講演・解説・ニュース音声の要約における表層的言語情報と韻律情報の検討

山口 優[†]　**中川 聖一**[†]

[†] 豊橋技術科学大学 情報工学系
e-mail: {yamaguchi, nakagawa}@slp.ics.tut.ac.jp

1　はじめに

音声情報は，その記録は容易であるが，記録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, 3]．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．

本稿では,「話し言葉工学」プロジェクト (CSJ) コーパスの講演音声，NHK のニュースの解説番組である「あすを読む」と「視点・論点」，およびニュース音声を，より多くの表層的言語情報と韻律情報を用いて自動的に重要文を抽出して要約する方法を検討し，人手による要約と比較した結果について述べる．韻律情報では発話時間長，表層的言語情報においては手がかり語や頻出単語などの特徴が重要文抽出において有用であることを示す．また，表層的言語情報と韻律情報の複数の特徴を組み合わせた結果について述べる．

以下 2 節では，実験に利用した試料や実験条件，本稿で用いる評価尺度について述べる．3 節では，人手の要約における被験者間の重要文の一致の度合について述べる．4 節では，韻律情報と表層的言語情報を利用した重要文抽出の手法，および抽出された要約文候補の棄却可能な特徴について説明する．5 節では，表層的言語情報と韻律情報を組み合わせた場合の自動的要約について説明する．6 節では評価実験と作成した要約の聴取実験について述べる．

2　音声資料と評価尺度

2.1　音声資料と書き起こし

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から 5 講演音声 (男性話者 5 名) ，NHK のニュースの解説番組である「あすを読む」と「視点・論点」，NHK のニュース音声を対象として実験を行なった．表 1 に講演音声と解説番組，ニュース音声の諸元を記す.

表 1: 要約対象音声の諸元

対象音声	対象音声数	平均文数	平均時間長
講演音声	5	259.2	13 分 15 秒
あすを読む	10	175.0	9 分 58 秒
視点・論点	5	155.4	10 分 00 秒
ニュース	8	57.1	2 分 49 秒

2.2　評価尺度

評価尺度として適合率と κ 値，F 値を用いた．それぞれの尺度について説明する．

- **κ 値**
  κ 値 [4] とは 2 者の判定の一致度を，偶然の一致を考慮して調整した指標であり，次式により定義されている．

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

$$P(A) = \frac{\text{重要文の一致数} + \text{非重要文の一致数}}{\text{文の総数}} \tag{2}$$

$$\begin{aligned} P(E) \quad &= \text{重要文の偶然の一致率} \\ &+ \text{非重要文の偶然の一致率} \end{aligned} \tag{3}$$

$$\text{重要文の偶然の一致率} = \frac{A \text{ が重要と判定した文}}{\text{文の総数}} \times \frac{B \text{ が重要と判定した文}}{\text{文の総数}} \tag{4}$$

$$\text{非重要文の偶然の一致率} = \frac{A \text{ が非重要と判定した文}}{\text{文の総数}} \times \frac{B \text{ が非重要と判定した文}}{\text{文の総数}} \tag{5}$$

- **一致率**
  人間が抽出した文集合 (正解文集合と見なす) を H とし，機械が抽出した文集合を M とすると，以下の二式が考えられる．(6) 式は機械から見た人間との適合率 (Precision)，(7) 式は人間から見た機械との再現率 (Recall) を表している [5]．本研究では (6) 式の適合率を用いる．

$$\text{適合率 } (Precision) = \frac{|M \cap H|}{|M|} \tag{6}$$

$$\text{再現率 } (Recall) = \frac{|M \cap H|}{|H|} \tag{7}$$

- **F 値**
  F 値は適合率と再現率の調和平均を示し，(8) 式により与えられる．

$$F \text{ 値} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

3　人間による重要文抽出

本研究で用いた講演音声は学会発表時のもので，講演内容が極めて専門性が高いため，人間による重要文抽出における被験者は，対象とする講演の内容を理解できる音声研究の専門家 6 名とした．解説音声とニュース音声については学生を被験者とした．各被験者は対象とする講演について，書き起こしを読み重要文の抽出要約を行った [6]．実際の聴講時のように講演音声を聴きながら逐次的に重要文を抽出する被験者

実験も行なっている [7] が，書き起こしテキストの場合とほぼ同等の結果である．本研究では要約の長さを対象音声の 3 分の 1 程度 (要約率 33%) を目標としており，各被験者に抽出文数は全体の 3 分の 1 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．また，ここで被験者 5 人中 3 人以上が重要と判断した文を人間の要約とし **man3** と呼ぶことにする．**man3** をとることで，被験者の ばらつきが小さくなり，より安定に重要な文を得ることが可能となる．

はじめに被験者間の重要文の一致の度合を見るため，被験者間の κ 値と適合率および各被験と人間 (その被験者を除いた 5 名による **man3**) の間の κ 値と適合率を求め，それぞれ図 1 と図 2 に示す．図 1 より講演音声とニュース音声における被験者間の適合率は約 0.55，κ 値は約 0.3 であったが，解説番組である「あすを読む」と「視点・論点」では講演音声とニュース音声と比較して，それほど高い値とは言えない．図 2 より，講演音声と「あすを読む」，ニュース音声については適合率が約 0.6，κ 値が約 0.35 から 0.4 と比較的高い値となった．この値が自動要約手法の目標値となる．以降，比較的一致度の高かった講演音声とニュース音声を対象に重要文抽出を行う．

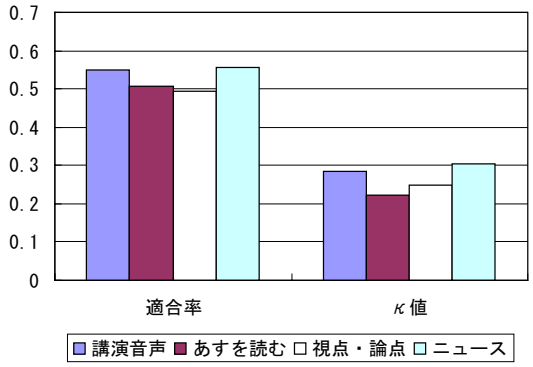

図 1: 被験者間の重要文の一致の度合い

4 韻律情報と表層的言語情報を利用した重要文抽出

4.1 韻律情報

音声には韻律的な情報を含んでおり，この韻律情報を用いて抽出要約の精度の向上を図ることができると考えられる．韻律情報として $F0$，パワー，発話時間長，ポーズを用いた．これらについて様々な特徴パラメータを抽出したが，以下で述べる特徴が比較的有用であった [6].

- $F0$
  各文の $F0$ の平均と標準偏差を求め，“平均 + 標準偏差 ” 以上の文を抽出する ($F0_{avg}$).
- **パワー**
  $F0$ と同様の手法により POW_{avg} を求め抽出を行う．
- **発話時間長**
  各文の発話時間長の平均を求め，平均の高い文から上位 3 分の 1 抽出する (LEN).
- **ポーズ**
  1 秒程度以上のポーズを長いポーズと定義する．その長いポーズを境に話題の転換が考えられ，この話題転換の境界付近に重要文の存在がいくつか確認された．長いポーズのベスト 10 を求め，長いポーズベスト 10 の前 5 文 (P_B10_{pre5}) と長いポーズベスト 10 の後 5 文 (P_B10_{pos5}) を抽出する．総文数の少ないニュース音声については長いポーズベスト 3(P_B3_{pre5}, P_B3_{pos5}) とする．

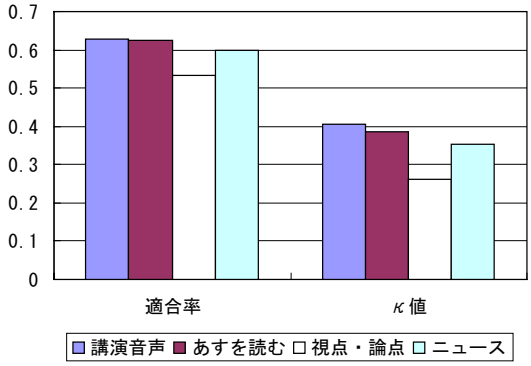

図 2: 被験者とその被験者を除いた **man3** との間の一致の度合い

4.2 表層的言語情報

表層的言語情報とは手がかり語や文の位置といった情報である．本研究では手がかり語，頻出単語，tf ベース，文の位置の各特徴を書き起こしテキストもしくは音声認識結果を基に利用した [5,6,8,13].

- **手がかり語**
  手がかり語とはその前後に重要文があるであろうという観点から定義されている．手がかり語としては “与えております” や “いう結果になってます”，“結果がこれです” など全 75 個を定めた．手がかり語は音声認識 [9] によって得られた結果を用いた．
  それぞれの手がかり語ごとに手がかり語を含む文とその前 5 文，手がかり語を含む文とその後 5 文，および手がかり語を含む文とその前後 5 文を抽出した 3 種類の文集合と人間による要約結果との適合率を求め，その値が 40%以上になる手がかり語と文の抽出法を決定した．
  決定後の手がかり語を含む文とその前 5 文 (CUE_{pre5}) と手がかり語を含む文とその後 5 文 (CUE_{pos5})，および手がかり語を含む文とその前後 5 文 (CUE_{in5}) を抽出する．これらによって抽出された文の集合を CUE_{all} とする．
- **頻出単語**
  頻出単語とは書き起こしテキストもしくは音声認識結果中に何度も現れる単語を意味する．頻出単語の検出には形態素解析システム『茶筌』[10] を使用し，不要語やフィラー，数詞を除いた一般名詞のみを頻出単語として定義する．類似したものとして情報検索で有用な tf·idf 法 [12] があり，tf では語頻度 (語の出現回数) を文のスコアとして文抽出を行う．しかし，長い文にはより多くの単語が含まれるため，長い文が抽出されやすいという傾向が見られた．そこで，頻出単語においては，1 文中にその単語が 2 語以上含まれている文を重要文とみなし文抽出を行う．書き起こしもしくは音声認識結果において頻出単語の出現回数を調べ，頻出単語を含む文を全体

の 3 分の 1 程度となるよう単語数の閾値を決め文を抽出する ($WORD_{rpt}$)．なお，講演音声の人手による書き起こしデータを用いた場合の単語数の閾値は 5 つの講演において 7, 8, 9, 3, 22 単語であった．また，ニュース 8 記事においては平均 13 単語であった．

- **tf(term frequency)**
  tf とは文書中に現れる単語の出現頻度のことである．今回 tf ベースでの重要文抽出には自動要約器 Posum [11] を用いた．本実験では不要語やフィラーを除去しカウントする品詞を名詞のみとして tf ベースで全体の 3 分の 1 程度となるよう抽出を行う (TF)．なお，TF による手法を本研究のベースラインとする．
- **文の位置**
  文の位置情報として，各講演とニュースのはじめの 10 文 ($LEAD_{10}$)，最後の 10 文 ($TAIL_{10}$) の抽出を行う．

4.3 棄却可能な特徴

これまで韻律情報と表層的言語情報を用いた重要文抽出の手法について述べてきたが，非重要文を棄却することでより良い要約が可能となると考えられる．様々な特徴を検討した結果，以下に棄却可能な特徴を示す．

- **発話時間の短い文の削除**
  各文の発話時間の平均値を 1 講演，またはニュース 1 記事の全ての文に対して求め，その “平均-標準偏差” 以下を発話時間の短い文 ($\#LEN_{short}$) と定義し抽出する．
- **長いポーズに挟まれた数文の削除**
  前節までに，長いポーズの前／後 5 文の抽出について述べた．その長いポーズの前後 5 文を見ると，その中に別の長いポーズの存在が確認された．その長いポーズ間の数文はデモの最中であったり，言い誤って言葉に詰まっているところが見受けられ重要文である可能性は低いと考えられる．ここでは長いポーズベスト 10 の間に挟まれた 5 文以内の文 ($BETWEEN$) を対象とし抽出する．

5 韻律情報と表層的言語情報を組み合わせた要約

書き起こしと音声認識による自動書き起こしに対して，韻律情報と表層的言語情報の中で比較的結果の良かった特徴を組み合わせて要約を行なう．組み合わせに用いる特徴は前節までに説明した韻律情報と表層的言語情報の中で，適合率が 40%以上の値を得ることのできたものとした．特徴 F_k により重要と推定された文 $S_i(1 \leq i \leq n)$ に対するスコア $Score_{F_k}()$ を (9) 式のように定義する．

$$Score_{F_k}(S_i) = \begin{cases} 1 & \text{(重要文として抽出された文)} \\ 0 & \text{(それ以外の文)} \end{cases} \quad (9)$$

ここで n は文書中の総文数を示す．
また，棄却可能な特徴 D_p により非重要と推定された文 S_i に対するスコア $Score_{D_p}()$ を (10) 式のように定義する．

$$Score_{D_p}(S_i) = \begin{cases} 1 & \text{(非重要文として抽出された文)} \\ 0 & \text{(それ以外の文)} \end{cases} \quad (10)$$

各特徴の値 $\left(Score_{F_k}(), Score_{D_p}()\right)$ に各特徴の寄与度 (α_k, β_p) を掛け合わせたものの線形和をとって各分 S_i のスコアとする．

$$TotalScore(S_i) = \sum_k \alpha_k Score_{F_k}(S_i) + \sum_p \beta_p Score_{D_p}(S_i) \quad (11)$$

ここで，寄与度 α_k は 0∼0.6 まで 0.2 刻み，β_p は 0 か $-\infty$ とした．$TotalScore()$ の高い文から順に全体の 3 分の 1 となるように抽出を行い，同スコアの場合には講演，またはニュースの後ろを重視し，後ろの文から順に抽出する．

6 評価実験

6.1 韻律情報と表層的言語情報における重要文抽出結果

韻律情報を利用した重要文抽出の結果を表 2 に示す．表 2 は韻律情報を用いて抽出したシステムの結果と人間 (**man3**) の結果との間の適合率である．また，表 3 に表層的言語情報を用いた際の結果，表 4 に棄却可能な特徴による結果を示す．表 2 から韻律情報では講演音声では P_B10_{pre5} と LEN が，ニュース音声では POW_{avg} が比較的高い適合率を得ていることが分かる．また，表層的言語情報においては講演音声とニュース音声ともに TF と $WORD_{rpt}$ が適合率が高い．また，講演音声では手がかり語 (CUE_{all})，ニュース音声では $TAIL_{10}$ の適合率が高い．表 4 から，LEN_{short} と $BETWEEN$ が重要文候補の棄却に有用であることが分かる．

表 2: 韻律情報を利用した重要文抽出結果

(a) 講演音声

特徴	κ 値	適合率	特徴	κ 値	適合率
P_B10_{pre5}	0.114	0.410	$F0_{avg}$	-0.108	0.148
P_B10_{pos5}	0.019	0.312	POW_{avg}	-0.072	0.183
LEN	0.219	0.441			

(b) ニュース音声

特徴	κ 値	適合率	特徴	κ 値	適合率
P_B3_{pre5}	-0.050	0.289	$F0_{avg}$	-0.011	0.348
P_B3_{pos5}	-0.135	0.221	POW_{avg}	0.043	0.445
LEN	-0.055	0.303			

表 3: 表層的言語情報を利用した重要文抽出結果

(a) 講演音声

特徴	κ 値	適合率	特徴	κ 値	適合率
CUE_{pre5}	0.154	0.522	TF(baseline)	0.243	0.459
CUE_{pos5}	0.094	0.531	$LEAD_{10}$	-0.033	0.160
CUE_{in5}	0.191	0.543	$TAIL_{10}$	0.043	0.460
CUE_{all}	0.228	0.522	$WORD_{rpt}$	0.287	0.486

(b) ニュース音声

特徴	κ 値	適合率	特徴	κ 値	適合率
$WORD_{rpt}$	0.144	0.463	$LEAD_{10}$	0.338	0.725
TF(baseline)	0.272	0.524	$TAIL_{10}$	0.010	0.363

6.2 組み合わせ実験結果

組み合わせパタンと各特徴の寄与度を表 5 に示す．‘#’ は削除する特徴であり，表の空欄は寄与度が 0 を意味している．表 6 に κ 値の最も良かった組み合わせの一致の度合いを示した．その結果，講演音声の場合，書き起こしを用いた際には κ 値=0.420，適合率=0.574，F 値=0.599 となり，ベースラインである TF と比較して高い値を得られた．これは，図 2 で示した **man3** と被験者間の κ 値と適合率とほぼ同等であり，人間並みの重要文抽出ができたと言える．また，音声認識による自動書き起こしデータを用いても，人手による書き

表 4: 削除する特徴の κ 値と適合率

(a) 講演音声

特徴	κ 値	適合率	特徴	κ 値	適合率
$F0_{low}$	-0.031	0.252	LEN_{short}	-0.175	0.104
POW_{low}	-0.101	0.164	$BETWEEN$	-0.036	0.050

(b) ニュース音声

特徴	κ 値	適合率	特徴	κ 値	適合率
$F0_{low}$	-0.011	0.377	LEN_{short}	-0.072	0.237
POW_{low}	0.047	0.394	$BETWEEN$	-0.008	0.000

起こしとほぼ同等の結果が得られた．また，ニュース音声においても複数の特徴を組み合わせることで κ 値=0.392，適合率=0.607，F 値=0.595 という値が得られた．これはベースラインである TF や $LEAD_{10}$ よりも高い値であり，**man3** と被験者間の結果 (図 2) と同等の結果が得られた．

表 5: 組み合わせパタンと各特徴の寄与度

(a) 講演音声

特徴	書き起こし	音声認識結果
CUE	○ 0.2	○ 0.2
$TAIL_{10}$	○ 0.4	○ 0.6
LEN		
TF	○ 0.2	○ 0.2
P_B10_{pre5}	○ 0.4	○ 0.6
$WORD_{prt}$	○ 0.4	○ 0.4
$\#LEN_{short}$	○	○
$\#BETWEEN$	○	○

(b) ニュース音声

特徴	書き起こし
$LEAD_{10}$	○ 0.6
$WORD_{rpt}$	
POW_{avg}	○ 0.4
LEN	○ 0.4
TF	○ 0.4
$\#LEN_{short}$	
$\#BETWEEN$	○

表 6: 組み合わせ実験結果

	κ 値	適合率	F 値
TF(baseline)(講演音声)	0.243	0.459	0.478
TF(baseline)(ニュース)	0.272	0.524	0.516
講演音声 (書き起こし)	0.420	0.574	0.599
講演音声 (音声認識結果)	0.384	0.550	0.574
ニュース音声	0.392	0.607	0.595

6.3 聴取実験

講演音声における最も良かった組み合わせ実験結果 (表 6) で得られた要約について要約された講演音声を作成した．これを機械による要約とする．被験者 8 名により機械の要約と人間の要約結果 (**man3**) を聴き比べ，要点 (講演の内容がつかみやすかったかどうか) と聴きやすさ (自然な音声に聴こえたかどうか) の 2 点について “人間の要約の方が良い”，“機械の要約の方が良い”，“どちらとも言えない” の 3 つの評価をした．

また，各被験者には人間と機械のどちらの要約音声を聴いているという情報を与えていない．一般に，人間は後に聴いた方が良いと判断する傾向があり，そういった傾向をなくすため，はじめに機械の要約を聴く被験者数と，はじめに人間の要約を聴く被験者数を同じにした．

表 7 に聴取実験結果を示す．この結果より要点については機械の要約よりも人間の要約の方が良いという意見が過半数を占め，人間の要約の方が優れている結果となった．この際，聴きやすさの面については，3 項目ともほぼ同数の意見となり，機械の要約が人間の要約とほぼ大差ないものとなったと考えられる．

表 7: 聴取実験結果

	人間 > 機械	どちらとも言えない	人間 < 機械
要点	21	10	9
聴きやすさ	13	15	12

7 まとめ

本研究では，人間による抽出要約の比較と，表層的言語情報および韻律情報を用いた講演音声と解説番組，ニュース音声の自動的な重要文抽出による要約を試みた．

人間が行った抽出要約においては，講演音声とニュース音声では被験者間の κ 値は 0.3 程度であり，解説番組では κ 値が 0.21 程度とさらに低い結果となった．このことから，解説番組は講演音声，ニュース音声と比較して話者の主張が不明瞭なため，あるいは同じ主張が複数の文に分散表現されておりどの文が重要であるかの判断が被験者によりばらついたのではないかと考える．

そこで，要約文集合の安定化を図るために，被験者 5 人中 3 人以上が重要文と認定した文を重要文とした要約 (**man3**) を基準とした．これと被験者の要約との κ 値は講演音声と「あすを読む」，ニュース音声で約 0.38 と少し高くなり，一方「視点・論点」においては κ 値が約 0.25 と依然低かった．

次に，韻律情報と表層的言語情報を用いた重要文抽出の手法について示した．また，韻律情報と表層的言語情報を組み合わせることによって，より良い要約ができることが示された．最終的に複数の特徴を組み合わせることにより，講演音声の書き起こしデータにおいて κ 値で 0.420，F 値で 0.599 という良い結果が得られ，人間並みの重要文抽出ができた．また，講演音声の音声認識による自動書き起こしデータにおいても F 値が 0.574 と従来手法 [13] と比べて比較的良い結果を得ることができた．ニュース音声についても κ 値が 0.392，F 値が 0.595 という良好な値が得られた．

参考文献

[1] Zechner, K.(2002). “Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Gence.” In ACL, pp.447-485.

[2] 長谷川将宏, 秋田祐哉, 河原達也 (2001). ”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, pp. 35-43.

[3] 下岡和也, 河原達也, 奥乃博 (2002). ”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.

[4] Siegel, S. and Castellan, N. (1988). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill.

[5] 野畑周, 関根聡, 井佐原均 (2003). ”異なるコーパスにおける重要文抽出の結果と素性の分析.”自然言語処理, 10 (5), pp. 93-120.

[6] 吉川裕規, 小林聡, 中川聖一 (2003). ”表層的言語情報と韻律情報を用いた講演音声の要約と評価.” 日本音響学会講演論文集, 3-Q-35, pp. 221-222.

[7] 小林聡, 吉川裕規, 中川聖一 (2002). ”表層情報と韻律情報を利用した講演音声の要約.” 情報処理学会音声言語情報処理研究会, 43-7, pp. 41-46.

[8] 伊藤山彦, 松本賢司, 谷田泰郎, 柏岡秀紀, 田中英輝 (2001). ”講演文を対象にした重要文抽出実験.” 話し言葉の科学と工学ワークショップ講演予稿集, pp. 157-164.

[9] 甲斐充彦, 廣瀬良太, 中川聖一 (1999). ”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, 40, pp. 1385-1394.

[10] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 (2003). 形態素解析システム『茶筌』version 2.3.3 使用説明書. 奈良先端科学技術大学院大学.
http://chasen.naist.jp/hiki/ChaSen/.

[11] 望月源 (2002). テキスト簡易要約器 Posum version 1.50.2 マニュアル. 北陸先端技術科学大学院大学.
http://www.tufs.ac.jp/ts/personal/motizuki/software/posumcl/.

[12] Salton, G. and Yang, C. S. (1973). ”On the specification of term values in automatic indexing.” In J.Documentation, Vol. 29, pp. 352-372.

[13] 北出裕, 南条浩輝, 河原達也, 奥乃博 (2003). “ 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, SLP-46, pp. 7-12.