

論文 / 著書情報  
Article / Book Information

論題(和文)	講演音声を対象とした要約手法の評価に関する検討
Title(English)	
著者(和文)	広畑 誠, 新中 庸介, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2004年秋季講演論文集, Vol. , No. 2-1-4, pp. 43-44
Citation(English)	, Vol. , No. 2-1-4, pp. 43-44
発行日 / Pub. date	2004, 9

## 1 はじめに

我々はこれまで様々な講演音声の自動要約手法を提案し、要約正解精度による客観評価を行ってきた [1-3]。しかし、客観評価と主観評価との関連性については十分に検討がなされていなかった。そこで本稿では、低い要約率の条件下での重要文抽出による要約において、様々な要約手法より得られた要約結果に対し、幾つかの客観評価と主観評価との関連について報告する。さらにその結果に基づいて、これらの客観評価値が、要約手法を評価する上で有効であるかについて検討する。

## 2 自動要約手法

講演音声の音声認識結果に対して、文境界判定を行う。自動的に区切られた結果を用いて、あらかじめ与えられた要約率に従い、以下に示す手法を用いて重要文抽出により要約する。

### 2.1 重要度スコアを用いた文抽出

重要文抽出を情報量に基づき各文に付したスコアを用いて行う [2]。各文のスコアは、その文に出現した内容語 (名詞、記号) の重要度スコアの総和を文の単語数で正規化することで求める。求められたスコアの高い文から順に抽出する (以下 SIG と表記)。内容語  $w$  における重要度スコア  $I_w$  は、音声認識結果中の内容語  $w$  の出現頻度  $f_w$  と、大規模コーパス中での内容語  $w$  の出現頻度に基づく値 ( $ICF$ ) の積で表される。

$$I_w = f_w \cdot ICF = f_w \cdot \log \frac{F_A}{F_w} \quad (1)$$

$F_w$ : 大規模コーパス中での内容語  $w$  の出現頻度

$F_A$ : 大規模コーパス中での総内容語数 ( $= \sum_i F_i$ )

大規模コーパスには話し言葉コーパス (CSJ) の講演書き起こし (約 8M 形態素) のテキストを用い、出現した約 50k 種類の単語の出現頻度を求めた。

### 2.2 潜在的意味分析を用いた文抽出

講演全体の構成を考慮するため、特異値分解を用いて潜在的意味分析を行い、文を抽出する [3]。まず音声認識結果について、内容語を行に、文を列に持つ行列を用意する。行列の要素は、その文における内容語の出現頻度と  $ICF$  の積で定義する。この行列に対し特異値分解を用いることで、単語と文のクラスタリングが行われ、文を特異値によって重み付けられたベクトル空間に射影できる。射影された文ベクトルにおいて、各特異値に対応する成分で最も大きな値を有する文を重要文と見なし抽出する (以下 LSA と表記)。

### 2.3 次元圧縮を用いた文抽出

文抽出手法 LSA と同様に特異値分解を用いる [3]。利用する特異値の数を減らすことで、各文が射影されているベクトル空間の次元を圧縮し、射影されたベクトルのノルムが大きい文から順に抽出する (以下 DIM と表記)。今回の実験では、圧縮後の次元数を 5 とした。

### 2.4 序論および結論からの文抽出

予備的な検討の結果、各講演が序論、本論、結論の 3 段落から構成されていると仮定すると、人間が要約する際には、序論、結論から多くの文を抽出する傾向があることが分かった。そこで、このような位置の情報を文抽出に利用する。まず、講演音声の認識結果に対し、話題の切れ目を

推定し、段落毎に分割する。Hearst 法を用い、各内容語境界の結束度を求める。結束度は前後の一定幅を持つ内容語列に対し、出現頻度ベクトル同士の余弦尺度で表す。結束度のうち、一定値以下であり、極小点である境界を話題の切れ目の候補とする。話題の切れ目の候補のうち、最初の候補を序論と本論の境界、最後の候補を本論と結論の境界と見なす。要約する際には序論および結論と推定した部分からのみ文を抽出する (以下 IC と表記)。この手法は、上記の 3 つの手法と組み合わせて利用する。

## 3 客観評価

自動要約結果の客観評価には、3 名の被験者が重要文を抽出することで作成した要約 (正解要約) を利用し、以下に示す方法で行う。

### 3.1 要約正解精度

被験者の作成した要約文をもとに、正解要約文の単語連鎖をまとめた正解要約文単語ネットワークを用いて評価を行う [2]。自動要約文に最も近い単語列をネットワーク中から抽出し、その単語列に対する単語正解精度 (SumACCY) を求める。また、ネットワークを用いずに被験者毎の正解要約に対する自動要約文の単語正解精度を求め、そのうち最も評価値の高いもの (SumACCY-E/max) と、平均値 (SumACCY-E/ave) についても検討を行う。

### 3.2 重要文抽出精度

自動要約により抽出された文を人手の文区切りにおける文に対応させ、その文を抽出したとして、文抽出の再現率、適合率を求める [4]。今回の実験では、各被験者の要約に対し F 値を求め、最も評価値の高いもの (F-measure/max) と、平均値 (F-measure/ave) について検討する。

### 3.3 ROUGE-N

自動要約と正解要約との単語列の重なりを調べるため、n-gram の再現率を求める [5]。今回は 1-gram, 2-gram, 3-gram について調べる。この再現率は以下に示す式で定義される。

$$\text{ROUGE-N} = \frac{\sum_{S \in S_H} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)} \quad (2)$$

$S_H$ : 被験者 3 名による正解要約の集合

$S$ : 各被験者の正解要約

$g_n$ : 単語列 n-gram

$C(g_n)$ : 正解要約における  $g_n$  の出現頻度

$C_m(g_n)$ : 正解要約と自動要約での  $g_n$  の共起頻度

## 4 評価実験

### 4.1 自動要約実験

CSJ の男性話者 20 名、女性話者 10 名の計 30 名による講演音声に対し、文献 [6] の音声認識システムを用いて得られた音声認識結果について、10% の要約率で重要文抽出による要約を行った。各講演に対し、SIG, LSA, DIM, SIG+IC, LSA+IC, DIM+IC の計 6 種類の文抽出手法 (要約手法) を適用する。なお、音声認識結果の単語正解精度は約 69% で、文境界判定の再現率/適合率は約 72%/75% であった。

### 4.2 主観評価試験

講演音声の要約に対し、要約としての相応しさを評価するため、主観評価試験を行った。評定対象である要約の数

\* A study on evaluation of summarization techniques for lecture speech

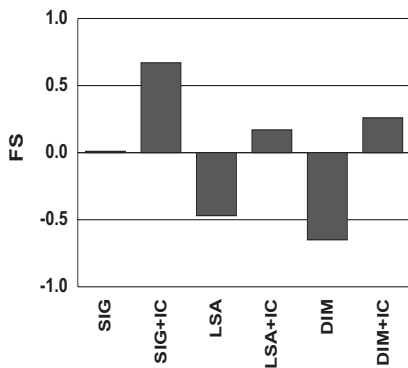


図 1. 要約手法の主観評価結果

は 180(30 講演 × 6 要約手法) で、評定者は 12 名である。各講演の要約に対し、人数の偏りが無いように評定を行った。評定者は各講演における評定対象の要約に目を通した後、講演の書き起こしを読み、「1. 非常に悪い - 5. 非常に良い」の 5 段階で要約の評定を行う。また、評定の基準を明確にするため、評定者にはあらかじめ、評定対象外の講演において、被験者が作成した要約 (認識誤りを含める) を「非常に良い」要約の例として提示した。その際、その講演の書き起こしも評定者に提示した。

主観評価実験から得られた評定得点に対し、評定者間の評価の偏りによる精度の劣化を考慮するため、因子分析 [7] により、因子得点 (FS) を算出した。

#### 4.3 要約手法の主観評価結果

各要約手法の主観評価結果を図 1 に示す。結果より、3 つの要約手法 SIG, LSA, DIM においては、それぞれ IC の導入があるか否かにより有意差が見受けられた。このことから、IC の導入により要約の精度が改善されることが示された。また、IC を利用した上で、3 つの要約手法を比較すると、SIG+IC と LSA+IC の間には有意差が見られたが、SIG+IC と DIM+IC, DIM+IC と LSA+IC の間に有意差が見られなかったため、3 つの要約手法に対して、明白な順位付けをすることは難しいと考えられる。なお、有意差検定には、5%有意水準による片側検定を用いた。

#### 4.4 客観評価と主観評価との関連

まず、計 6 種類の要約手法における 3 で示した客観評価と主観評価結果 (30 講演に対する因子得点の平均値) との関連性について考察する。比較のため客観評価とは別に、要約結果に対する音声認識精度 (単語正解精度: ACCY) との関連性についても調査する。客観評価と主観評価から求められた単相関の値を図 2 に示す。検定の結果、各客観評価と主観評価との間に相関があることが確認できた。また、各客観評価において、ACCY の影響を取り除き主観評価との偏相関係数を求めた。その結果、ROUGE-1 以外の客観評価と主観評価との関連においては、見かけ上の相関では無いことが確認された。今回の実験条件下では、ROUGE-1 以外の客観評価は、要約手法の評価として有効であると考えられる。一方で、客観評価間での有意差は認められなかった。

次に、客観評価と講演および要約手法の組合せの異なる総数 180 の要約結果に対する主観評価結果 (因子得点) との間の単相関の値を図 3 に示す。この結果から、強い相関は確認されなかった。これは、客観評価と主観評価との関連性に講演毎のばらつきがあることが特に影響していると考えられる。

### 5 まとめ

本稿では、10%要約率の条件下での講演音声の自動要約において、計 6 種類の要約手法の主観評価実験を行い、幾つかの客観評価と主観評価との関連について検討した。その結果、要約手法の比較を行う上で、ROUGE-1 以外の

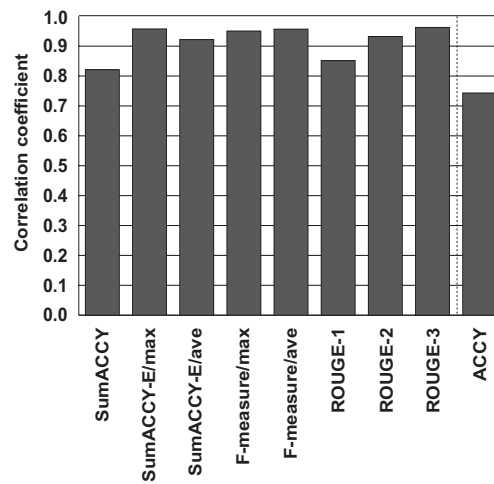


図 2. 各要約手法に対する客観評価と主観評価の単相関

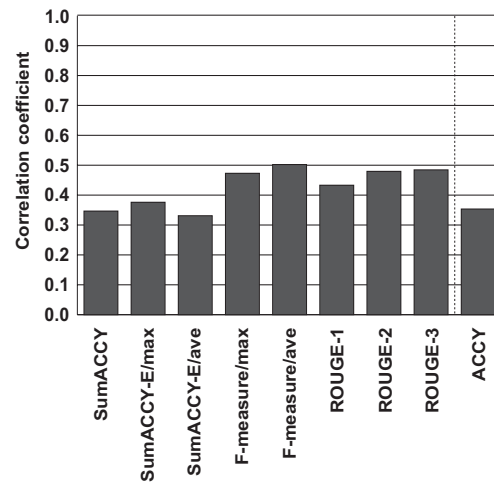


図 3. 講演および要約手法の組合せの異なる要約結果に対する客観評価と主観評価の単相関

客観評価の有効性が確認できた。

今後の課題としては、評定結果を増やすことによる信頼性の向上やより多くの文抽出手法との比較、他の客観評価法と主観評価との関連の検討があげられる。また、単語抽出による要約手法における検討も今後の課題である。

謝辞 主観評価方法について貴重なご助言を頂いた当大学院人間行動システム専攻中川正宣教授に感謝致します。

#### 参考文献

- 堀智織, 古井貞照, “講演音声の自動要約の試み,” 話し言葉の科学と工学ワークショップ講演予稿集, pp.165-171 (2001-3).
- 菊池智紀, 古井貞照, 堀智織, “重要文抽出と文圧縮による音声自動要約,” 信学技報, SP2002-158, pp.61-66 (2002-12).
- 広畑誠, 新中庸介, 古井貞照, “音声自動要約における SVD を用いた重要文抽出手法の検討,” 日本音響学会 2003 年秋季講演論文集, 2-6-17, pp.93-94 (2003-9).
- 北出祐, 南條浩輝, 河原達也, “談話標識と話題語を用いた重要文抽出手法の CSJ の学会講演における評価,” 話し言葉の科学と工学ワークショップ講演予稿集, pp.111-118 (2004-2).
- C.Y.Lin, “Looking for a few good metrics: ROUGE and its evaluation,” Proc. NTCIR-4, vol.2, pp.1-8 (2004-6).
- T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui, “Benchmark test for speech recognition using the corpus of spontaneous Japanese,” Proc. SSPR2003, pp.135-138 (2003-4).
- P.D.Allison, “Estimation of linear models with incomplete data,” Sociological Methodology, pp.71-103 (1987).