示す。言語モデルはLM3-2を用いている。この図におけるAB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を適用した結果を図11に示す。クロストークを含むテストセットは女性対談19文中9文、男性対談21文中16文であった。

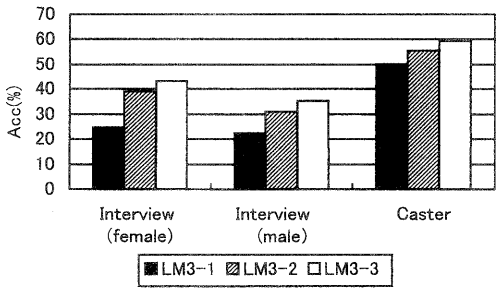

図9　言語モデルの性能評価

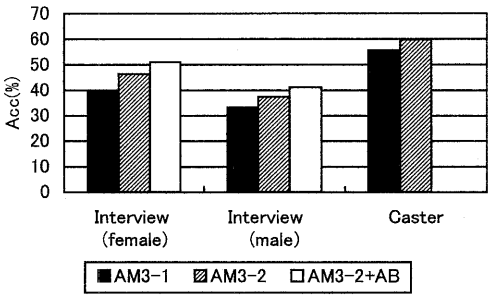

図 10　音響モデルの性能評価

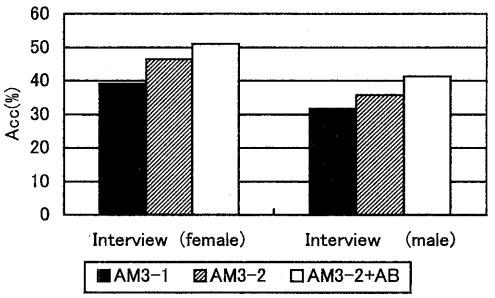

図 11　音響 Back-off の性能評価

3.5 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。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。

4．対話型議事録作成システムのための予備的検討

4.1 タスクとシステム構成

会議などにおいて書き起こしを作成しておくと音声での記録に比べ後からの参照、検索が容易となる。書き起こし作成は手間のかかる作業であるので、音声認識による自動生成が望まれるが、話し言葉を対象とすると、十分な認識率が得られないのが現状である。そこで人が書き起こしを作る作業を支援する形で音声認識を利用する、コンピュータ支援型書き起こし支援システムを提案する。システム構成を図 12 に示す。

このシステムでは、始めに認識エンジンが音声から認識文を作成する。次にユーザが誤りのうち幾つかを指摘すると、システムはそれに基づき内部のモデルを更新し、より精度の高い認識文を出力する。このような対話を繰り返すことにより、少ない労力で書き起こしが作成できる。対話を成り立たせるためにはシステムが認識をやり直す時間が短くなくてはならない。そこで始めに一回だけ音声からスケッチを作成して保存しておき、その後はスケッチ上で認識処理をする。スケッチは音声認識の中間表現や抽出した話者情報などの集合で、対象とするタスク全体に対して作成し、保存する。システムは Julius を元に作成している。スケッチに用いる中間表現としては単語トレリス形式とワードグラフ形式を検討した。現在の Julius では全体として高速なトレリス形式を中間表現に用いているが、対話的なシステムとしては第 2 パスに相当する部分の計算時間が短いことが必要であることから、ワードグラフ形式を採用した。本システムでは音響モデルが逐次更新されることになる。中間結果上でリスコアをする場合、音響モデルの変更は音素境界などに影響し、認識率を低下させることが考えられる。しかしあらかじめ教師なし適応を行うことによりこの影響を減らすことを試みた。

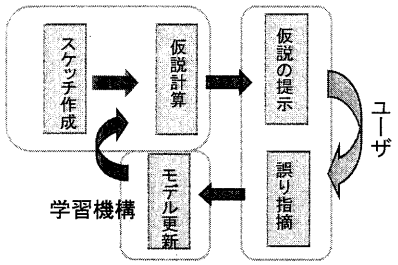

図 12　対話型議事録作成システム構成の構成

実験はラジオとテレビで放送された政治家の討論音声「日曜討論」（全部で 60 分）の内、話者 1 名(m1)の音声（217 文）を切り出して用いた。前半の 100 文を学習（話者適応）に、後半の 117 文を評価に用いた。

4.2 言語モデルと音響モデル

言語モデルには、LM2-1 を用いた。ただしフィラーは m1 の前半の情報を用いて追加した。音響モデルには、次のモデルを用いた。

AM4-1：IPA の男性モデル。2k 状態 16 混合。

AM4-2：後半 117 文を用い、教師なし適応を行ったモデル。

AM4-3：**AM4-1** を初期モデルとし、適応用 100 文を用いて教師あり話者適応（MAP+VFS）した HMM。

4.3 実験結果

同じ認識精度を得る条件で、トレリス形式と単語グラフ形式で第 1 パスと第 2 パスそれぞれにかかる時間の例を図 13 に示す。図では提案システムの動作を想定し、第 1 パスと第 2 パスの間で音響スコアのキャッシュをクリアしてある。3 種類の AM を用いたときの単語グラフの精度、第 1 パスと第 2 パスで同じ音響モデルを使用した場合(Same AM)、第 2 パスから AM4-3 を用いた場合(Different AM)の認識率を図 14 に示す。第 1 パスに AM4-1 を用い第 2 パスに AM4-3 を用いた場合では、ともに AM4-1 を用いた場合と比べて認識率はよくならなかった。しかし第 1 パスに AM4-2 を用いた

場合、第2パスからAM4-3に切り替えることで始めからAM4-3を使った場合に近い認識率を得ることができた。

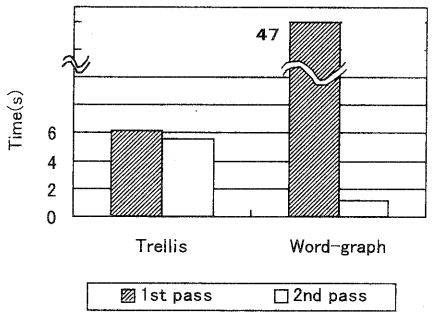

図13　認識処理にかかる時間

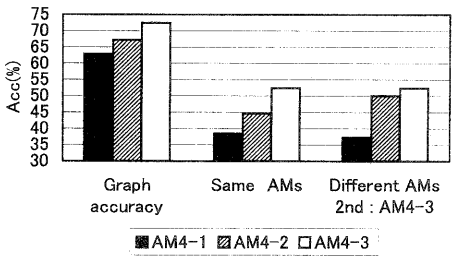

図14 音響モデルの組み合わせと認識精度

4.4 考察

音声認識システムとユーザが対話をしながら書き起こしを作成するシステムについて検討した。対話におけるシステムのレスポンスを早くするために、スケッチ上で処理をする方法を検討した。その際音響モデルのミスマッチが問題となるが、教師なし適応をしておくことで対応することができることを示した。今後は未知語や、始めに中間結果上に現れなかった単語に後から対処する方法、誤り指摘をより有効に利用する方法、システムの側から質問をすることでより効果的に書き起こし作成を進める方法などを検討したい。

5. むすび

本論文では、最近開始された話し言葉の音声認識理解のためのプロジェクトに関連して東工大で進めている研究状況について報告した。

学会における講演音声、対談、討論の音声を対象として、話し言葉の音声認識の予備的検討を行い、講演書き起こしコーパスを用いた言語モデルや音響モデル、音素モデルの適応化、クロストークにおける音響back-off、未知語を話題に応じて言語モデルに組み込む方法などが単語誤り率の削減に有効であることを確認した。しかし、まだ単語誤り率が50%近くに上っていて極めて大きく、今後の大幅な改善が必要である。

このためには、大規模な話し言葉コーパスの構築が急務である。話し言葉のコーパスと音響および言語モデルに関連した今後の課題には、

- 話し言葉の書き起こし法
- 話し言葉の発音辞書
- 話し言葉の形態素解析法
- 精密かつ一般的な間投詞モデル
- 言い直しや言い淀みへの対処
- 未知語を含む言語モデルのタスク適応
- 音声認識に用いる文単位の抽出法
- 話し言葉の認識に適した音響単位とモデル化

などがある。

話し言葉の認識においては、言い直し、言い淀み、間投詞などは、そのまま文字化するよりも、むしろ認識結果から除去した方がよい。この意味で、音声認識から内容の理解、さらに要約への展開が重要と思われる。音声認識結果から、重要語をキーとする要約文を生成する方法[7]は、音声理解の方法の一つとして位置付けることもできる。

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、プロジェクトの推進研究者各位、特に貴重な討論と協力をいただく京都大学の河原助教授、ほか研究室員の協力に感謝する。

参考文献

[1] S. Furui et al.: "Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –", Proc. ICSLP, Beijing, pp. 518-521 (2000)

[2] K. Maekawa et al.: "Spontaneous speech corpus of Japanese", Proc. LREC 2000, Athens, Greece, pp. 947-952 (2000)

[3] 加藤、李、河原："講演ディクテーションのための話題独立言語モデルと話題適応"，音声言語情報処理研資、26-2 (1999)

[4] 古井：音声情報処理、森北出版 (1998)

[5] 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)

[6] 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)

[7] 堀、古井："話題語と言語モデルを用いた音声自動要約法の検討"、信学技報、SP99-110 (1999)