社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49 (14)
2003／12／18

連続音声認識による言語情報と韻律情報を利用した講演音声の重要文抽出

井上 章† 三上 貴由† 山下 洋一†

† 立命館大学理工学部情報学科
〒 525-8577 滋賀県草津市野路東 1-1-1
E-mail: †{pigman,mikami,yama}@slp.cs.ritsumei.ac.jp

あらまし 講演音声を要約するために、音声の韻律情報を利用して重要文を抽出する手法について述べる。文字テキストから得られる言語情報だけでなく、韻律情報を利用した重回帰モデルにより、文の重要度を予測する。文境界は人手で決定した。要約実験で決定した文重要度とモデルによる予測値との相関係数および重要文認定度の二つの尺度によって提案手法を評価した。韻律情報を利用することによって、文の重要度の予測精度が向上した。特に、言語情報を得るための書き起こしテキストを音声認識システムによって作成した場合に、その効果が大きくなる。
キーワード 音声要約、韻律情報、重要文抽出、文重要度、講演音声

Extraction of Important Sentences for Speech Summarization Using Prosodic Information and Linguistic Information by Continuous Speech Recognition

Akira INOUE†, Takayoshi MIKAMI†, and Yoichi YAMASHITA†

† Dept. of Computer Science, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu-shi, Shiga, 525-8577 Japan
E-mail: †{pigman,mikami,yama}@slp.cs.ritsumei.ac.jp

Abstract This paper describes the speech summarization based on the extraction of important sentences using prosodic information. A multiple regressive model using linguistic scores and prosodic parameters predicts the importance score of the sentence. The boundaries of the sentence are manually identified. The proposed method is evaluated both on the correlation between the predicted sentence importance and the preference scores by human subjects and on the accuracy of extraction of important sentences. Prosodic information improved the quality of speech summary, and it is more effective when the speech is transcribed by automatic speech recognition because speech recognition errors damage linguistic information.
Key words Speech summarization, Prosody, Sentence extraction, Sentence importance, Spoken lecture

1. はじめに

蓄積された情報コンテンツのデータベースから欲しいデータを捜し出す時、一般に、蓄積されたデータ量が増えれば増えるほど、欲しいデータを捜し出すことが難しくなり、検索技術やコンテンツへのアノテーションが重要になってくる。情報技術の進歩によって、画像、音声、文字テキスト、さらにそれらを組み合わせたメルチメディア情報が大量に蓄積できるようになっている。情報コンテンツの表現においてマルチメディア化が進むにつれて、音声に代表される言情報を含むコンテンツも増大しており、講演や演説など音声言語が中心的な役割を果たすコンテンツも多い。今後も、文化的／歴史的に意義の大きい講演や演説が多数蓄積されていき、大学での授業や学会講演などの学術的コンテンツのディジタルカーカイブ化も進むものと思われる。

データがどのような内容なのかを容易に把握することを一つの目的として、これまでに、文字テキストに対する自動要約の研究が広く行なわれてきている[1],[2]。近年、連続音声認識の性能が向上したことにより[3]、音声データに対する自動要約の研究も始まっている[4]~[10]。音声データは、文字テキストと比べてスキャニング (拾い読み／拾い聞き) が難しく、講演などの音声に対する要約自動生成に対する期待は大きい。捜し出した音声データが必要なものかどうかを判断するのに、要約された結果を聴いたり読んだりできれば非常に有用である。また、音声の自動要約に関する技術は、会議の収録音声からの議事録自動作成などへ発展していくことが考えられ、重要な要素技術として位置付けられる。

音声データの自動要約は、図1(a)に示すように、連続音声認識とテキスト要約の単純な組合せによっても実現することが可能である。すなわち、音声を連続音声認識によって文字テキストに変換し、得られた文字テキストに対してテキスト要約を行ない要約結果を得る。しかし、このような処理は音声の持つ言語的な情報のみに注目しており、非言語的な (パラ言語的な) 情報を無視されることになる。音声によるコミュニケーションでは、意図、感情、強調、微妙なニュアンスなどの非言語的情報が韻律情報 (声の高さ、声の大きさ、発話速度) によって表現されることがよく知られている。音声の要約でも、図1(b)に示すように、音声波形の持つ韻律情報を言語情報と併せて利用することによって、要約の精度を向上させられる可能性がある。そこで本報告では、講演音声を対象として、言語的な情報に加えて韻律情報を利用した要約生成について述べる。

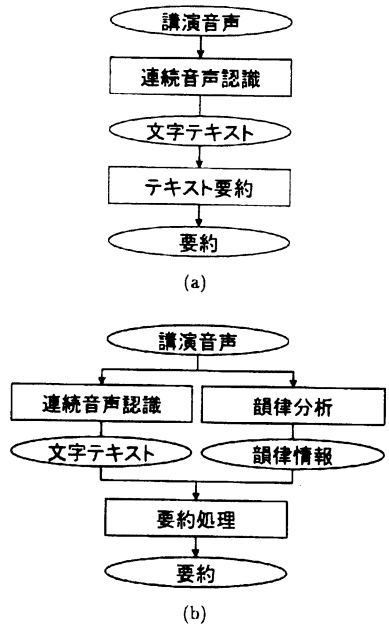

図1 講演音声の要約処理過程

2. 手　　法

2.1 要　　約

人間が文章を要約するときには、まず全文を読んで内容を理解してから重要な箇所を取り出し、それを頭の中で再構成し要約を完成させる。しかし、現在の情報処理研究では、十分な意味理解ができるところまで技術が進んでいないため、計算機が人間と同じような処理過程で要約を作り出すことは難しい。これまでに行われてきたテキスト自動要約では、図2に示すようにあらかじめ決められた文などの単位のうち、重要な部分を抽出することによって抄録する方法が多い[1],[2]。このような処理による講演音声の要約は、以下の過程からなる。

1. 講演音声を文などの単位に分割する。
2. 1文毎に重要度を算出する。
3. 重要度の高い文から必要数の文を抜き出す。

重要部分を抜き出すことで要約を生成する場合には抜き出す単位が問題になる。テキスト要約では文が単位として用いられることが多いが、音声要約では文の単位を決定することすら簡単ではない。本研究では、抽出単位の決定は重要度の決定とは別の問題と考え、人手で決めた文を単位として重要文の抽出を行う。このため、文間のつながりの良さなどの結束性は無視される。このような重要文抽出による要約の生成では、文の重要度の決定が本質的な問題となる。

2.2 言語情報

本研究では言語情報と韻律情報を組み合わせて文の重要度を算出する方法を検討する。従って韻律情報だけでなく、文テキストからの言語情報の獲得が必要となる。これまでの研究から重要な単語（重要語）が多く含まれる文は重要度が高く、出現頻度が中程度の単語は重要語である確率が高いことが知られている。これより、文章中での単語の出現頻度を見ることで各文の重要度を算出する事が可能であると言える。ほかにも重要文の検出について言語情報の有用な手がかりとしてこれまでにいくつかの方法が提案されている。要約に用いられる言語情報として、

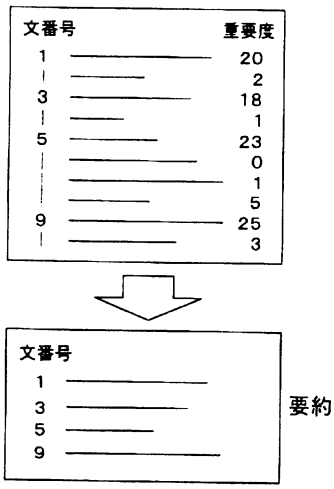

図 2　重要文抽出による要約

- 文中の位置（冒頭、段落頭、文章末など）
- 重要語の出現頻度
- 原文の構造を解明
- 文と文のつながり具合
- 手がかり語（「ようするに」「つまり」など）

などが試みられている。このような言語情報の利用に関しては、本研究では公開されているテキスト要約システム Posum [2], [11] を用いる。Posum は、テキスト中の単語の重要度や単語間のつながりを利用する基本的な要約エンジンで、テキストを入力とし、各文の重要度を出力することができる。本研究では、この重要度スコアを言語情報として利用し、以下 $LING$ と表記する。

2.3 韻律パラメータ

文の重要度を決定するために、韻律情報を利用することを考える。以下に述べる時間長、パワー (声の大きさ)、基本周波数 (声の高さ) に関するパラメータを文ごとに算出する。

2.3.1 基本周波数

基本周波数に関するパラメータとして、以下に示す文中の最小値 F_{min}, 最大値 F_{max}, レンジ F_{range}, 平均値 F_{avg} の 4 つのパラメータを文ごとに算出する。

$$
\begin{aligned}
F_{avg} &= \frac{1}{L}\sum_{i=1}^{L} f_i \\
F_{min} &= \min\{f_1, f_2, \cdots, f_L\} \\
F_{max} &= \max\{f_1, f_2, \cdots, f_L\} \\
F_{range} &= F_{max} - F_{min}
\end{aligned}
$$

ここで、L はその文のフレーム数、f_i はその文の i 番目のフレームの基本周波数である。基本周波数の算出には、Entropic 社の音声分析ライブラリ ESPS を用いた [12]。

2.3.2 音素時間長

文の j 番目の音素 ph_j の音素長 D_j を次式で正規化し、正規化された音素時間長 d_j を求める。

$$
d_j = \frac{D_j - \bar{D}(ph_j)}{\sigma_D(ph_j)} \tag{1}
$$

ここで、$\bar{D}(ph)$ と $\sigma_D(ph)$ はそれぞれ音素 ph の時間長の平均と標準偏差である。各音素の時間長 D_j は、音声認識ツール HTK [13] を用いた強制整列によって求めた。音素時間長に関するパラメータとして、上記の正規化時間長を用いて以下に示す文中の最小値 DUR_{min}、最大値 DUR_{max}、レンジ DUR_{range}、平均値 DUR_{avg} の 4 つのパラメータを文ごとに算出する。

$$
\begin{aligned}
DUR_{avg} &= \frac{1}{N}\sum_{j=1}^{N} d_j \\
DUR_{min} &= \min\{d_1, d_2, \cdots, d_N\} \\
DUR_{max} &= \max\{d_1, d_2, \cdots, d_N\} \\
DUR_{range} &= DUR_{max} - DUR_{min}
\end{aligned}
$$

ここで、N はその文の音素数である。

2.3.3 パワー

文の j 番目の音素の中心 20ms の区間の平均パワー P_j を、音素時間長に対する式 (1) と同様に、音素毎の平均値と標準偏差を用いて正規化した値を p_j とおく。パワーに関するパラメータとして、文中の最小値 POW_{min}、最大値 POW_{max}、レンジ POW_{range}、平均値 POW_{avg} の 4 つのパラメータを 2.3.2 節の音素時間長と同様に文ごとに算出する。

2.3.4 発話時間長

文の発話時間長は、狭い意味での韻律情報にはあたらないが、予備的な検討から文重要度との関連性が高いことが明らかとなったため、パラメータとして利用する。以下では、これを LEN と表記する。

3. 実験と結果

3.1 音声データ

講演音声データとしては約 10 分のＮＨＫ論説番組「あすを読む」の 5 回分を用いた。表 1 に用いたデータの内容、話者の性別、文数を示す。講演音声の文単位への分割は、人手で行なった。音声認識の性能評価、さらに重要文抽出における認識誤りの影響の分析を行なうために、まず、人手でよる書き起こしテキストを作成した。

3.2 音声認識

言語情報を得るために文ごとに音声認識を行う。音声認識は、CSRC [14]2001 年度版のシステムを用いて行なった。具体的には、デコーダは julius-3.3p3 高精度版 [15]、音響モデルは 64 混合分布、3000 状態の PTM モデル、言語モデルは語彙数 20K の 3-gram を用いた。5 つの講演音声データに対する平均単語認識精度は 64.6% であった。

3.3 文の重要度の決定

重要文抽出の性能は、人手で決定した文の重要度に基づいて評価する。人手で作成した書き起こしテキストを用い、以下の要領で重要文抽出の要約実験を行なった。

(1)　番組のビデオの視聴し、概要を理解する。

表 1　講演音声データ

データ番号	データ 1	データ 2	データ 3	データ 4	データ 5
内容	東海村臨海事故	高齢者パワーをどう活かしていくか	砂浜の再生	原発と老朽化	ヤコブ病訴訟和解へ
話者	男声 A	女性 A	男声 B	男声 C	女性 B
文数	65	68	71	76	71

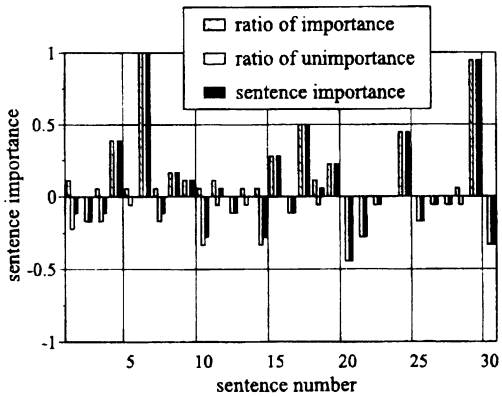

図 3　要約実験から決定された文重要度の例

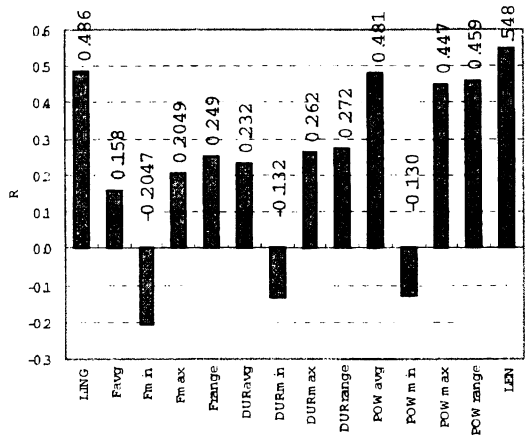

図 4　重要度と各韻律パラメータの相関係数

(2)　書き起こしテキストを見ながら音声を聴取し、書き起こしテキストから重要文／非重要文をそれぞれ 10 文程度抽出する。

被験者数はデータ 1〜5 に対して、それぞれ、14, 18, 13, 14, 15 人である。

この結果から、i 番目の文の重要度 $SI(i)$ は、

$$SI(i) = R(i)_{imp} - R(i)_{unimp} \tag{2}$$

で求める。ここで $R(i)_{imp}$、$R(i)_{umimp}$ は、それぞれ i 番目の文を重要文として選んだ人の割合、非重要文として選んだ人の割合である。図 3 に要約実験から決定された文重要度の例を示す。

3.4　文重要度と各韻律パラメータとの相関

複数の韻律パラメータを組み合わせて用いる前に、各韻律パラメータと人手による文重要度との相関を調べた。結果を図 4 に示す。ここでは、5 つのデータに対する相関係数の平均値を示している。これより、言語情報 ($LING$) の相関が高いほか、LEN, POW_{avg}, POW_{max}, POW_{range} においても高い相関が見られる。

3.5　複数の韻律パラメータの組合せによる文重要度の予測

次に、言語情報と複数の韻律パラメータを組み合わせて用い、文の重要度を予測する。予測には重回帰モデルを利用する。言語情報は必ず用いることとし、i 番目の文の重要度を

$$SI(i) = a_0 + a_{LING} \times LING(i) + \sum_{j=1}^{M} a_j \times B(i)_j \tag{3}$$

で予測する。$LING(i)$ は Posum から出力される言語情報のみによる i 番目の文の重要度スコア、$B(i)_j$ は i 番目の文における j 番目の韻律パラメータ、M は組み合わせる韻律パラメータの数である。学習データを用いて、モデルパラメータ a_0, a_{LING}, a_j を決定することによってモデルが作成される。

パラメータの組合せを考える時、2.3 節で述べた全ての韻律パラメータを用いて重回帰モデルを作成することもできるが、基本周波数に関する F_{avg}, F_{min}, F_{max}, F_{range} など、同じ種類のパラメータ間では相互の相関が高い場合があり、用いるパラメータ数を増やすことが必ずしも精度の高いモデルの作成にはつながらない。そこで、3.5 節での結果をもとに、韻律パラメータの種類 (基本周波数、音素時間長、パワー) ごとに文重要度との相関の高いものから順にパラメータを選択して用いることを考える。用いるパラメータの組合せとして、表 2 に示す 3 つのセットを試みた。$C0$ は、言語情報のみを用いる場合で、これがベースラインの性能を与える。$C1$ は、これに発話時間長を加え、韻律パラメータの種類ごとに、文重要度との相関が最も高いパラメータを一つずつ加えている。$C2$ では、さらに韻律パラメータの種類ごとに、相関が 2 番目に高いパラメータも加えている。

3.5.1　相関係数による評価

各パラメータセットにおける重回帰モデルの重相関係数の値を、書き起こしテキストの作成方法および評価の仕方に分けて図 5 に示す。重相関係数は 0 から 1 の間の値をとり、値が大きいほどモデルによる現象の説明がうまくいっていることを示す。

ここで trans- と CSR- はそれぞれ人手による書き起こし

表 2　パラメータの組合せ

パラメータセット	用いるパラメータ
$C0$	$LING$
$C1$	$LING, LEN, F_{range}, DUR_{range}, POW_{avg}$
$C2$	$LING, LEN, F_{range}, F_{min}, DUR_{range}, DUR_{max}, POW_{avg}, POW_{range}$

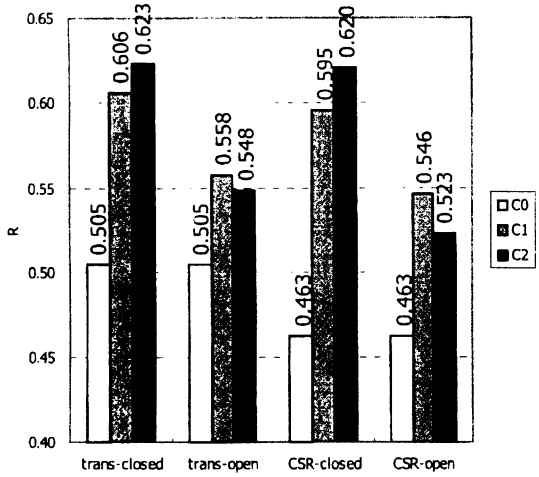

図 5　重相関係数による評価

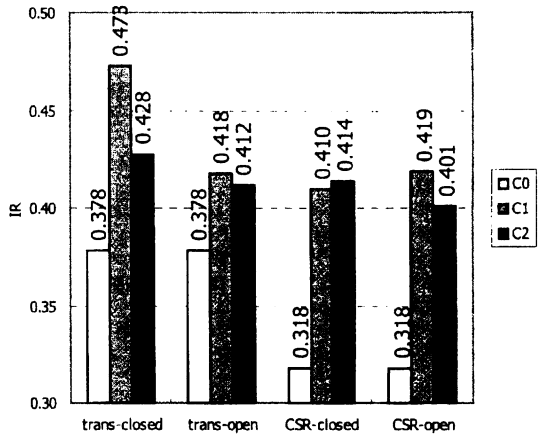

図 6　重要文認定度による評価

テキストと自動音声認識によるテキストから言語情報スコア $LING$ を算出した場合を示しており、また、-closed、-open はそれぞれクローズド評価、オープン評価を示している。オープン評価では、表 1 の 5 つの講演音声データのうち 4 つのデータで重回帰モデルを作成し、残りの一つのデータに適用して評価する処理を 5 回行ない、その平均値を求めた。

図 5 を見ると $C1, C2$ ともベースラインの $C0$ よりも重要関係数が大きくなっている。$C1$ と $C2$ の比較では、オープンな評価において $C2$ の方がやや重相関係数が小さくなっており、必ずしも用いる韻律パラメータが多い方が良いとはかぎらないことを示している。しかし、学習に用いたデータ数がそれほど多くはないため、今後、学習データ量をさらに増やした場合には、$C2$ の方が重相関係数が大きくなることも考えられる。

次に、書き起こしテキストを人手で作成した場合と、連続音声認識の結果を利用した場合を比較する。$C0$、すなわち言語情報だけで重要度を決定する場合には、連続音声認識を用いることによって重相関係数がやや小さくなっている。これは、連続音声認識による認識誤りのために言語情報が劣化したためと考えられる。一方、$C1, C2$ の韻律パラメータを利用して文重要度を決定した場合には、連続音声認識を用いても重相関係数はそれほど変化しておらず、韻律パラメータを利用することによる効果は連続音声認識を使った場合の方が顕著であることがわかる。

3.5.2　重要文認定度による評価

重要文抽出による要約の精度を評価するために、次式で定義する重要文認定度 IR を評価尺度として新しく導入する。

$$IR = \frac{1}{4}(IR_5 + IR_{10} + IR_{15} + IR_{20}) \tag{4}$$

$$IR_n = \frac{C(n)_{imp} - C(n)_{umimp}}{n} \tag{5}$$

ここで $C(n)_{imp}$, $C(n)_{umimp}$ は文重要度に基づいて抽出された上位 n 文と、人手による重要度におけるそれぞれ上位 n 文／下位 n 文との一致文数である。IR_n は n 文の重要文抽出を行なうとき、抽出されるべき 1 つの重要文が実際に抽出される「見込み」を表しており、$n = 5, 10, 15, 20$ の時の結果を平均した値を IR としている。

重要文認定度 IR による各パラメータセットに対する評価結果を図 6 に示す。ここで trans, CSR, closed, open の意味は、図 5 と同じである。

図 6 を見ると、ベースラインの $C0$ より $C1, C2$ の精度が高くなっており、重要文認定度の尺度においても韻律情報を利用した効果が確認できる。さらに、韻律パラメータを利用することによる効果が連続音声認識を使った場合の方が顕著であることも重相関係数による評価の場合と同様である。

4. おわりに

重要文抽出における講演音声の自動要約において、韻律情報の利用が有効であることを示した。連続音声認識システムを用いて講演内容の書き起こしテキストを作成した場合に、その効果が大きくなることが確認された。本報告では重要箇所抽出の単位となる文を人手で決定している。今後、自動抽出を実現するためには、文など抽出単位の自動決定を行なう必要である。また、要約精度の評価手法の再検討や他の講演音声データでの評価も今後の課題として挙げられる。

謝辞

研究を行うに当たって言語要約システム Posum の使用を快諾してくださった望月源氏に深く感謝する。本研究の遂行には、科学研究費補助金 (特定領域研究 (B)(2)「韻律と音声処理」、No.12132203)、および 21 世紀 COE「京都アート・エンタテインメント創成研究」の支援を受けた。

文　献

[1] I. Mani and M. Maybury : "Adavances in Automatic Text Summarization", The MIT Press (1999).
[2] 奥村学, 望月源 : "テキストを自動的に要約する技術 —第 1 回— テキスト中の重要な文を抜き出す", コンピュータサイエンス誌 bit2 月号, 共立出版, pp.37-42 (2000).
[3] 中川聖一 : "音声認識研究の動向", 電子情報通信学会論文誌, **J83-DII**, 2, pp.433-457 (2003).
[4] 笠原力弥, 山下洋一 : "講演音声における重要文と韻律的特徴の関係", 情報処理学会研究報告, SLP-35-5 (2001).
[5] 井上章, 三上貴由, 山下洋一 : "複数の韻律パラメータを用いた音声要約のための文重要度予測", 日本音響学会春季研究発表会講演論文集, 2-4-6, pp.69-70 (2003).
[6] 堀智織, 古井貞熙 : "単語抽出による音声要約文生成法とその評価", 電子情報通信学会論文誌, **J85-DII**, 2, pp.200-209 (2002).
[7] 小林聡, 吉川裕規, 中川 聖一 : "表層情報と韻律情報を利用した講演音声の要約", 情報処理学会研究報告, SLP-43-7 (2002).
[8] 北出祐, 南條浩輝, 河原達也, 奥乃博 : "談話標識と話題語に基づく統計的尺度による講演からの重要文抽出", 情報処理学会研究報告, SLP-46-2 (2003).
[9] S. R. Maskey and J. Hirschberg : "Automatic Summarization of Broadcast News using Structural Features", Proc. of Eurospeech 2003, pp.1173-1176 (2003).
[10] B. Wrede and E. Shriberg : "Spotting "Hot Spots" in Meetings: Human Judgments and Prosodic Cues", Proc. of Eurospeech 2003, pp.2805-2808 (2003).
[11] http://www.tufs.ac.jp/ts/personal/motizuki/ software/posumcl/
[12] http://www.entropic.com/
[13] http://htk.eng.cam.ac.uk/
[14] http://www.lang.astem.or.jp/CSRC/
[15] http://julius.sourceforge.jp/