自動アーカイブ化のための講義音声の区分化

4U－5

河原達也　野村和弘　堂下修司　(京都大学・情報学研究科)
飯塚重善　辻本雅彦　(NTT 情報通信研究所)

1　はじめに

計算機・メディア技術の発展に伴い、大学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可能になってきた。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。

本稿ではまず、講義音声を講義で使用されるスライドと対応づけることによりアーカイブ化を行う方式について説明し、さらに韻律情報を利用して講義音声を文単位に区分化する方法について検討する。

2　講義音声のスライドとの対応づけ

第一段階として、講義で使用されるスライドを利用して、講義音声の大まかな区分化及びインデキシングを行う。

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。そこで、講義で使用された各スライドに講義音声の区間を対応づけることにより、講義音声の自動インデキシングを実現する。

NTT の教材作成支援システム CALAT では、講義中において、講師が (マウスかキーボードで) スライド表示ソフト上のスライドを切り替えたイベントを検出・記憶しておくことで、このスライドと音声の対応づけを行っている。

さらに、京都大学で開発された音声操作プロジェクタを利用すれば、講師は (講義室のどこからでも) 講義に使用しているマイクロフォンでスライドの切替え指示を送ることができ、かつこのスライド切替えイベントを自動的に保存することが可能になる [1]。

オンラインスライドを用いた講義を対象とした自動アーカイブ化システムの処理の流れを図 1 に示す。オンラインスライドを用いる講義は近年多くなっており、今後も増加するものと予想される。

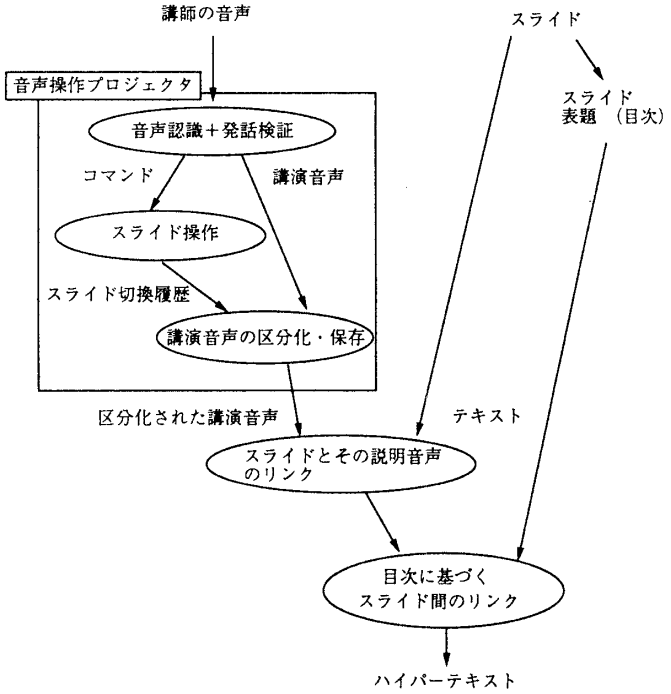


図 1: 講義音声とスライドの自動アーカイブ化

音声操作プロジェクタは、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 ブラウザ及び xdvi(LaTeX) プレビューアが利用できる。

講師は同一のマイクロフォンで講義を行いながら、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、有限状態文法で記述されている。音声認識エンジンには京都大学で開発された JULIAN[2] を用いている。

講義音声の区間でスライド表示ソフトが誤動作しないように、音声認識結果の検証を行っている。このために、多くの講演の書き起こしテキストから学習した講演調スタイルの言語モデルを使用している。

なお音声の切り出しのために、コマンドの前後にポーズを入れることを前提としている。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 (“オペレータ”) を発声する仕様も考えた。

評価実験において、マジックワードを用いることにより誤り率 (=誤受理率+誤棄却率) は 0、マジックワードを用いなくても誤り率は 3%程度となっている [3]。

Segmentation of Lecture Speech for Automatic Archiving
T.Kawahara, K.Nomura, S.Doshita (Kyoto Univ.) and
S.Iizuka, M.Tsujimoto (NTT)

3 講義音声における文境界の抽出

前章の処理により、講義音声はスライド単位に区分化されるが、1枚のスライドの説明音声は通常1〜5分と長い。アーカイブから音声を再生する際の自然性や利便性を考えると、この音声がさらに文単位などに適切に区分化されていることが望ましい。

そこで、音声の韻律的特徴、具体的にはポーズ長やピッチの変化などを手がかりに、講義音声の文境界を抽出することを検討する [4]。

講義音声は言い淀みや間投語を多く伴うかなり自由な発声であるので、単純にポーズ長のみ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。

そこで、ポーズ長以外の韻律情報としてピッチパターンに着目する。一般に日本語の平叙文の発話について、ピッチが文の始端では高く、終端では低下することが知られているので、この特徴を利用して文境界の抽出を試みる。ピッチパターンはアクセントの影響も受けるが、十分に長い範囲でその平均値や回帰直線を求めることにより、より大局的な特徴をとらえる。

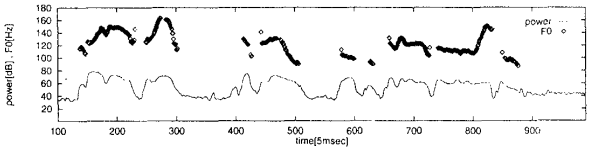
自己相関係数のピークを用いて基本周波数 $F0$ を抽出し、倍ピッチ/半ピッチの補正、孤立点消去などを行った結果得られるピッチパターンの例を図 2(a) に示す。さらに、十分に長いポーズ (=ピッチが抽出されない区間) で区切られた区間 (=文境界の候補) について、単回帰直線による近似を行った結果を図 2(b) に示す。

これからポーズ前のピッチの終端値と平均値、ポーズ後のピッチの始端値と平均値に基づいて判別関数を構成し、そのポーズが文境界であるか判定する。

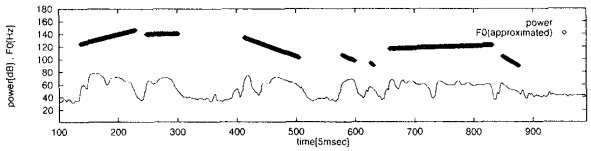
実際の講義から約 30 分の区間を取り出して、文境界の抽出を試みた。正解の文境界は 69 箇所 (+スライド切替えは 6 箇所) であり、ポーズに基づく境界候補の数は 835 であった。抽出された文境界の数と再現率・適合率を表 1 に示す。誤抽出を正解と同程度許容する (適合率 1/2) 場合で、およそ正解の 2/3 が抽出されている。誤抽出には、文境界 (句点) ではないものの、実際に聴取してみた場合には区切りとして自然なものが多く含まれており、アーカイブとしては十分な精度であると考えられる。ただし、間投語 ("あのー" 等) や咳払いの直後で誤って区切られた例が多数あったので、これについては対策が必要である。

4 おわりに

講義の中で主に音声情報を対象として、自動アーカイブ化の方式を設計・実装した。音声操作プロジェクタを用いたアーカイブ化システムは、現在実際の講義において試験的に運用しているが、今後評価を行いながら、黒板や講師・聴衆の映像等の画像情報との統合を進めていく予定である。



(a) ピッチパターン



(b) ピッチパターンの回帰直線による近似

図 2: 音声の例 "バイグラムというのは、ひとつ前の、直前の単語のみを見る"

表 1: 文境界の抽出結果

抽出数	再現率 (%)	適合率 (%)	平均時間 (sec.)
34	30.4	61.7	34.0
43	37.6	60.4	26.9
50	44.9	62.0	23.1
66	52.1	54.5	17.5
77	59.4	53.2	15.0
96	66.6	47.9	12.0
113	73.9	45.1	10.2

謝辞： 日頃より議論をして頂く京都大学 美濃研究室・池田研究室の皆様に感謝します。

参考文献

[1] 石塚健太郎, 河原達也, 堂下修司. 音声操作プロジェクタを用いた講義音声・テキストの ハイパーメディア化. 人工知能学会研究会資料, SIG-J-9701-1, 1997.

[2] 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. 電子情報通信学会技術研究報告, SP98-110, NLC98-46 (98-SLP-24-15), 1998.

[3] 河原達也, 石塚健太郎, 堂下修司. 音声操作プロジェクタのためのドメイン独立フィラーモデルの評価. 日本音響学会研究発表会講演論文集, 1-1-12, 秋季 1998.

[4] 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. 電子情報通信学会技術研究報告, SP98-80, 1998.