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自動ハイパーテキスト化

河原達也† 石塚健太郎† 堂下修司†

音声認識技術を一般の機器操作に用いるためには，コマンド発話以外の音声を正しく棄却できる必要がある．本研究では，使用される状況に依存した話し言葉の特徴をとらえた発話検証用モデルを用いることを提案し，講演をしながら音声で操作できるスライドプロジェクタを設計・実装した．検証用モデルは，講演の書き起こしテキストからドメインに依存した名詞を除去しながら，頻出音節系列を抽出することにより学習した．このモデルにより，従来の音節連接モデルに基づく手法に比べて，はるかに高い発話検証性能を得ることができ，音声操作プロジェクタの安定な動作を実現した．さらにこのプロジェクタにより，講演音声をオンラインスライドテキストと対応づけた形式で自動ハイパーテキスト化するシステムを作成した．

Voice-operated Projector Using Utterance Verification and Its Application to Hyper-text Generation of Lectures

TATSUYA KAWAHARA,† KENTARO ISHIZUKA† and SHUJI DOSHITA†

In order to apply speech recognition to operation of electronic devices, the system needs capability to correctly reject irrelevant inputs. We propose a verification model depending on the speaking-style when the device is used, and then develop a slide projector that can be operated via voice commands during a lecture. The verification model is trained with transcription text of oral presentations by extracting frequent syllable sequences after filtering out topic-dependent nouns. It achieves much better verification performance than the conventional methods, thus makes the voice-operated projector practical. Furthermore, we develop a system that automatically generates hyper-text of lecture speech by aligning it with the on-line slides.

1. はじめに

近年の音声認識技術の進展にともない，種々のアプリケーションが開発されるようになってきた．音声による機器の操作は，その最も初歩的かつ典型的な例である．そのようなシステムはかなり以前から作成されてきたが，実環境において頑健に動作しない場合が多い．実環境における頑健性には，雑音や歪み等への対策とならんで，想定したコマンド以外の発話を正しく棄却できる能力が重要である[1]．この発話検証の能力が十分でないと，ユーザにコマンド発声のたびに音声入力用のボタンを押してもらう等の負担を課したり，きわめて小語彙のシステムしか実現できないことになる．

本研究では，講演中に（講演自体に使用しているのと同一の）マイクロフォンで操作できるスライドプロジェクタを主に想定して，高精度の発話検証手法を提案する[2]．本手法は，このプロジェクタが使用されるような講演調の話し言葉の特徴を，講演の話題とは独立にモデル化するものである．どのような講演にでも頻繁に出現するような言い回し（＝フィラー）のパターンを抽出することにより，これを実現する[3]．

さらに，この音声操作プロジェクタを利用して講演の音声とオンラインスライドテキストの自動ハイパーテキスト化を行うシステムを提案する．これは，スライドの切換えコマンドによって，音声とスライドの対応づけが行えることを利用したものである．これにより，スライドやその目次を手がかりにその説明箇所の音声を容易に取り出すことが可能になり，膨大な講演音声データの効率的なアーカイブ化が実現できる．

以下，2章で発話検証の定式化とフィラーモデルの役割について述べた後に，3章で話題に独立で話し言葉に依存したフィラーモデルの構築方法を説明する．

† 京都大学大学院情報学研究科知能情報学専攻
Graduate School of Informatics, Kyoto University

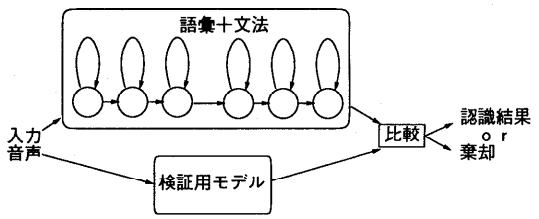


図 1 発話検証の処理の概要
Fig. 1 Overview of utterance verification.

4章では，このモデルの音声操作プロジェクタへの適用と実験的評価について述べる．5章では，これに基づいて作成した講演音声の自動ハイパーテキスト化システムについて説明する．

2. 競合モデルを用いた発話検証

2.1 発話検証

発話検証は，音声認識結果（W）が受理できるか否かを判定し，未知語やシステムの想定外の発話を棄却する処理である．通常はそのためのモデル（λ_V）を用意しておく．入力 X に対する W と λ_V の尤度比 LR（対数スケールでは尤度差）を計算して，しきい値より上回れば認識結果を受理する．

$$LR = \frac{P(X|W)}{P(X|\lambda_V)}$$

$$\log LR = \log P(X|W) - \log P(X|\lambda_V)$$

この処理の概要を図 1 に示す．

発話検証用モデル λ_V は，コマンド以外の（文法外）音声入力に対して高い尤度を与え，かつコマンドの（文法内）入力に対しては低い尤度を与えることが望ましい．音節モデルを用いる方式[4),5)]が最も一般的であるが，いずれの音声に対しても尤度が高くなるので，両者を識別する能力が十分とはいえない．

そこで，我々は競合サブワードモデル（anti-subword model）[6)]を利用して得られる認識結果のサブワードごとの尤度比を総合して，キーフレーズの検証を行う手法を提案し，その有効性を示した[7)]．

2.2 競合フィラーモデル

これに対して，キーワードスポッティングにおいて，キーワード以外の区間を近似するフィラーモデルを利用することが有効であることが知られている[8)~10)]．本手法は，図 2 (a) に示すように，発話全体をモデル化することにより入力全体を評価したうえで，キーワードが含まれる尤度を計算するものである．言語的制約により湧き出し誤り（False Alarm）を抑制するとともに，長さの異なるスコアの正規化や入力に応じたしきい値の設定を自動的に行うことができる．

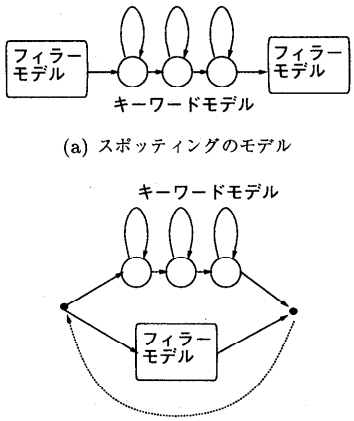


図 2 フィラーモデルのスポッティング・発話検証への適用
Fig. 2 Filler model for spotting and verification.

我々はこのフィラーモデルに関していくつかの比較を行い，以下の傾向を確認している[10)]．

- 音節連接モデルは頑健だが不十分である．
- タスクの語彙のモデルが有効かつ頑健である．
- 単語対（文法）モデルは強力だが頑健でない．

特に，数百の語彙を登録していれば，音節連接モデルを用意しなくても，未知語が音響的に類似した単語モデルで近似されるために，未知語を含む発話に対しても動作することを示した．また文献 9) では，これに加えて，語彙サイズを数百から数千に増やしてもあまり効果がないことも報告されている．

図 2 (a) は入力中にキーワードがちょうど 1 回出現するモデルとなっているが，実際には複数回出現したりまったく出現しない場合もあるので，図 2 (b) のようなループを持つモデルを使用する．フィラーモデルは，キーワード以外の区間の近似および尤度正規化を行うものであり，特に図 2 (b) の最下部点線のループの遷移を除去すると，入力がキーワードのみであるか，キーワードを含まないかという特殊な場合に相当し，図 1 で示した発話検証のためのモデルと同等になる．

そこで本研究では，このような特徴を持つフィラーモデルを発話検証に利用する．すなわち，キーワードに競合させる検証用モデルとしてフィラーモデルを用いる．特に，語彙レベルの知識が有効であるので，その利用を考える．我々は以前，限られたタスクドメインにおいて，語彙の制約を用いることで文法外発話を棄却する方式を検討した[11)]．しかしながら，必ずしも入力発話の語彙が予測可能であるとは限らない．特に，講演や会話中に特定の機械にコマンドを発声するようなアプリケーションにおいては，一般の講演や会話の

ドメイン（話題や語彙）を予測することはできない．

3. 話し言葉依存のフィラーモデル

したがって，ドメインに独立で話し言葉の特徴に依存した競合モデルを構成することを考える．講演においては，話題（ドメイン）に関係なく一定の表現が多用されるので，このような講演調発話に特徴的なフィラーモデルを学習する．固有の言い回し等のモデル化を目標とするが，統計的に頻出する音節系列を抽出することにより，これを実現する．すなわち，日本語においては単語の定義は曖昧であるので，単語の単位にこだわることなく，講演調発話に普遍的なパターンを抽出する．

このモデルの構築には，1995 年 5 月に行われた情報処理学会音声言語情報処理研究会（SLP）のパネル討論における 5 名の講演の書き起こしテキストを使用した．このデータは，「重点領域研究音声対話コーパス☆ Vol.4」に含まれている．テキスト中の総音節数は 18109 である．

フィラーの抽出は以下のように行う．

(1) 形態素解析（JUMAN）を行い，講演の話題に関係する普通名詞と固有名詞を除去する．

(2) 形態素解析の区分に関係なく，ただし除去された形態素の区間を含まないように，n 音節連鎖（$n = 3 \sim 6$）の系列を求めて，出現頻度がしきい値以上の音節系列を抽出する．

(3) 抽出されたすべての音節系列集合を統合し，長い音節系列の一部分にマッチングする（サブストリングとなる）短い音節系列を除去する．

出現頻度は，テキスト中の総文字数に対する比で正規化し，そのしきい値は 0.1%と 0.05%の 2 通りを使用した．各音節長 n における頻出系列の種類を**表 1** に示す．これらをマージすることによって最終的に，0.1%の場合で 70 個，0.05%の場合で 230 個の系列（長さ $n = 3 \sim 6$）が抽出された．その一部を**表 2** に示す．

個人差もあると考えられるが，講演調の発話において典型的な言い回しが抽出されていることが分かる．

なお本研究では，講演の話題（ドメイン）に関連する単語を除去するために形態素の属性情報を利用したが，この処理を統計的に行うことも可能である．

4. 音声操作プロジェクタ

4.1 仕　　様

以上のモデルを用いて，音声操作プロジェクタを設計・実装した．これは，計算機の画面をそのまま講義室の大スクリーンに投影できる環境を利用しており，計算機上のスライド表示ソフトを音声で操作することにより，スライドの切替えを行う．スライド表示ソフトとして，Netscape ブラウザおよび xdvi（LaTeX）プレビューアが利用できる．これらのページ切替えコマンドを音声で入力する．

ユーザは，同一のマイクロフォンで講演を行いながら，プロジェクタに対するコマンドを発声する．コマンドは，“次”，“2 ページ前” といったキーフレーズで，有限状態文法（FSA）で記述されている．コマンド用文法の語彙サイズは 56 である．

このアプリケーションでは，コマンド以外の入力発話の話題や語彙を限定できない．ただし，講演調の発話であることは仮定できるので，前章で構築した講演調スタイル依存のフィラーモデルをコマンドキーフレーズの検証に利用する．

さらに安定した認識を行うために，コマンドの前にマジックワードを発声する仕様も考えた．マジックワードは，発声の自然さを考慮して，“オペレータ” とした．なお，音声の切り出しのために，コマンド（マジックワードを含みうる）の前後にポーズを入れることのみを前提としている．

本システムの言語モデルは図 **3** のようになる．

4.2 発話検証実験

音声操作プロジェクタを実現するにあたって，本論文で提案する発話検証手法の評価を行った．

男性 5 名に（音声操作プロジェクタでなく通常の OHP を使用して）自由なプレゼンテーションを行ってもらった．音声データは，いったん DAT に 48 kHz, 16 bit で録音したうえで，16 kHz にダウンサンプリ

表 1 n 音節系列の総数と頻出した系列の種類
Table 1 Number of syllable sequences.

構成音節数 n	総数	しきい値 0.05%	しきい値 0.1%
3	15987	325 種	118 種
4	14288	145 種	63 種
5	12988	90 種	31 種
6	11813	53 種	13 種

表 2 抽出されたフィラーの例
Table 2 Examples of fillers.

しきい値 0.05%	だいたい，である，ようするに，ですけれども，かもしれない，もので，なんですけど，なので，ただし，というふうに，とおもいます，だから
しきい値 0.1%	ではないか，かもしれない，みたいなもの，おもうんです，というような，やっぱり

☆ http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus

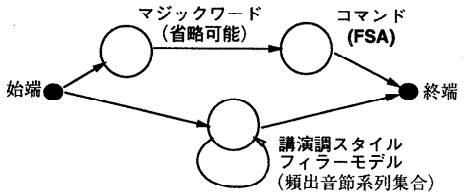


図 3 音声操作プロジェクタ用の言語モデル
Fig. 3 Language model for voice-operated projector.

ングした．これをポーズ（0.25 秒以上）で区切って，講演音声のサンプルを得た．これとは別に，同一の話者に音声操作プロジェクタ用のコマンドを発声してもらった．コマンドを講演とは別に発声してもらったのは，できるだけ多くの種類のコマンドのサンプルを多数収集するためである．この結果，講演音声 646 サンプル，コマンド音声 199 サンプルを得た．なお話者とプレゼンテーションの話題は，フィラーモデル学習用のコーパスとは異なっている．

本節の実験では，コマンドの前にマジックワードをおいていない．この場合，“次”，“前” といった 2 音節程度の音声に対しても判定を行うことになるので，かなり難しいタスクである．

これらの音声データに対して，コマンド文法に基づく認識を行った．同時に発話検証用モデルによる尤度を計算した．比較のために，2 種類のフィラーモデル（filler 70 語/230 語）に加えて，毎日新聞記事データ（91～94 年）における 3 音節以上からなる上位頻出語（newspaper 70 語/230 語/5000 語），および音節連接モデル（syllable）を用いた．参考までに，認識結果の尤度の絶対値から判定する方法も試みた．音響モデルには音素環境独立 HMM（16 混合）を利用しており，コマンドとフィラーモデルの両方に共通のサブワードモデルとなっている．

尤度差（$\log LR$）に対するしきい値を設定することにより，コマンド音声に対しては誤棄却（False Rejection: FR），講演音声に対しては誤受理（False Acceptance: FA）の誤り率が算出される．

FA=1%および FA=2%のときの誤棄却（FR）の割合を**表 3** に示す．音声操作プロジェクタでは，講演音声で誤動作する方が，コマンド音声で正しく動作しない場合よりも重大なエラーであると考えられるので，FA が小さい範囲のこれらの値に着目した．

発話検証用モデルをまったく用いないで，認識尤度のみで検証を行う方法は，ほとんど機能していない．典型的な従来手法である音節連接モデルを用いる方法では，かなり改善されるが，FA=1%において FR=16.1%であり，これは講演の切れ目の 100 個中に

表 3 特定の誤受理率における誤棄却率（マジックワードなし）
Table 3 FR rate at specific FA rate (without magic word).

発話検証用モデル	誤棄却率 FR @FA=1%	誤棄却率 FR @FA=2%
なし（認識尤度のみ）	66.8%	53.8%
音節連接モデル	16.1%	12.1%
新聞記事上位 70 語	8.5%	6.5%
新聞記事上位 230 語	10.6%	5.0%
新聞記事上位 5000 語	20.1%	10.6%
フィラーモデル 70 語	3.5%	1.0%
フィラーモデル 230 語	2.5%	1.0%

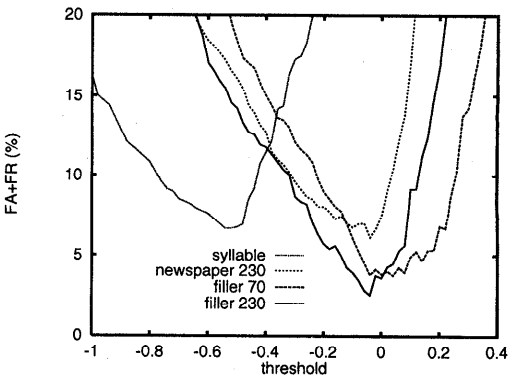


図 4 誤受理率+誤棄却率（マジックワードなし）
Fig. 4 FA rate + FR rate (without magic word).

1 回の誤動作を許容しても，ほぼ 6 回のコマンド入力のうち 1 回は認識されないことを意味し，音声操作プロジェクタとして実用的なレベルといえない．また，新聞記事の語彙を用いる手法は，音節連接モデルに比べて若干良い程度である．これは，新聞記事の頻出語が講演においてあまり用いられず，良いモデルになっていないためである．なお，表 3 において，語彙を大きくするほど FR が増加しているが，FA=3%以上の区間においては逆転しており，(本アプリケーションに限らない）一般的な意味で，語彙が大きい方が性能が悪いというわけではない．これに対して，提案するフィラーモデルによって，誤りを大幅に削減することができた．この場合は，230 語のモデルの方が 70 語のモデルより有効であった．

このうちの代表的なものについて，しきい値の変化に対する 2 種の誤り率の和（FA+FR）を**図 4** に示す．

音節連接モデルおよび新聞記事の語彙を用いる手法に比べてフィラーモデルの方が，誤りがかなり少ないことが分かる．230 語のフィラーモデルが最小の誤り率（の和）2.5%を実現した．なお，語彙のモデルを用いた場合は最適なしきい値が 0 付近になるが，音節連接モデルでは単語モデルより必ず認識尤度が良くなる

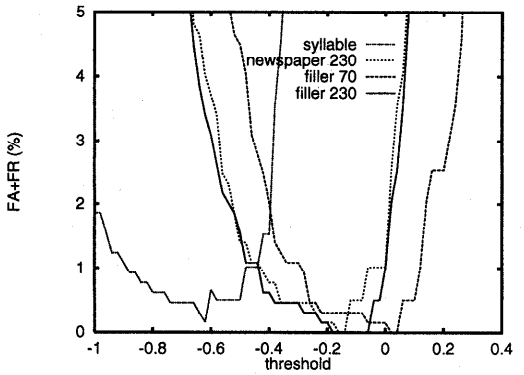


図5 誤受理率+誤棄却率（マジックワードあり）
Fig. 5 FA rate + FR rate (with magic word).

ために，0よりかなり小さいしきい値（=ペナルティ値）を設定する必要があり，その調整は容易でない．

これらより，提案するフィラーモデルを用いた発話検証手法が最も効果的であり，またそのサイズも大きい方が良いことが示された．

4.3 マジックワードの効果

次に，コマンドの直前にマジックワード（“オペレータ”）を含めて発声されたサンプルに対して実験を行った．話者は前節の実験と同一で，コマンド音声のサンプルは197個である．講演音声のサンプルは前節の実験と同一である．

その結果，どのような発話検証用モデルを用いても，しきい値を調整すれば2種の誤り率の和（FA+FR）が0.2%未満になり，マジックワードの効果が絶大であることが分かった．しかしながら図5において，しきい値の変動に対する誤り率の変動（=頑健性）を調べると，提案するフィラーモデルを用いる場合が，誤り率が0となる範囲が最も大きく，すなわち最も安定した発話検証が実現されることが確認できる．

なお，コマンド音声自身の誤認識は5.6%あったが，そのほとんどが，“いち”，“ひち”，“しち”といった数字の誤りに起因するものであった．

4.4 動 作 例

以上の結果より，提案手法が命令音声とそれ以外の講演音声の識別に非常に有効であることが示されたので，これに基づいて音声操作プロジェクタを実装した．マジックワードの使用は識別精度の点では有効であるが，ユーザによっては抵抗感を感じる場合があるので，発声してもしなくてもよい仕様となっている．

音声は，Silicon Graphics社のO2あるいはIndyワークステーションのマイク入力を用いて，16 kHz, 16 bitで入力される．音声認識およびスライド操作も同ワークステーション上で動作し，指示からスライド切替えまで1〜2秒を要するが，実際の使用には支障はない．

前節までの実験で用いたサンプルの話者とは異なる男性2名に，それぞれ約11分，約12分のプレゼンテーションを自由に行ってもらった．その結果，コマンド音声が認識されずに発声し直す場面は数回あったものの，プレゼンテーションの音声区間をコマンドとして誤識別し誤動作を起こすことはなかった．コマンド音声をポーズに基づいて検出しているが，その際の誤りはなかった．切り出し誤りが生じた場合，コマンドとして認識されずに棄却されるので，ユーザは発声し直すことになる．なお，実際のプレゼンテーションで用いられた命令の大多数は，1枚次のスライドに切り替える指示（“次”）であった．

5. 講演の自動ハイパーテキスト化システム

本章では，音声操作プロジェクタを利用して，講演の自動ハイパーテキスト化を行うシステムについて述べる．

5.1 講演音声の自動インデキシング

近年の計算機・メディア技術の発展にともない，大容量のマルチメディアデータをオンラインデータとして記憶することが可能になってきた．たとえば，大学で行われているすべての講義音声をランダムアクセス可能なディジタルメディアに記憶することも不可能でない．しかしながら，このような音声データが蓄積されても，適切なインデックスが付与されていないと，膨大なデータから望みのものを検索することが困難であり，有意義なアーカイブとはいえない．人手でインデックスを付与するのは大変な労力を要するし，テキストデータと異なり，音声データの自動インデキシングは容易でない．多くのビデオ・オン・デマンドシステムにおいても，講演のタイトルからそのビデオを取り出せても，そこから特定の説明等の区間を探すには，そのビデオをユーザ自身が先頭からすべて走査しなければならない．

近年，ニュース音声[12)〜14)]やボイスメール[15)]を対象として，自動インデキシングの研究が行われている．これらは主として，音声メディアに対して大語彙連続音声認識やワードスポッティングを適用して，キーワードを認識することを基本としている．したがって，タスクドメインに依存した語彙や言語モデルを仮定しており，また性能面からも実用レベルではない．

講演は，ニュース等に比べて話題が特定的で，説明が長時間かつ詳細である．すなわち1つの話題について，少なくとも5〜10分程度の説明を行うのが普通

である．そのため，このような長い単位でないと話題を同定できないし，厳密に話題の区切りを検出するのが困難である．我々は以前，音声を10秒程度のセグメントに区切って話題の変化を求める手法を検討したが[16]，予稿テキストから抽出したキーワードを用いてもインデキシングの性能は十分ではなかった．

実際に講演音声を書き起こすと，話し言葉特有の種々の現象が観察される[17]ために，ニュース音声等と比較して，話題を示すキーワードはそれほど頻繁に出現しない．また，話し言葉に対する音声認識が研究途上であるので，現段階で単純に音声認識に基づいて自動インデキシングを行うのは非常に困難である．

そこで本研究では，音声データの中身からではなく，講演で使用されるスライドを用いて話題の区分化を行うことを考える．近年，オンラインのスライドを用いる講演が多くなってきた．また，スライドをワープロソフト等で作成する場合が多いことから，潜在的にオンラインスライドを使用できる講演の割合は大きいと予想される．

スライドには表題がつけられていることが一般的であるので，これをそのままインデックスとすることが妥当である．また前述のとおり，講演音声においてはスライド1枚分に対応する単位（1〜5分程度）より細かいインデックスを付与することにはあまり意味がないと考えられる．

したがって，講演で使用された各スライドに講演音声の区間を対応づけることにより，自動インデキシングを実現する．

このスライドと音声の対応づけを，音声操作プロジェクタにより実現する．音声コマンド以外でスライドの切替えを行った場合，講師はスライド交換の途中や前後で'つなぎ語'を用いて話すことが多いので，機械的に区切りを入れることが容易でない．実際に講演音声を保存したうえで，単純にポーズ等の情報に基づいて区分化を行うと，文や文節の途中での区切りが多数生じ，結果として不自然な再生となってしまう[18]．またOHPのスクリーンをビデオカメラで撮影して，スライドの切替えを画像情報に基づいて自動検出することも試みたが，処理量が大きいうえに，スライドを切り替えたのか位置をずらしたのかの判別が困難であった．音声操作プロジェクタを用いて音声コマンドで切替えを行うことにより，そこで講師の説明が中断され，スライドとそれに対応する説明音声の同期が保証される．

5.2 対象とする講演

具体的に，自動ハイパーテキスト化の対象とする講演を以下のように定める．

まず，インデキシングおよび検索の観点から次の3点が条件となる．

- 講演で使用されるスライドはオンラインテキストに限る．
- 講演が主として，スライドを用いた説明によりなされる．黒板やビデオ等のメディアは補助的に使用してもよいが，スライドのみで講演の筋が構成されるものとする．
- スライドすべてに表題がついており，かつそれらの目次が用意される．ただし，目次が自動的に抽出されるソフトウェアもあるし，人手で用意するのもそれほど面倒でない．

さらに，話題の区切りを適切に入れるために以下の制約を課す．1つの話題は，1枚もしくは複数枚のスライドから構成される．なお，以下は自然な設定であると考えられるが，絶対的な条件ではない．

- 同一話題のスライドが複数枚にわたる場合は，それらが同一の話題であることを表題と目次から知ることができる．
- スライド1枚が複数の話題を持たない．話題が変わればスライドも替わるようにする．

5.3 自動ハイパーテキスト化システムの構成

以上の条件を満たす講演を対象とした自動ハイパーテキスト化システムの処理の流れを**図6**に示す．

本システムは，講師の音声およびオンラインスライドとその目次を入力とし，これらを対応・関連づけたひとまとまりのハイパーテキストとして出力する．

本システムは，音声操作プロジェクタによる音声の識別・区分化と，音声とテキストを対応づける処理から構成され，以下のような手順で実行される．

(1) 入力音声区間が，講演の一部であるかプロジェクタ操作コマンドであるかを識別する．

(2) 講演の一部であると識別されれば，その音声をファイルとして保存する．プロジェクタ操作コマンドとして認識されれば，スライドの切替えを行うとともに，講演の区切りであるインデックスを付与し，それまでに区分化・保存された（そのスライドに対応する）講演音声のファイルをまとめて対応づける．

(3) スライドテキストとそれに対応づけられた音声ファイルをハイパーテキストとしてリンクする．

(4) スライドの目次に従って，個々のスライドをハイパーテキストとしてリンクする．また，1つの話題が複数枚のスライドにわたっている場合は，ハイパーテキスト上では1つにまとめる．

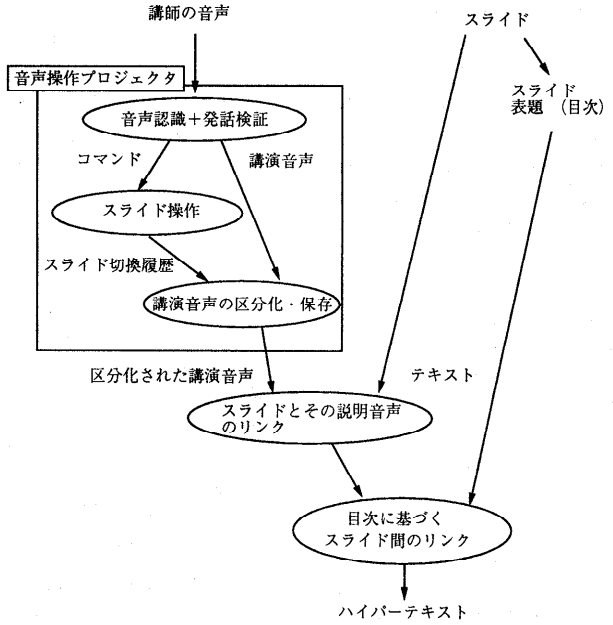


図 6 講演の自動ハイパーテキスト化システムの処理の流れ
Fig. 6 Flowchart of hyper-text generation of lectures.

ハイパーテキストとしてアーカイブ化された講演については，まず目次によってその概要を知ることができ，そこから興味を持ったスライドを取り出して，さらに説明が聞きたい際には音声を再生することができるようになる．

本システムを Netscape ブラウザを利用して実装し，動作を確認している．なお，元のスライドが HTML 形式でなくても，それに自動変換できる場合が多い．本システムは，本学での実際の講義において現在試験的に運用しているところである．

6. おわりに

本論文では，話題や語彙等に独立で，使用される状況の話し言葉の特徴に依存した発話検証用モデルを提案した．本モデルは，講演の書き起こしテキストから自動的に構築され，従来の発話検証手法と比べてはるかに高い性能（誤受理率 FA=1%において誤棄却率 FR=2.5%）を得た．これにより，音声操作プロジェクタを作成し，安定に動作させることができた．さらに，このプロジェクタにより自動的に講演音声とスライドの対応づけができることを利用して，講演の自動ハイパーテキスト化を行うシステムを作成した．

今後は，実際の講義において運用・評価を進めながら，講義室等での種々の機器操作にも適用していく予定である．また，本研究では主に音声情報しか保存していないが，講演の再現性においては画像情報も重要であるので，スライド以外に講演の際に使用される黒板やビデオ等のメディア，さらには講演中の講師や聴衆の映像等についても，統合的にアーカイブ化を行えるようにしていく予定である．

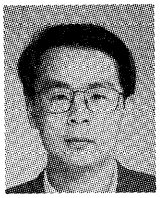
謝辞 講演・講義の自動アーカイブ化に関して，日頃より議論をしていただく京都大学美濃研究室，池田研究室および堂下研究室の皆様に感謝いたします．なお，音響モデルと新聞記事データの単語辞書は，「日本語ディクテーション基本ソフトウェア 97」[19] を利用した．

参考文献

1) 嵯峨山茂樹：音声認識技術実用への課題，情報処理，Vol.36, No.11, pp.1047–1053 (1995).
2) 河原達也，石塚健太郎，堂下修司：話し言葉依存の競合言語モデルを用いたキーフレーズの検出・検証，日本音響学会研究発表会講演論文集，2-1-19 (1997).
3) Kawahara, T., Doshita, S. and Lee, C.H.: Speaking-Style Dependent Lexicalized Filler Model for Key-Phrase Detection and Verification, 情報処理学会研究報告，97-SLP-19-11 (1997).
4) Asadi, A., Schwartz, R. and Makhoul, J.: Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System, *Proc. IEEE-ICASSP*, pp.305–308 (1991).

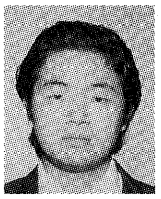
5) 渡辺隆夫，塚田 聡：音節認識を用いたゆう度補正による未知発話のリジェクション，信学論，Vol.J75-DII, No.12, pp.2002–2009 (1992).
6) Sukkar, R.A. and Lee, C.-H.: Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword based Speech Recognition, *IEEE Trans. Speech & Audio Process.*, Vol.4, No.6, pp.420–429 (1996).
7) Kawahara, T., Lee, C.-H. and Juang, B.-H.: Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification, *IEEE Trans. Speech & Audio Process.*, Vol.6, No.6, pp.558–568 (1998).
8) Rohlicek, J.R., Jeanrenaud, P., Ng, K., Gish, H., Musicus, B. and Siu, M.: Phonetic Training and Language Modeling for Word Spotting, *Proc. IEEE-ICASSP*, Vol.2, pp.459–462 (1993).
9) Weintraub, M.: Keyword-Spotting using SRI's DECIPHER Large-Vocabulary Speech-Recognition System, *Proc. IEEE-ICASSP*, Vol.2, pp.463–466 (1993).
10) 河原達也，宗続敏彦，堂下修司：ヒューリスティックな言語モデルを用いた会話音声中の単語スポッティング，信学論，Vol.J78-DII, No.7, pp.1013–1020 (1995).
11) 三木清一，河原達也，堂下修司：タスクの構文的制約から逸脱した発話のリジェクション，日本音響学会研究発表会講演論文集，1-Q-1 (1994).
12) Wactlar, H.D., Hauptmann, A.G. and Witbrock, M.J.: Informedia: News-On-Demand Experiments in Speech Recognition, *Proc. DARPA Speech Recognition Workshop* (1996).
13) James, D.A.: A System For Unrestricted Topic Retrieval From Radio News Broadcasts, *Proc. IEEE-ICASSP*, pp.279–282 (1996).
14) 横井謙太朗，河原達也，堂下修司：キーワードスポッティングに基づくニュース音声の話題同定，情報処理学会研究報告，95-SLP-6-3 (1995).
15) Jones, G.J.F., Foote, J.T., Jones, K.S. and Young, S.J.: Video Mail Retrieval: The Effect of Word Spotting Accuracy on Precision, *Proc. IEEE-ICASSP*, pp.309–312 (1995).
16) 三村正人，河原達也，堂下修司：パネル討論音声の話者と話題に関する自動インデキシングの検討，情報処理学会研究報告，96-SLP-11-3 (1996).
17) 峯松信明，片岡嘉孝，中川聖一：講演調の話し言葉に対する言語的解析，情報処理学会研究報告，95-SLP-8-7 (1995).
18) 野村和弘，河原達也，堂下修司：講義の自動アーカイブ化のための韻律情報を用いた講義音声の文境界の抽出，信学技報，SP98-80 (1998).
19) 河原達也，李 晃伸，小林哲則，武田一哉，峯松信明，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（97年度版）の性能評価，情報処理学会研究報告，98-SLP-21-10 (1998).

(平成10年9月30日受付)
(平成11年2月 8 日採録)



河原 達也（正会員）

1987年京都大学工学部情報工学科卒業．1989年同大学院修士課程修了．1990年同博士後期課程退学．同年同大学工学部助手．1995年同助教授．現在，同大学情報学研究科助教授．1995年から96年まで米国ベル研究所客員研究員．1998年からATR音声翻訳通信研究所客員研究員．音声認識・理解の研究に従事．京都大学博士（工学）．1997年度日本音響学会粟屋賞受賞．電子情報通信学会，日本音響学会，人工知能学会，IEEE各会員．



石塚健太郎

1997年京都大学工学部情報工学科卒業．現在，同大学院情報学研究科修士課程在籍．音声・映像等のマルチメディア情報処理の研究に従事．

堂下 修司（正会員）

1958年京都大学工学部電子工学科卒業．1960年同大学院修士課程修了．1963年同博士課程単位取得退学．同年同大学工学部助手．1965年同助教授．1968年東京工業大学理工学部助教授．1973年京都大学工学部教授．現在，同大学情報学研究科教授．1996年から同大学大型計算機センター長（併任）．その間，音声の分析と認識，オートマトンの学習的構成，自然言語処理，人工知能等知的情報処理の研究・教育に従事．京都大学工学博士．1959年通信学会稲田賞受賞．1988年人工知能学会論文賞受賞．1990年情報処理学会創立30周年記念論文賞受賞．人工知能学会元会長．電子情報通信学会，日本音響学会等会員．