音声言語情報処理 34－11
(2000. 12. 21)

会議音声の自動アーカイブ化システム

秋田 祐哉　河原 達也

京都大学大学院 情報学研究科 知能情報学専攻
〒606-8501 京都市 左京区 吉田本町
e-mail: akita@kuis.kyoto-u.ac.jp

あらまし　会議音声に対して階層的なアーカイブを構築し，議事録の作成支援を行うシステムを提案する．まず会議音声に対してGMMによる話者識別を行い，その結果により音声を分割するとともに話者IDや時間情報などのインデックスを生成する．また談話標識を含むキーフレーズの検出により議論の結論となる発話を特定し，議事次第や会議の配布資料などに含まれる話題依存語彙を利用してこれを自動的に書き起こし，議事録のドラフトとする．以上により音声・インデックス・テキストの3層からなるアーカイブの構築を試みる．本稿ではGMMによる話者識別とキーフレーズ検出に関して評価実験を行い，その有効性を確認した．

キーワード　音声認識，会議音声，話者識別，キーフレーズ検出，アーカイブ，議事録

Automatic Archiving System for Meeting Speech

Yuya Akita　Tatsuya Kawahara

Graduate School of Informatics
Kyoto University, Kyoto 606-8501, Japan
e-mail: akita@kuis.kyoto-u.ac.jp

Abstract　In this paper, automatic archiving system for meeting speech is presented. At first, meeting speech is segmented by the result of speaker identification using speaker GMMs, and indexed with speaker-ID and beginning/end time information. Secondly, concluding utterances are detected by spotting key-phrases including discourse markers, and automatically transcribed using topic-dependent words in the agenda or material of the meeting, and finally made into drafts of minutes. In this way, a three-layer archive including speech, indexes, and texts is generated. We made experiments on speaker identification using GMMs and key-phrase detection, and confirmed the feasibility.

key words　speech recognition, meeting speech, speaker identification, key-phrase detection, archive, minutes

1 はじめに

私たちが日常行う会議は議事録として記録することが一般的である．しかし公平・公正な議事録の作成は負担が大きい．会議の客観的な記録として音声を録音することもよく行われるが，長時間の録音音声を効率よく参照するためには適切な索引(インデックス)が必須であり，その抽出・付与作業には大変な手間がかかる．

一方，近年では計算機の性能やマルチメディア関連技術が急激に進歩し，音声を計算機上のディジタル・アーカイブとして保存できるようになった．また音声処理技術の発展がめざましく，話者認識や音声認識の技術が実用的なレベルまで達している．

本研究ではこうした技術を会議音声の自動処理に応用することを考える．まず音声に話者などのインデックスを付与してアーカイブ化する．さらに処理した音声の中から議論の展開上重要な発話を選んで書き起こすことができれば，これを編集することで実際の発話を元にした議事録を従来より効率的に作成することができる．

本研究ではこのように，検索・参照が容易なインデックスが付与された二次的な資料としての音声アーカイブの生成を目指す．また議事録作成を支援するために，議事録に求められる要素を会議の音声から自動的に抽出して書き起こすことも目的とする．抽出した議事録要素からは，各議題に対応する音声アーカイブの区間に対してリンクを設けることで有機的に結合する．

以下，2においてアーカイブの枠組みについて論じ，3にてシステムの詳細を述べる．またシステムの要素技術について行った評価実験について4で述べる．

2 アーカイブの枠組み

2.1 概要

本研究では，数名から十数名で行われる小規模の会議を対象として想定する．システムには入力として会議の録音音声とその会議の議事次第や配布資料を与える．会議音声は発話ごとに話者が識別され，話者ごとに分割されて保存される．このとき話者IDや発話の開始時間・終了時間がインデックスとして生成される．こうしたインデックスは一覧に整理されて，音声リストとして保存される．さらに，議事録の要素となるような発話はシステムによって自動的に書き起こされて，テキストとしても保存される．計算機による書き起こしでは完全に正しいテキストを得ることはできないため，これは議事録のドラフトとして編集に供される．修正や編集を容易にするために，書き起こしテキストは発話音声とリンクされる．議事次第などは議題の取得の他に，テキストの自動書き起こしの際の未知語対策として話題依存語を登録するためにも用いられる．

システムが生成するアーカイブの例を図1に示す．1次アーカイブは会議の記録として残される議事録である．2次アーカイブは，音声からシステムが自動的に抽出したデータであり，議事録の編集に利用される．また議事録から，その記述の根拠を探すための索引としても用いられる．3次アーカイブは会議の音声そのものである．

議事録はシステムが書き起こしたテキストを元に人手で編集される．議事録はその元となったテキスト(議事録ドラフト)と，その議題に対応する音声リストにリンクされる．議事録ドラフトは，それに含まれる各発話のテキストが，それに対応する音声リスト中の発話データにリンクされる．音声リストは議題の中で発された発話のそれぞれのインデックス情報が含まれ，それは実際の音声ファイルとリンクされる．

2.2 ドラフト生成における問題点

議事録ドラフトを生成する際には，議論の内容の要約，音声の書き起こしの2つの処理が必要である．人が議事録を作成する場合，議論の内容を聞きながらメモを取り，後程メモと記憶や録音音声をもとに作成するのが普通である．しかしながら計算機で行う場合には，入力音声の意味を即座に把握してこのような記憶を行うことは不可能である．一方，先に音声をすべて文字に書き起こしてから要約処理を行うことも，会議音声は良好な録音環境とは言い難く，また計算機による書き起こし自体も完全ではないため困難である[1]．したがって本研究では要約処理は人手に頼らざるを得ないものとして，これを支援する方向で考える．

本研究では，会議の際に議論のまとめとなるような発話に着目する．この場合問題となるのは，そのような要約発話が存在することが自然であるかどうか，この発話をどうやって抽出するかの2点である．そこで本研究では，会議のプロセスを以下のようにモデル化して考える．

1. 議題の説明
   議論の冒頭では，その議題に関する担当者の説明が行われる．これはその議題を検討するに至った経緯や検討対象となる人・物の状況といったことである．議題が議論ではなく報告を目的とする場合は，この説明部分で議題が終了する．
2. 問題提起
   説明を受けて，現在問題となっていることは何か，検討課題を明らかにする．議論の最初は担当者によって提起され，また対応の原案(たたき台)が提示されることが多い．

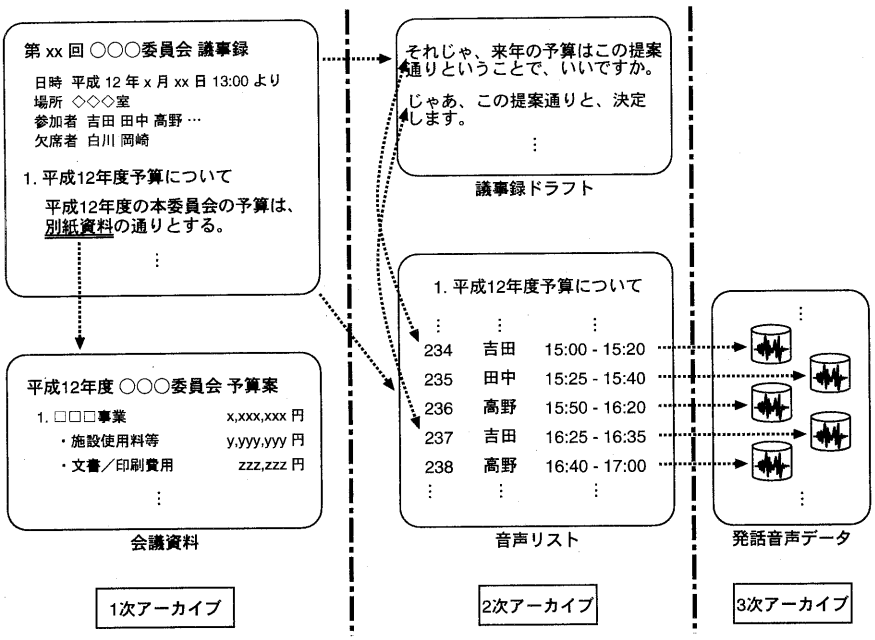

図 1: アーカイブの例

3. 議論
   問題に対して，その解決法・対応などを検討する．
4. 結論
   議論の結果得られた意見・方策をまとめ，会議としての結論とする．結論は議長によって確認される．

議事録は議論の要点をまとめ，会議の結果どういった対応をとるかを示すものである．上記の考察から，結論の部分が議事録に対応し，したがって議長のまとめの発話を書き起こしのターゲットとするのが妥当であると考えられる．

次に，議長の発話から目的の結論部分を抽出する手法について検討する．議長は結論のみ発話するわけではないので，結論とそれ以外を区別する必要がある．

結論の発話はそれまでの議論を踏まえて行われるものであるので，それまでの議論を受ける意味を持つ言葉が含まれる．具体的には，文頭に「(それ) では」や「じゃあ」という言葉が現れる．これらを談話標識と呼ぶ．本システムでは，このような談話標識が検出できた場合は要約発話と見なすことにする．

2.3 手法の検証

以上の議事録ドラフト生成手法が妥当であるかどうか，実際の会議音声を用いて検証した．検証に用いた音声は1999 年に京都大学情報工学教室にて行われた教室の計算機に関する会議である．この結果，次のような点が明らかとなった．

- 議題 5 例のうち 4 例で，完全ではないものの議事録ドラフトとして十分なテキストを得ることができた．
- 発話に出現する指示代名詞が指し示す語句・文脈の情報が欠落する．例えば,「では，これを 5 セット購入することにします」という発話に対して，この発話だけでは「これ」の指し示すものが分からない．これらの語句・文脈情報は，談話標識を含む発話の前に存在するからである．ただし議題などから明らかなものもあり，必ずすべての語句を抽出する必要はない．
- 説明の段階での発話は抽出できないが，これを抽出できると議事録が詳細になる．
- 議論の要約が議長以外の参加者によって行われ，決定事項が抽出できなかった例があった．進行が議長以外に委ねられた場合でも，要約は最後に議長によって行われるよう留意してもらうことで，ある程度の情報は得られると期待される．

指示代名詞の展開や議事説明者の発話の抽出に関しては今回のシステムでは検討せず，今後の課題とする．

3 システムの構成

システムの処理の流れを図 2 に示す．会議音声は議長と議長以外の一般参加者とで分離して別チャネルで録音され，処理もそれぞれ異なる．

議長の発話は一定時間の無音で区切られて音声ファイルとして保存する．さらに談話標識に対応する要約発話を示すキーフレーズが検出された場合は音声認識による自動書き起こしを行って議事録ドラフトとして保存する．

議長以外の一般参加者の音声は，無音により分割したのち話者識別を行い，話者の代わる点で区切られて音声ファイルとして保存される．

3.1 収録環境

本システムでは簡便さと実現性を考慮して集音マイクで全員の音声を録音する．ただし議長の発話については音声認識の対象となるため，議長発話のみ接話型マイクを用いて，他の参加者と独立に録音する．すなわち，音声は 2 チャンネルで収録する．

3.2 セグメンテーション・話者識別

一般参加者の発話は話者ごとに分割して保存される。まず、入力音声に対して無音により分割する。分割の対象となる無音の長さは 0.5 秒である。次に、分割された音声に対し話者識別を行う。ここでは GMM による手法を用いる [2]。GMM の学習は会議の冒頭などで録音された 30 秒程度の任意の音声により行い、混合数は 16 である。話者識別の結果、同一話者の音声区間が連続する場合はそれらを連結し、単一の音声ファイルとして扱う。

3.3 キーフレーズ検出

議長の発話のうち，議事録として抽出する発話は，特定のキーフレーズが含まれるかどうかで判断する．具体的なキーフレーズは「それでは」などといった談話標識である．しかし，談話標識のみによる抽出は単語が短く誤検出が多くなると考えられるため，本システムでは談話標識を含むもう少し長いキーフレーズ (1 文) を検出対象とする (図 3)．これはいくつかのパターンを文法で記述する．議長にはこの文を要約発話に先立って発してもらう．やや自然性に欠けるが，検出を安定にする．

本システムでは，発話検証を用いたキーフレーズ検出を行う [3],[4]．これは文法ベースの認識パーザ Julian[5] にキーフレーズ文法とフィラーからなるモデルを与えて実現する．

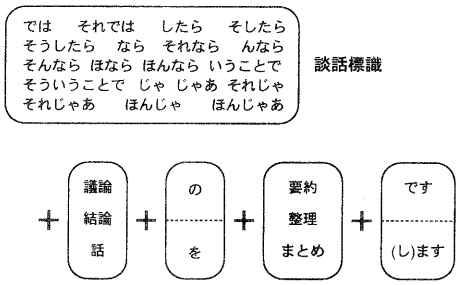

図 3: キーフレーズ検出文法

3.4 自動書き起こし

議長発話のうちキーフレーズが検出されたものに対しては，議事録ドラフトとするための自動書き起こしを行う．書き起こしは次に述べる話題語追加済みの言語モデルと認識デコーダ Julius[6] を用いて行う．Julius は当研究室で開発された，単語 trigram を用いた大語彙連続音声認識エンジンである．

Julius に用いる言語モデルは大規模な学習データが用意できる新聞記事を用いている．このため，新聞記事に出てくるような語句による読み上げ音声については高い精度で認識でき，最高 95%とされている．しかし，新聞記事ベースの言語モデルは特定分野の専門用語などあまり一般的でない単語 (ドメイン依存語) には対処できない．ドメイン依存語を Web 上から取得する手法も提案されているが [7]，本システムでは議事次第や配布資料から抽出した話題語を追加することで対処する．

4 評価実験

本章では，自動アーカイブ化システムの実現に不可欠な，話者識別と議事録ドラフト生成について評価実験を行った結果について述べる．

4.1 話者識別実験

話者識別は音声のアーカイブを生成する上で重要な処理である．本システムでは GMM による手法を用いる．この実験の目的は，会議音声に対するこの手法の性能を明らかにすることである．

実験の対象は，実際の会議の録音音声と，本研究室で収録された模擬会議の音声である．実際の会議の音声は議長の発話も含めてすべて集音マイクで録音された．S/N 比は 0-10 dB 程度である．一方模擬会議の方は，すべての話者が接話型のマイクを用い，S/N 比は 30-40 dB 程

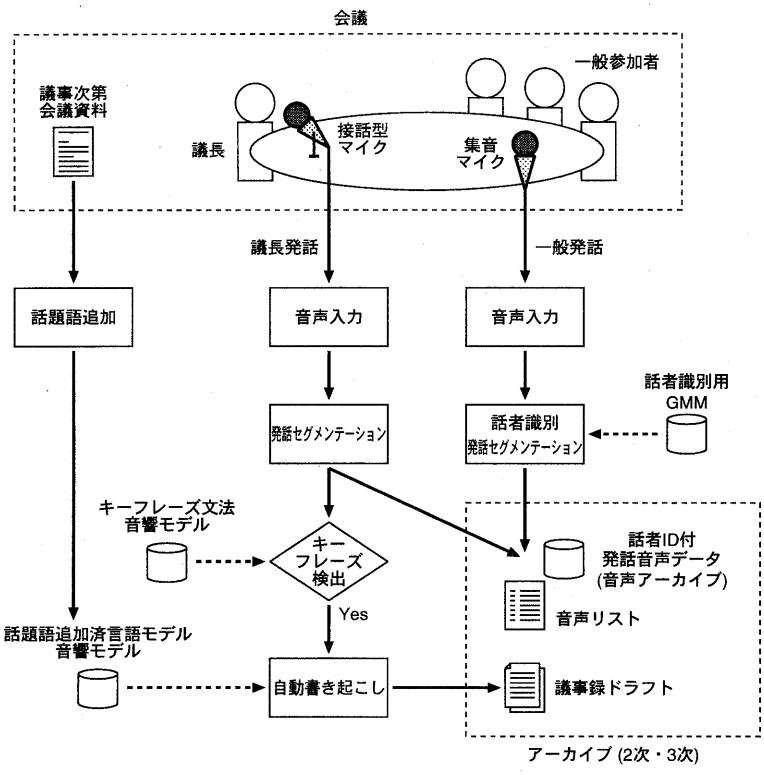

図 2: 処理の流れ

表 1: 話者識別実験結果

会議	話者	総時間	発話	正解数	識別率
実際 (1)	6	3 分 30 秒	84	70	83.3%
実際 (2)	6	8 分 30 秒	166	129	77.7%
模擬 (1)	7	8 分 20 秒	50	45	90.0%
模擬 (2)	8	8 分 00 秒	37	34	91.9%

度である．また模擬会議ではあらかじめ会議のシナリオが準備され，それに沿った議論が行われている．議題数はともに 2 である．

識別率は無音ごとに区分化された区間単位である．ただし，実際の会議の音声は，相槌などのように同時に複数の参加者が発話している部分が存在する．この場合はもっとも大きな声で発している話者がその音声区間の話者と見なしている．識別用の GMM は，実際の会議の方は会議の別の区間の音声を，また模擬会議の方は会議とは別に収録した学習用音声を参加者それぞれ 30 秒程度用いて学習している．

実験結果を表 1 に示す．実際の会議でも 80% 程度の識別率が得られた．しかし，実際の会議における識別率は，同一話者でも発話によって識別率が大きく異なる．これはまず話者の声の大きさ，すなわち S/N 比が主な原因と考えられる．例えば何かを強く主張するときなどは高い S/N 比が得られるため識別率がよいが，近くの人と相談するようなひそひそ声，またぼそぼそ声などは逆に S/N 比が悪く，識別率も悪くなる．このほか，咳をする音，紙をめくる音，机の振動など，大きな雑音によっても識別率は大きく低下する．また相槌や笑い声など，同時に複数の話者が発話している区間も識別は難しい．

一方，話者間で識別率に差が出る原因は，性別や個人差のほか，マイクと座席の距離による収録音声レベルの差がある．これに対しては，マイクロフォンアレイを用いて音声強調することも考えられる [8]．

模擬会議においては，各話者とマイクの距離は数十 cm であるため，S/N 比が高く識別率はよく，発話単位ではほぼ完全に識別できた．しかしやはり個人差が存在する．

4.2 キーフレーズ検出実験

前節で用いた模擬会議の音声のうち議長の発話に関して，議事録ドラフトを生成するために必要な処理のうち，キーフレーズ検出に関する実験を行った．検出はキーフレーズを含む文法と音声認識パーザJulianにより行う．なお音響モデルはtriphone HMM (3,000状態/16混合分布/男性) のものを用いている．

対象となった発話は議題1・議題2ともに11発話である．うちキーフレーズ文の発話は各議題1発話ずつである．キーフレーズ文以外の発話には談話標識は含まれない．実験の結果，キーフレーズ文の発話は正しく検出し，またそれ以外の発話から検出されることはなかった．

比較のため，文法が談話標識 (「それでは」など) のみからなる発話検証モデルにより上記の発話についてキーワード検出を行ったところ，いくつかの誤検出が見られた．このことから，長いキーフレーズによる検出がこれらの湧き出し誤りを防ぐ効果が確認された．

なお，キーフレーズ文の発話と他の発話の間に十分なポーズが置かれず，キーフレーズ文を一文として区分化できなかった場合は検出に失敗する．すなわち，キーフレーズ文はそれに続く発話とはっきり区別して発話しなければならない．このキーフレーズの発声の自然性と負荷については，実際の会議での検証が必要である．

5 結論

本研究では会議音声の自動アーカイブ化システムを提案した．話者識別やキーフレーズ検出，音声認識などを用い，アーカイブは音声・インデックス・テキストの3層で構成される．自動アーカイブ化の中心となるこれらの処理について評価実験も行い，実現可能性を確認できたが，実際の会議音声を処理する上で以下のような問題点も明らかになった．

- 議長の発話において指示代名詞がもたらす情報の欠落
- 議事説明者の発話の抽出
- S/N比の悪い発話に対する話者識別精度の低下
- キーフレーズの自然性・キーフレーズ発声の負荷

このほか，本来会議の内容は公開する性質のものではないため，研究に利用できる会議音声データがきわめて少ないことも重大な問題として挙げられる．

今後は，会議運営上の制約ができるだけ小さくなるようこれらの問題を解決してシステムを改善し，その後は実際の様々な会議に適用してみることで向上を図っていく予定である．

謝辞

本研究では京都大学情報工学教室の会議を実験に利用しました．また，模擬会議は京都大学大学院情報学研究科音声メディア研究室の皆さんにご協力いただきました．関係各位に謝意を表します．

参考文献

[1] Hua Yu, Cortis Clark, Robert Malkin, and Alex Waibel. Experiments in Automatic Meeting Transcription Using JRTk. In *Proc. ICASSP*, pp. 921–924, 1998.

[2] 松井知子. HMMによる話者認識. 電子情報通信学会技術研究報告, SP95-111, 1996.

[3] Tatsuya Kawahara and Shuji Doshita. Topic Independent Language Model for Key-Phrase Detection and Verification. In *Proc. ICASSP*, pp. 685–688, 1999.

[4] 河原達也, 石塚健太郎, 堂下修司. 発話検証に基づく音声操作プロジェクタとそれによる講演のハイパーテキスト化. 情報処理学会論文誌, Vol. 40, No. 4, pp. 1491–1498, 1999.

[5] 李晃伸, 河原達也, 堂下修司. 文法カテゴリ対制約を用いたA*探索に基づく大語彙連続音声認識パーザ. 情報処理学会論文誌, Vol. 40, No. 4, pp. 1374–1382, 1999.

[6] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (99年度版) の性能評価. 情報処理学会研究報告, 2000-SLP-31-2, 2000.

[7] Hua Yu, Takashi Tomokiyo, Zhirong Wang, and Alex Waibel. New Developments in Automatic Meeting Transcription. In *Proc. ICSLP*, Vol. 4, 2000.

[8] 山田武志, 中村哲, 鹿野清宏. マイクロホンアレーを用いた話者方向検出による音声認識. 電子情報通信学会技術研究報告, SP95-65, 1995.