音声言語情報処理 43－8
（２００２．１０．２５）

講演音声の自動要約のための韻律情報の利用

井上 章† 三上 貴由† 山下 洋一‡

†立命館大学大学院 理工学研究科 情報システム学専攻
‡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 講演音声の音声要約を行うために、音声の韻律情報を利用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、音声から得られる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。対象とした講演音声は人手で書き起こし文単位も人手で決定した。基本周波数、パワー、継続時間長に関して、いくつかのパラメータを分析し、人手による要約実験で決定した文重要度との相関を調べた。重要度予測における重相関関数および文の重要度抽出において、韻律情報を利用することによって精度が向上することを示した。

キーワード 音声要約、韻律情報、重要度、講演音声

.

Use of prosodic parameters for automatic summarization of lecture speech

Akira Inoue†, Mikami Takayoshi†, and Youichi Tamashita‡

†Graguate school of Science and Engineering,Ritsumeikan University
‡Dep.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 proposes a new method of automatic scoring of sentence importance using prosodic parameters in lecture speech for speech summarization. The linear multiple regression model predicts sentence importance using not only linguistic information from written text but also prosodic parameters obtained by speech wave. Lecture speech is transcribed by hand and the sentences in the lecture speech is manually identified. The correlation coefficients to the sentence importance score by human preference are investigated for several types of parameters of F0, power, and duration, to identify efficient prosodic parameters. The introduction of prosodic parameters increases the multiple correlation coefficients to human preference and extraction accuracy of important sentence.

Key words Speech summarization, Prosody, Sentence importance, Lecture speech

1. はじめに

近年、大容量なハードディスク、ＣＤやMO等のメディアが広く普及し、音声データや画像データ、テキストデータなど様々なデータが容易に蓄積できるようになった。人間の情報処理能力は限られているため、蓄積されるデータ量が増大するにつれ必要な情報を取り出すことが難しくなる。そのため、情報検索や自動要約などの技術の研究が非常に重要になる。音声データについても容易に大きなデータをコンピュータに保存ができるようになった。しかし、講演や講義など長時間の話の内容（発話）を要約するという技術はあまり進んでいない。音声データの特徴として、素早いスキャンニングが難しいということがあり、要約による効果はテキストより大きいと考えられるため、今後はこのような音声データに対する要約技術が期待されている。

現在ではテキストにおける要約技術がかなり進んでおり、いくつかの市販アプリケーションにもこのような機能が組み込まれている。一方では、精度の高い連続音声認識によって音声から容易にテキストにすることも可能になってきている。このような技術を組み合わせることにより、図１の実線矢印で表すような手順、すなわち、つまり、音声を連続音声認識によってテキスト化し、言語情報のみを用いたテキスト要約を行うことによっても音声の自動要約は可能である。しかし、文字テキストで表現された文章とは異なり音声には韻律的特徴により非言語的な情報が表現されると考えられる。従って音声を対象とした要約では、図１の点線で表した手順、つまり言語情報だけでなく音声の韻律的特徴を利用することにより要約の品質を向上できる可能性がある。

本研究では、講演音声から抽出した様々な韻律パラメータを用いて重回帰モデルにより文の重要度を決定する手法について考察する。文の重要度の予測には、韻律パラメータだけでなく、言語情報から得られた情報も合わせて利用し、韻律パラメータを利用することによる効果を明らかにする。

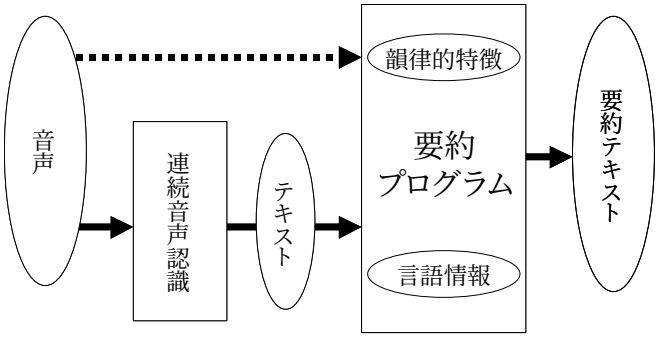


図１ 音声要約

2. 手法

2.1 要約

テキストを対象とした自動要約では言語情報を用いて自動要約を行っている。人間が文章を要約する際には、まず全文を読んで理解し、それを頭の中で再構成し要約を完成させる。しかし、現在の計算機技術は、意味理解に関してまだ十分に研究が進んでいないため計算機が人間と同じ作業によって要約を作り出すのは難しい。現在行われている多くのテキスト自動要約は、文中の重要な部分を抽出しようとする方法である[1]。一般に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。このような要約の導出過程は以下の３過程からなる。

1. 講演音声を文単位に分割
2. １文毎に重要度を算出
3. 重要度の高い文から必要文数抜き出す

ここで、講演音声は人手で書き起こし、文の決定も人手で行う。このような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 言語情報を用いた要約手法

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、

- 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（「要するに」「つまり」など）

などが挙げられる。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][3]を利用することとした。このシステム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。

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 の論説番組「あすを読む」を用いた。今回は表 1 に示す放送 3 回分の内容に対して分析を行った。

表 1　音声データ

	データ1	データ2	データ3
内容	高齢者パワーをどう生かしていくか	砂浜の再生	東海村臨界事故
話者	女性	男性A	男性B
全文数	68 文	71 文	65 文
全文節数	816 文節	788 文節	830 文節

3.2 文の重要度の決定

人手による文の重要度を決定するために、人手による要約実験を行った。書き起こしテキストは実験前に人手により作成した。被験者数はデータ 1,2,3 それぞれ 18 人、13 人、14 人で以下の要領で実験を行った。

1. 番組のビデオの視聴
2. 書き起こしテキストを見ながら音声の聴収
3. テキストを見ながら、重要文／非重要文をそれぞれ 10 文程度抽出

i 番目の文の重要度 $SI(i)$ は、式(1)で計算した。

$$SI(i) = R(i)_{inp} - R(i)_{unimp} \qquad (1)$$

ここで $R(i)_{imp}$、$R(i)_{unimp}$ 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。

3.3 韻律パラメータ

今回講演音声の中から以下のように時間長、パワー、基本周波数に関するパラメータを各文毎に算出した。

3.3.1 基本周波数

基本周波数のパラメータとして以下の 3 つのパラメータを用意した。

$$F_{\min} = \min(f_1, f_2, ..., f_N)$$

$$F_{\max} = \max(f_1, f_2, ..., f_N)$$

$$F_{range} = F_{\max} - F_{\min}$$

ここで N はその文のフレーム数、f_i はその文での i 番目のフレームの基本周波数を示す。基本周波数の算出には ESPS を使用した。

3.3.2 音素時間長

文の i 番目の音素 ph_i の音素長 D_i を式(2)で正規化し、正規化された音素時間長 d_i を求める。

$$d_i = \frac{D_i - \overline{D}(ph_i)}{s_D(ph_i)} \qquad (2)$$

ここで、$\overline{D}(ph)$ は音素 ph の時間長の平均、$s_D(ph)$ は音素 ph の時間長の標準偏差であり、データ 1,2,3 別々に算出した。音素長 D_i の決定は HTK による強制配列によって行った。

音素時間長を表すパラメータとして以下の 4 つのパラメータを算出する。

$$DUR_{avg} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$DUR_{\min} = \min(d_1, d_2, ..., d_n)$$

$$DUR_{\max} = \max(d_1, d_2, ..., d_n)$$

$$DUR_{range} = DUR_{\max} - DUR_{\min}$$

ここで n はその文の音素数を示す。

3.3.3 発話時間長

文の発話時間長を LEN とした。

3.3.4 パワー

文の i 番目の音素の中心 20ms の区間の平均パワー P_i を3.3.2の音素時間長の式(2)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_i とおく。

パワーを表すパラメータとして以下の４つのパラメータを算出する。

$$POW_{avg} = \frac{\sum_{i=1}^{n} p_i}{n}$$

$$POW_{\min} = \min(p_1, p_2, ..., p_n)$$

$$POW_{\max} = \max(p_1, p_2, ..., p_n)$$

$$POW_{range} = POW_{\max} - POW_{\min}$$

3.4 重要度と各韻律パラメータとの相関

各韻律パラメータと人手による文重要度 $SI(i)$ との相関を図 2 に示す。データ 1,2,3 に対する個別の相関係数と平均の値を示している。この結果を見ると、基本周波数、時間長、パワーの内で、それぞれ $F_{\min}$, DUR_{range}, POW_{avg} が高い相関を示している。

3.5 重回帰分析

次に、言語情報と複数の韻律パラメータを組み合わせて文の重要度を予測するために、重回帰分析を行った。重回帰式は、

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

とし、$LING(i)$ は言語情報のみによる i 番目の文の重要度、$B(i)_j$ は i 番目の文における組み合わせる 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。

3.6 言語情報と韻律情報の組み合わせ

まず１つの韻律パラメータを単独で用いて言語情報による重要度を組み合わせ、韻律パラメータを用いる効果を調べた。各韻律パラメータと言語情報のみによる文の重要度 $LING$ との組み合わせを重回帰分析によって行った。その結果得られた重相関係数を図 3 に示す。この結果を見ると、基本周波数、時間長、パワーの内で、それぞれ F_{range}, $DUR_{\max}$, POW_{avg} が高い相関を示している。図 3 を見るとデータ 2 が全体的に高い相関を見せているが、これは、各韻律パラメータと $LING$ の相関が低かったためであると考えられる。

3.7 韻律パラメータの組み合わせ

言語情報のみによる文の重要度 $LING$ 、発話時間長 LEN 、そして韻律情報を組み合わせることを考える。組み合わせる韻律情報のセットとして、以下の２つを用意した。

- セット 1

  $F_{\min}$, DUR_{range}, POW_{avg}

- セット 2

  F_{range}, $DUR_{\max}$, POW_{avg}

セット 1 は3.4、セット 2 は3.6の結果より、基本周波数、時間長、パワーのそれぞれの内で文重要度 SI との相関が高かったため選択した。

表 2に挙げた８つのパターンについて組み合わせを行った。

表 2　韻律パラメータの組み合わせパターン

組み合わせパターン	C0	C1	C2	C3	C4	C5	C6	C7	C8
LING	○	○	○	○	○	○	○	○	○
LEN	×	○	○	○	○	×	×	×	○
基本周波数	×	○	×	×	×	○	×	×	○
音素時間長	×	×	○	×	×	×	○	×	○
パワー	×	×	×	○	×	×	×	○	○

図 4(a) (b)にそれぞれセット 1、セット 2 の各組み合わせパターンの重回帰分析結果における重相関係数を示す。図 4(a)を見ると、POW_{avg} を使用しているパターン C3,C7,C8 が他の韻律パラメータを組み合わせたときよりも高い相関があることがわかる。図 4(b)を見ても同様の事が言える。

3.8 重要文の検出

次式で定義された重要文認定度 S を用いて重要文

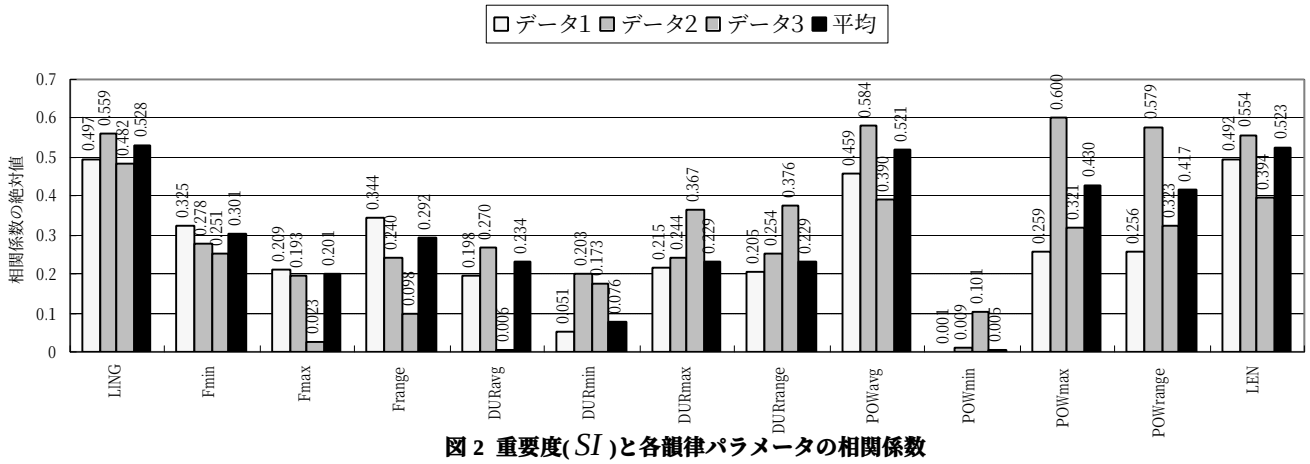


図 2 重要度(SI)と各韻律パラメータの相関係数

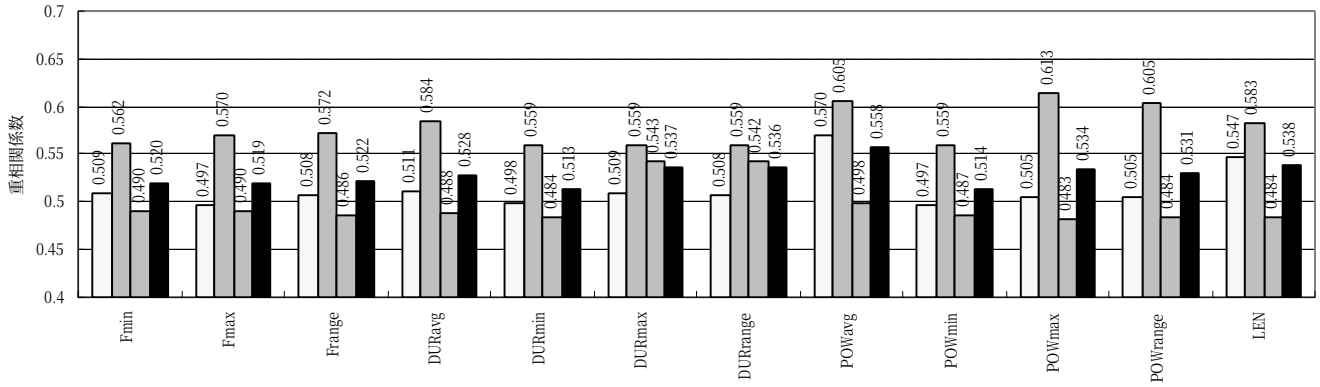
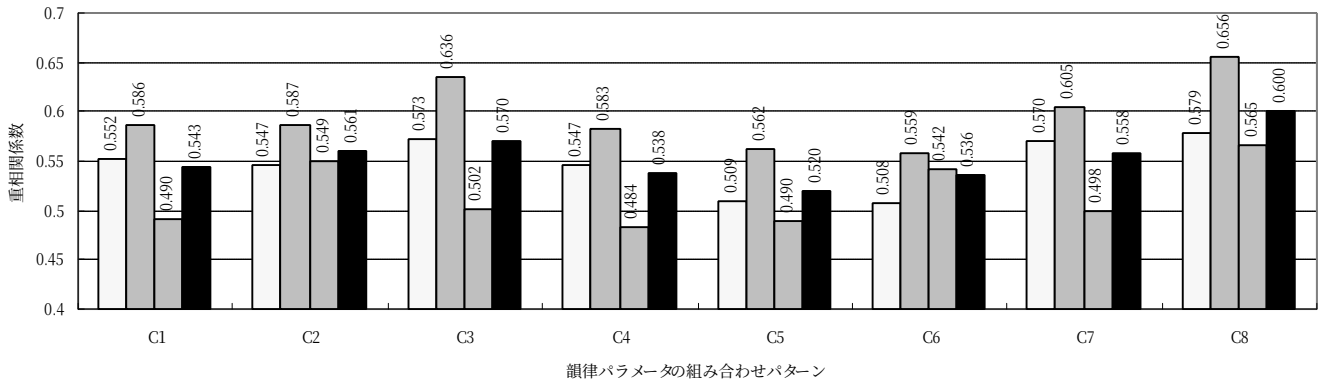
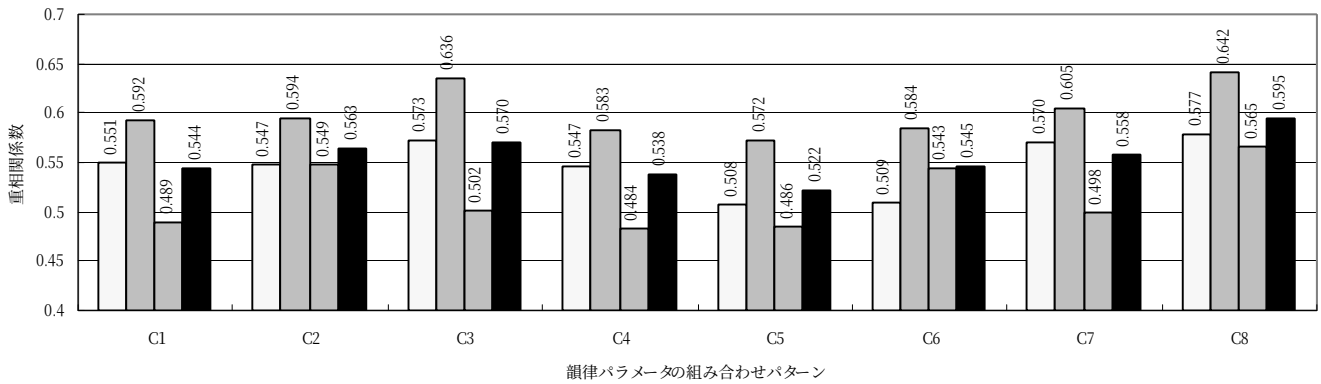


図 3 言語情報と一つの韻律パラメータを用いた文重要度予測における重相関係数



(a) セット 1 (基本周波数: $F_{\min}$,音素時間長: DUR_{range} ,パワー: POW_{avg})



(b) セット 2 (基本周波数: F_{range} ,音素時間長: $DUR_{\max}$,パワー: POW_{avg})

図 4 言語情報と複数の韻律パラメータを用いた文重要度予測における重相関係数

検出の精度を評価する。

$$S = \frac{S_5 + S_{10} + S_{15} + S_{20}}{4}$$

$$S_n = \frac{C(n)_{imp} - C(n)_{unimp}}{n}$$

ここで $C(n)_{imp}$, $C(n)_{unimp}$ は式(3)で予測された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。S_n は n 文の重要文抽出において抽出されるべき 1 つの重要文が抽出される「見込み」を表しており、n = 5,10,15,20 の時を平均した値を S としている。セット 1,2 の重要文認定度 S による評価をそれぞれ図 5(a),(b)に示す。図 5 を見ると韻律パラメータを組み合わせた C1～C8 のデータ 1,2,3 平均が言語情報のみのときの C0 の値を上回っている。これは言語情報のみならず韻律情報も利用することにより要約の質が向上した事を示している。

4. まとめ

今回は言語情報のみならず韻律情報も組み合わせて使用することにより要約の精度が向上する事を示した。また韻律情報の中でも平均パワーの利用が文重要度との相関を高めることがわかった。

今後の予定として、単語や文節単位での分析によってより細かく分析を行うことが考えられる。

謝辞

今回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します。

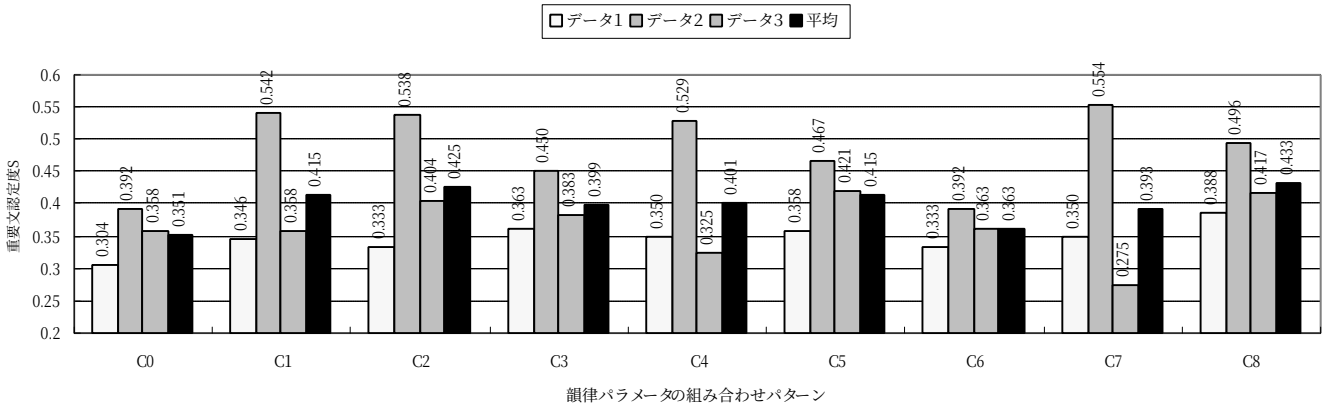
文　献

[1] 奥村学, 難波英嗣,「テキスト自動要約に関する最近の話題」北陸先端科学技術大学院大学情報科学研究科 Research Report, IS-TM-2000-001,2000.

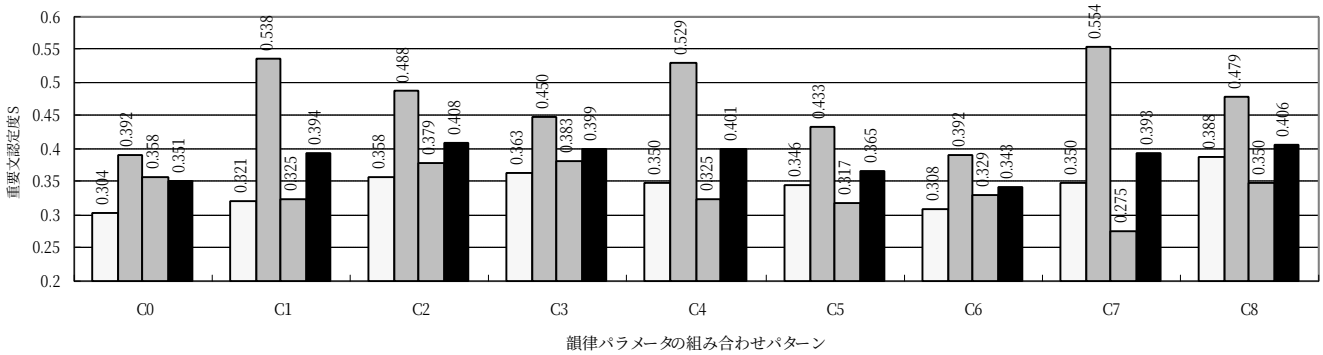
[2] 奥村学，望月源,「テキストを自動的に要約する技術-第 1 回- テキスト中の重要な文を抜き出す」，コンピュータサイエンス誌 bit 2 月号，共立出版，pp.37-42,2000.2.

[3] http://www.tufs.ac.jp/ts/personal/motizuki/software/posumcl/

[4] 有馬哲,石村貞夫,「多変量解析のはなし」 東京図書 ISBN4-489-00231-9 C0041 P1860E



(a) セット 1 (基本周波数: $F_{\min}$,音素時間長: DUR_{range} ,パワー: POW_{avg})



(b) セット 2 (基本周波数: F_{range} ,音素時間長: $DUR_{\max}$,パワー: POW_{avg})

図 5 文重要度予測に基づく重要文認定度