音声言語情報処理 43－7
（２００２．１０．２５）

表層情報と韻律情報を利用した講演音声の要約

小林 聡, 吉川 裕規, 中川 聖一

豊橋技術科学大学
〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘１－１
e-mail; skoba@cc.tut.ac.jp, {yoshikawa,nakagawa}@slp.ics.tut.ac.jp

あらまし: 音声情報は記録が容易であるが、後の参照は必ずしも容易ではない。音声情報をインデクス化したりそのまま要約することが可能となれば、音声情報の参照も容易かつ簡便なものとなる。本研究では、音声の自動抽出要約を目指し、まずは人間による重要文抽出結果の比較を行なった。次いで表層情報による重要文抽出結果と人間による結果との比較を行ない、音声の抽出要約に有用と思われる表層情報を得た。ここで得た表層情報を利用し、実際に抽出要約音声を作成し、聴取実験を行なった。また、韻律情報として F_0 とパワーに着目し、同様に抽出要約音声を作成し、聴取実験を行なった。いずれの場合も人間による結果にはやや劣り、今後の検討を要する。

キーワード: 講演音声、要約、表層情報、韻律情報、手がかり語、ポーズ

Extracting Summarization of Lectures Based on Linguistic Surface and Prosodic Information

Satoshi KOBAYASHI, Noriki YOSHIKAWA, Seiichi NAKAGAWA

Toyohashi University of Technology
Hibarigaoka 1-1, Tenpaku, Toyohashi, Aichi 441-8580
e-mail; skoba@cc.tut.ac.jp, {yoshikawa,nakagawa}@slp.ics.tut.ac.jp

Abstract: It is easy to record speech, but it is not easy to refer to audio recordings. If it is usable to index or summarize speech, refering audio recordings will become easier. In this paper, we aim for automatic extracting summarization of lectures. For this purpose, at first we compared results of extracting summarization by human subjects. Then we investigate relations between linguistic surface information and human results and we got the useful surface information. Next, we made summarized audios based on this information, and we compared them with human results. Additionaly, we focused on prosodic features, F_0 and power. We did same expariments on them. As the result, both of them behinded human results. We need furthur research.

Keywords: Lectured Speech, Summarization、Linguistic Surface Information, Prosodic Information, Cue Word, Pause

1　はじめに

音声情報は、その記録は容易であるが、後の参照は必ずしも容易ではなく、そのためにはインデクス化や文書化しておく必要がある [1, 2]。音声情報をインデクス化したりそのまま要約することが可能となれば、音声情報への参照も容易かつ簡便なものとなる。従来テキストからの自動的な重要文抽出や要約技術の研究は行なわれているが [3, 4, 5]、音声を対象としたものは少ない。

講演音声の効率的利用に関しては、長谷川らによる研究 [1] がある。これは、談話標識となる単語を自動的に抽出するものであるが、文境界の検出などに多量の学習データを必要とし、またインデキシングされるのみであるため講演内容の概略の把握が難しい。

ニュース音声の自動要約に関しては堀らの研究 [6, 7] が有る。自動書き起こし結果の比較的精度の良かった音声を対象としているため、より広汎な音声を対象とすることは考慮されていない。

Waibel らは、disfluency の検出と除去、文の境界検出、質問-応答ペアの検出などを用いた会議の議事録の自動作成について報告している [8]。Reithinger らはホテル予約などのタスクに対話において統計的な dialogue act 推定を用いて、自動要約について報告している [9]。Koumpis らは、トピックに応じた単語や固有名詞、日時の表現、単語の持続時間が有用な情報であると報告している [10]

また、笠原や三上らにより、韻律的特徴と重要文との関係が検討されており [11, 12]、文単位ではパワーに関して文の重要度とやや強い相関が報告されているが、韻律情報だけからの要約は難しい。そこで、井上らは、ピッチ情報と言語情報を用い、言語情報の利用により要約精度が向上することを示している [13]。

日高らは、声の高さ、強さ、速さを用いた強調個所の自動抽出により、音声要約を試み。会議音声を対象として良い結果が得られたと報告している [14]。

我々もまた、対話音声における F_0、パワー、発話速度について、キーワード、述部、文の中央における特徴を調査し、キーワード部では F_0 とパワーが相対的に高く、大きくなるという結果を得ている [16]。

現在、講演音声のコーパスが利用可能となり、講演音声という自発発話に近い音声を対象とする研究も道が開けた。そこで、本研究では、まず「話し言葉工学」プロジェクトコーパスの講演音声を、表層情報と韻律情報を用いて自動的に要約し、人手による要約と比較した。

以下 2 節では、実験に利用した試料や実験条件について述べる。3 節では、表層情報を利用した自動的要約結果の評価実験について述べる。4 節では、韻律情報を利用した自動的要約結果の評価実験について述べる。5 節では、まとめとともに研究の展望について述べる。

2　音声試料と前処理

はじめに、本研究で利用した音声試料等について説明する。本研究では「話し言葉工学」プロジェクトにより提供されている日本語話し言葉コーパスの内、特に日本音響学会における 5 つの講演音声 (話者 5 名) を対象として実験を行なった。表 1 に貢献音声の諸元を記す。講演の書き起こしの例を図 1 に挙げる。

人による重要文抽出実験の被験者は、対象とする講演の内容を理解できる、音声研究の専門家 5 名とした。

なお、以下で用いる「文」の認定についてはコーパスの書き起こしテキストに従った (原則として 200msec. 以上のポーズにより分割)。

各被験者は、対象とする講演について、書き起こしを元に重要文を抽出する要約、および聴取により重要文を抽出する要約を行なった。この際、講演全体の 1/3 程度の分量になるように指示した。そのため、被験者により抽出文数は異なる。

3　被験者間での重要文の一致度

各被験者間において、抽出された重要文の一致の程度を調査した。対比較による一致数、重要文として選択された文数に対する一致した文数の割合 (%)、そしてカッパ (κ) 値 [4] を表 2 に示す。ここで、一致率と κ 値は次式により定義される。

$$\text{一致率} = \frac{\text{一致した重要文数} * 2}{\text{被験者 1 による重要文数} + \text{被験者 2 による重要文数}}$$

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$$P(A) = \frac{\text{重要文の一致数} + \text{非重要文の一致数}}{\text{文の総数}}$$

$$P(B) = \text{重要文の偶然一致率} + \text{非重要文の偶然一致率}$$

$$\text{重要文の偶然の一致率} = \frac{\text{被験者 1 が重要と判定した文数}}{\text{文の総数}} \times \frac{\text{被験者 2 が重要と判定した文数}}{\text{文の総数}}$$

$$\text{非重要文の偶然の一致率} = \frac{\text{被験者 1 が非重要と判定した文数}}{\text{文の総数}} \times \frac{\text{被験者 2 が非重要と判定した文数}}{\text{文の総数}}$$

なお κ 値については $0.4 < \kappa < 0.6$ を fair、$0.6 < \kappa < 0.75$ を good、$0.75 < \kappa$ を excellent とする指標が提案されている [17]。また、$0.7 < \kappa$ を信頼性の目安とするものもある [18]。

被験者間では、新聞やニュースを対象とした結果 [4] と比べて、平均で、一致率が 52%程度かつ κ 値が 0.28 程度と低い値となっている。書き起こしテキストからの抽出要約と、音声の聞き取りによる抽出要約との差異は見られない。これより、韻律情報による重要文抽出の効果は小さいと推察される。被験者間での一致率が低くなったのは、約 13 分間に限られた時間での研究発表という講演では、冗長な個所が少なく、内容が高度に専門的であったため、どこが重要かは被験者によってかなり違うことによる。

表 1: 講演音声の諸元

講演音声	文の総数	時間長	長いポーズの数	手がかり語の数
講演 A: a01m0037	194	13'35"	10	24
講演 B: a01m0086	314	14'04"	11	15
講演 C: a01m0088	281	14'02"	25	19
講演 D: a01m0124	185	13'14"	7	12
講演 E: a01m0157	333	11'22"	5	15
平均	261.4	13'15"	11.6	17

ところでもって (F えー)
(F んー) 限界があると (F ま) そういう欠点がありますでまた (F あ) 少量の学習データの場合でありましても
(F えー) 次元 (F あのー) 回帰行列の次元以下になりますとやはり不良条件になるって (F ま) そういった (D て) 欠点になる訳ですでこの発表では (F えー) ノルムを最小にするという拘束条件付きで
(F えー) 解を求めるそういう (F えー) 検討をした結果を (F えー) 発表いたします。

図 1: 書き起こしの例

また書き起こしを対象とした重要文抽出と音声の聴き取りによる重要文抽出の被験者ごとの平均一致率は 53.5%、κ 値は 0.34 であった。

4 表層情報を利用した自動要約

4.1 半自動化手法

表層情報としては、講演の最初の 10 文、最後の 10 文、長いポーズの前後 5 文、手がかり語を含む前語 6 文を用いた評価を行なった [19]。長いポーズと手がかり語の講演当りの出現頻度は表 1 に示す。長いポーズの判定は聴取により行なったが、概ね 1 秒以上のポーズである。手がかり語の例を図 2 に示す。実際には、手がかり語によってその手がかり語の前後のどちらに重要文があるかは異なると考えられるが、今回は便宜上、一括して扱った。

これら表層的な情報によって抽出した文に対して、被験者が重要と判断した文の一致の割合を表 3 に示す。被験者がランダムに重要文を抽出すると偶然の一致率は 1/3 程度であるが、講演の最後 10 文、長いポーズの前 5 文、手がかり語の前 6 文が、被験者の抽出した重要文との一致率が 40% 強という結果が得られた

表 2: 被験者間での一致度

		書き起こしを対象			音声を対象		
講演	一致尺度	最低	最高	平均	最低	最高	平均
a01m 0037	文数	33	51	43	45	37	35.1
	一致率 [%]	46.5	60.4	57.8	68.2	58.8	55.6
	κ 値	.09	.49	.33	-.62	.39	.21
a01m 0086	文数	49	48	44	37	49	39.8
	一致率 [%]	46.2	29.4	47.6	39.0	56.6	45.0
	κ 値	.03	.27	.26	.12	.40	.23
a01m 0088	文数	67	51	48.8	49	43	44.8
	一致率 [%]	59.2	56.4	53.0	56.4	51.4	51.6
	κ 値	.27	.28	.24	.28	.21	.18
a01m 0124	文数	40	38	31.5	29	50	38.1
	一致率 [%]	49.4	62.2	48.8	44.2	64.2	54.4
	κ 値	-.09	.44	.22	.15	.38	.30
a01m 0157	文数	77	59	59	49	57	49.9
	一致率 [%]	58.6	63.2	58.0	54.2	62.6	57.8
	κ 値	.14	.49	.38	.37	.49	.44
平均	文数	53.2	49.4	45.3	49	57	41.5
	一致率 [%]	52.0	58.4	53.0	52.4	58.6	52.8
	κ 値	.09	.39	.28	.06	.37	.27

この結果を元に、講演の最後の 10 文、長いポーズの前 5 文、手がかり語の前 6 文を表層情報として利用する。この表層情報を用い、半自動的に要約音声を作成した (**半自動化手法**)。この際、フィラーは除去している。表 4 に、抽出された音声の長さや、人手による抽出要約との κ 値を示す。なお、人手による重要文の抽出結果は、被験者 5 人中 3 人以上が重要と判断した文とした。

半自動的手法による抽出要約と、人手による結果に基づいた抽出要約を、聴取により比較実験を行なった。基準としては、「聴き易さ」および「良い要約かどうか」を求めた。被験者は、重要文抽出者と異なる、音声関係に従事している大学院生 5 名である。結果を表 5 に示す。

聴き易さについては、「どちらとも言えない〜半自動化手法が良い」という結果になり、良い要約か

表 3: 表層情報による抽出結果と被験者の結果との一致率

	講演の最初と最後		長いポーズの前後		手がかり語の前後	
講演	最初の 10 文	最後の 10 文	前 5 文	後 5 文	前 6 文	後 6 文
講演 A: a01m0037	23%	41%	39.9%	33.0%	37.2%	37.7%
講演 B: a01m0086	34%	47%	34.8%	24.8%	42.8%	31.7%
講演 C: a01m0088	22%	35%	36.7%	29.9%	43.1%	34.9%
講演 D: a01m0124	47%	50%	46.6%	47.4%	43.6%	48.2%
講演 E: a01m0157	9%	59%	50.8%	31.2%	39.8%	29.3%
平均	27%	46%	41.8	33.3	41.3	36.4

強調したい
考えております
評価結果です
まとめさせていただきます
得ました
以上です
今回の背景の説明
ポイントになります
発表します。
ご報告いたします
注目しており
課題になりまして
結果は

図 2: 手がかり語の例

表 4: 半自動化手法による音声データ

	講演 B	講演 D	講演 E
文の総数	314	185	333
人間による抽出文数	77	52	60
半自動による抽出文数	98	72	84
一致文数	34	20	33
人間による抽出区間の長さ	3'15"	3'56"	2'21"
半自動による抽出区間の長さ	3'05"	3'45"	2'35"
κ 値	0.157	-0.006	0.314

どうかについては「どちらとも言えない」という結果になった。これより、音声の自動要約の可能性が示された。

4.2 自動化手法

手がかり語を音声認識 [15] によって得た場合の結果を示す。今回の実験では、手がかり語の認識率を向上させるため言語モデルの中の手がかり語に関する 1-gram、2-gram、3-gram の値を 3 倍にしている。これにより、全体の精度は下がるものの (単語認識率 52%から 49%)、手がかり語の認識率は 76%から 90%にまで向上できた。なお、手がかり語の沸き出し誤りは生じなかった。

音声認識を利用した手がかり語検出を用い、最後の 10 文+長いポーズの前 5 文+ 手がかり語を含む文の前 6 文 (**自動化手法 1**) による抽出時間長などを表 6 に示す。

また、自動化手法 1 による要約と人間による要約との聴取比較実験結果を表 7 に挙げる。聴き易さでは人間による要約が優れているが、良い要約か否かについては判断が分かれる結果となった。総合的評価は、表 5 と表 7 から、人間による要約 $\simeq$ 半自動化 $>$ 自動 1 という順序関係になった。

表 5: 半自動化手法と人間による要約との聴取比較実験

評価項目		人間 > 半自動	どちらとも言えない	人間 < 半自動
聴き易さ	講演 B	1	1	3
	講演 D	3	1	1
	講演 E	1	3	1
良い要約か	講演 B	0	4	1
	講演 D	1	3	1
	講演 E	1	3	2

表 6: 自動化手法 1 による音声データ

	講演 B	講演 D	講演 E
文の総数	314	185	333
人間による抽出文数	77	52	60
自動 1 による抽出文数	91	70	81
一致文数	32	19	33
人間による抽出区間の長さ	3'15"	3'56"	2'21"
自動 1 による抽出区間の長さ	2'54"	3'35"	2'30"
κ 値	0.157	-0.016	0.239

表 7: 自動 1 と人間による要約との聴取比較実験

評価項目		人間 > 自動 1	どちらとも言えない	人間 < 自動 1
聴き易さ	講演 B	3	1	1
	講演 D	1	3	1
	講演 E	3	2	0
良い要約か	講演 B	1	1	3
	講演 D	2	3	0
	講演 E	2	2	1

表 8: 韻律による結果と人手による結果との一致率

抽出範囲	±3	± 1	±0
F_0	29.5%	27.8%	29.8%
パワー	29.4%	28.3%	32.0%

5 韻律情報を利用した自動要約

5.1 韻律情報による重要文抽出

今回、韻律情報としては F_0 とパワーを用いた。それぞれ各文の平均値を取り [16]、その上位 n 文の前後 m 文を抽出結果とした。この抽出結果と人手による抽出結果との一致率を表 8 に挙げる。いずれも約 30% 程度であり、偶然一致率にほぼ等しい。

5.2 F_0 による要約

まず、F_0 を対象とし、各文の平均値を取り、その上位 n 文 (講演 A では 9 文、講演 B では 15 文) の各々前後 3 文を抽出した (**自動化手法 2**)。この際、自動化手法 1 で用いた表層情報は用いていない。前後 3 文としたのは、聞き易さを考慮したためである。

自動化手法 2 による抽出時間長などを表 9 に示す。なお講演 B における人間による抽出文数が自動化手法 1 等の場合と異なるのは、自動化手法による音声抽出時間と人間による結果に基づく音声抽出時間を合わせるために、人間による抽出の条件を変えているためである (自動化手法 1 では 3 名以上が一致したもの、自動化手法 2,3 では 4 名以上が一致したもの)。

また、手法 2 による要約と人間による要約との聴取比較実験結果を表 10 に挙げる。聴き易さでは判断が分かれるが、良い要約か否かについては人間の方が勝る結果となった。

表 9: 自動化手法 2(F_0) による音声データ

	講演 B	講演 D	講演 E
文の総数	314	185	333
人間による抽出文数	41	30	60
自動 2 による抽出文数	57	36	92
一致文数	6	7	6
人間による抽出区間の長さ	1'56"	2'04"	2'21"
自動 2 による抽出区間の長さ	2'00"	2'04"	2'21"
κ 値	-0.035	0.043	-0.178

表 10: 自動化手法 2 と人間による要約との聴取比較実験

評価項目		人間 > 自動 2	どちらとも言えない	人間 < 自動 2
聴き易さ	講演 B	1	2	2
	講演 D	1	3	1
	講演 E	2	2	1
良い要約か	講演 B	3	2	0
	講演 D	2	3	0
	講演 E	3	1	1

5.3 パワーによる要約

パワーに着目し、同様の実験を行なった。各文の平均値を取り、その上位 n 文 (講演 A では 8 文、講演 B では 12 文) の各々前後 3 文を抽出した (**自動化手法 3**)。

自動化手法 3 による抽出時間長などを表 11 に示す。

また、手法 3 による要約と人間による要約との聴取比較実験結果を表 12 に挙げる。聴き易さ、要約の善し悪しともに人間の要約の方が優れているという結果になった。

表 11: 自動化手法 3(パワー) による音声データ

	講演 B	講演 D	講演 E
文の総数	314	185	333
人間による抽出文数	41	30	60
自動 3 による抽出文数	55	38	83
一致文数	5	7	16
人間による抽出区間の長さ	1'56"	2'04"	2'21"
自動 3 による抽出区間の長さ	2'01"	1'57"	2'07"
κ 値	-0.148	0.030	0.018

表 12: 手法 3 と人間による要約との聴取比較実験

評価項目		人間 > 自動 3	どちらとも言えない	人間 < 自動 3
聴き易さ	講演 B	5	0	0
	講演 D	3	1	1
	講演 E	5	0	0
良い要約か	講演 B	2	2	1
	講演 D	2	2	1
	講演 E	3	1	1

6 まとめ

本研究では、人間による抽出要約の比較と、表層情報および韻律情報を用いた、講演音声の自動的な抽出要約を試みた。

人間が行なった抽出要約においては、被験者間での κ 値は 0.28 程度とあまり高い値は得られなかった。これは、対象とした講演が、約 13 分間の研究報告ということもあり、冗長な部分が少なく、また内容が高度に専門的でありどこが重要であるかの判断は被験者によるばらつきが大きいためと考えられる。

自動的な抽出要約については、3 つの手法を試みた。講演の最後の 10 文、長いポーズの直前の 5 文、手がかり語を含む直前の 6 文という表層情報を利用する方法の方が、韻律情報単独による方法に比べて、人間による抽出結果との κ 値は高い結果が得られた。

聴き易さ、良い要約か否かの聴取比較実験においては、良い要約かどうかについては人間による要約と自動要約には大差はなかったが、聞き易さにおいて人間による抽出要約の方がやや良い結果を得ており、今後の研究課題と言える。

今後は、これまでとは異なる表層情報の使用、自動書き起こし結果の利用、複数の韻律情報の組み合わせや、韻律情報と表層情報の組み合わせを検討し、より良い音声抽出要約の自動化手法の実現を目指す。また、フィラーの除去が聴き易さに効果があるかどうかの検討も行なう。

参考文献

[1] 長谷川, 秋田, 河原, 談話標識の抽出に基づいた講演音声の自動インデキシング, 情報処理学会,SLP-36-6,pp35-43,2001.

[2] 下岡, 河原, 奥乃, 講演の書き起こしに対する統計的手法を用いた文体の整形, 情報処理学会,SLP-41-3, pp.17-24,2002.

[3] 野畑, 関根, 井佐原,Grishman, 自動獲得した言語パタンを用いた重要文抽出システム, 言語処理学会第 8 回年次大会発表論文集,pp.539-542,2002.

[4] 野畑, 関根, 内元, 井佐原, 話し言葉コーパスにおける文の切り分けと重要文抽出, 第 2 回話し言葉ワークショップ, pp.93-100, 2002.

[5] 奥村, 難波, テキスト自動要約に関する研究動向, 自然言語処理, vol. 6, no.6, pp.1-26,1999.

[6] 堀, 岩崎, 古井, 話題語に着目したニュース音声の要約方の検討, 日本音響学会秋季研究発表会、vol. 1,3-1-11,pp117-118, 1999.

[7] 堀, 古井, 英語ニュース音声を対象とした音声自動要約, 情報処理学会,SLP-39-26, pp. 153-158, 2001. 1999.

[8] Waibel, Bett, Metze, Ries, Schaaf, Schultz, Soltau, Yu, Zechner, Advances in Automatic Meeting Record Creation and Access, Proc. of ICASSP, pp.597-600, 2001.

[9] Reithinger, Kipp, Engel, Alexandersson, Summarizing Multilingual Spoken Negotiation Dialogues,Proc. of ACL,pp.310-317,2000.

[10] Koumips, Remals, Miranjau, Extractive Summarization of Voicemail Using Lexicaland Prosodic Feature Subset Selection, Proc. of EuroSpeech, pp.2377-2380,2001.

[11] 笠原, 山下, 講演音声における重要文と韻律的特徴の関係, 情報処理学会,SLP-35-5,pp,25-30,2001.

[12] 三上, 井上, 山下, ポーズで分割した発話単位の韻律パラメータと文重要度の相関, 日本音響学会秋季研究発表会, 3-10-1,pp.329-330,2002.

[13] 井上, 三上, 山下, 音声要約のための重要文検出における韻律パラメータの利用, 日本音響学会秋季研究発表会, 2-9-20,pp101-102,2002.

[14] 日高, 水野, 中島, 発話の強調自動抽出による音声要約技術, 日本音響学会秋季研究発表会,2-9-19,pp99-100,2002.

[15] 甲斐, 廣瀬, 中川, 単語 N-gram 言語モデルを用いた音声認識システムにおける未知語・冗長語の処理, 情報処理学会論文誌 , Vol.40 , No.4, pp.1385-1394, 1999.

[16] 中川, 甲斐, 小林, 音声対話における韻律情報の分析, 文部省科学研究費補助金 特定領域研究 (B) 韻律に着目した音声言語情報処理の高度化研究成果報告書,pp.139-146, 2001.

[17] Fleiss, J.L., Measuring Nominal Scale Agreement Among Many Rater, Psychological Bulletin, Vol. 76, pp. 378-382, 1971.

[18] Bakemann, B and Gothman , J.M, Observing Interaction (2nd edition), Cambridge Universitty Press, 1997.

[19] 片岡, 吉川,, 小林, 中川, 講演の聴き取りと書き起こしテキストからの重要文抽出, 日本音響学会秋季研究発表会,3-Q-23, pp.199-200,2002