論文 / 著書情報
Article / Book Information

論題(和文)	客観尺度を用いた音声自動要約手法の評価
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)	音声音響学会2005年春季講演論文集, Vol. , No. 1-5-2, pp. 3-4
Citation(English)	, Vol. , No. 1-5-2, pp. 3-4
発行日 / Pub. date	2005, 3

1—5—2 客観尺度を用いた音声自動要約手法の評価 *

◎広畑 誠　新中 庸介　岩野 公司　古井 貞熙 (東工大)

1 はじめに

我々はこれまで講演音声の自動要約手法として，講演の構成を考慮した，特異値分解による次元圧縮を用いた手法 [1] や講演内の位置情報を利用した手法 [2] を提案してきた．また，音声認識誤りの少ない文を抽出するため，信頼度，言語スコアを利用した手法についても検討してきた [3]．そこで本稿では，この [1], [2] の手法に対して，信頼度，言語スコアの導入を試みることで，要約の高精度化を図る．自動生成された要約に対し，主観評価との相関が高いと考えられる客観尺度を用いて評価を行った結果を報告する．

2 重要文抽出による音声自動要約

講演音声の認識結果を自動的に文単位に分割した後，あらかじめ与えられた要約率に従い，文を抽出することで要約を行う．以下では，本研究で用いた文抽出手法について説明する．

2.1 次元圧縮を用いた文抽出

話題単語を多く含み，かつ，他の多くの文とも意味的に類似した文を重要文とみなし抽出する手法として，特異値分解による次元圧縮を用いた手法を提案している [1]．まず，各文 i のベクトル $A_i = [a_{1i}, a_{2i}, \ldots, a_{Ji}]^\top$ の j 番目の要素 a_{ji} を式 (1) を用いて求める．ただし，J は語彙数である．

$$a_{ji} = B(f_{ji}) \cdot \log(F_A/F_j) \tag{1}$$

f_{ji} : 文 i での内容語 j の出現頻度
F_j : 大規模コーパス中での内容語 j の出現頻度
F_A : 大規模コーパス中での総内容語数 $(= \sum_w F_w)$
B は，値が 0 より大きい場合，常に値 1 を返す関数である．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．

図 1 のプロセスに従い，各文のベクトルは特異値分解により，特異値 σ_k，右特異行列の要素 v_{ik} を用いて K 次元空間で表現できる．ここでは 1 次元空間におけるノルムを文のスコアとし，スコアの高い文から順に抽出する (以下この手法を DIM と表記)．

$$D(i) = |\sigma_1 v_{i1}| \tag{2}$$

2.2 信頼度，言語スコアを利用した文抽出

次元圧縮を用いた手法により求められたスコア D に信頼度スコア C，言語スコア L を重み付けて加えることで，各文 i に統合スコア S を与える (式 (3))．要約は，この統合スコアによって行う (以下この手法を COMB と表記)．

$$S(i) = D(i) + \lambda_C C(i) + \lambda_L L(i) \tag{3}$$

なお，λ_C，λ_L は，各スコアのバランスをとるための重み係数である．以下に信頼度，言語スコアについて説明する．

信頼度スコア

各文 i が N_i 個の単語からなる認識単語列 $W_i = w_1, w_2, \ldots, w_{N_i}$ のとき，各単語 w_n における信頼度 $c(w_n)$ を用い，信頼度スコアを求める．なお，$c(w_n)$ は，音声認識デコーダから出力された音響尤度および言語尤度に基づく事後確率の対数値で定義される．

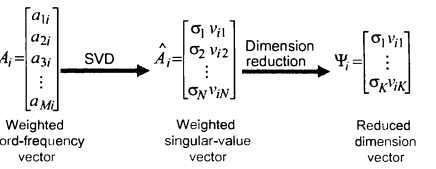

図 1. 特異値分解を用いた次元圧縮のプロセス

$$C(i) = \frac{1}{N_i} \sum_{n=1}^{N_i} c(w_n) \tag{4}$$

言語スコア

CSJ から作成した単語 trigram を用い，単語連鎖の適正度として各文の言語スコアを求める．

$$L(i) = \frac{1}{N_i} \sum_{n=1}^{N_i} \log P(w_n | w_{n-2} w_{n-1}) \tag{5}$$

2.3 序論および結論からの文抽出

要約率が 10%のように低い場合，被験者は序論および結論と考えられる部分から多くの文を抽出する傾向がある．その傾向を要約手法に反映させるため，位置の情報を文抽出に利用する [2]．Hearst 法を用い，講演における各内容語境界の結束度を求めることで，話題の切れ目を推定する．序論および結論部分を最初と最後の話題の切れ目の候補によって定める．そして，序論および結論と定めれた部分からスコア上位の文を抽出する (以下この手法を IC と表記)．この手法を，2.1, 2.2 節で述べた手法と組み合わせて利用する．

3 音声自動要約の客観評価方法

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) を利用する．要約率 10%の条件下で主観評価との相関が特に高かった客観尺度で要約手法の評価を行う．以下，客観評価方法について説明する．

3.1 要約正解精度

複数の被験者が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 [3]．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 (以下 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) も評価値として用いる．

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，

* Evaluation of speech summarization techniques using objective metrics
By Makoto Hirohata, Yousuke Shinnaka, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

表 1. 要約率 50%の条件下での各要約手法の評価結果

Technique	SumACCY	SumACCY-E	F-measure	ROUGE-3	ACCY
RDM	33.7%	31.8%	0.525	0.319	70.4%
DIM	39.9%	35.6%	0.572	0.349	71.4%
COMB	42.2%	37.2%	0.556	0.371	75.9%
C only	39.9%	35.6%	0.551	0.361	76.1%
L only	38.9%	34.0%	0.529	0.356	74.5%

表 2. 要約率 10%の条件下での各要約手法の評価結果

Technique	SumACCY	SumACCY-E	F-measure	ROUGE-3	ACCY
RDM	-7.3%	14.7%	0.102	0.083	70.3%
DIM	2.9%	16.7%	0.137	0.105	71.3%
DIM + IC	11.1%	21.3%	0.193	0.132	72.0%
COMB + IC	11.5%	23.8%	0.199	0.144	76.6%
C only	0.7%	17.4%	0.126	0.116	83.3%
L only	1.0%	17.6%	0.101	0.112	81.9%

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，その平均値 (F-measure) を用いる．ただし，この評価は要約文中に現れる認識誤りを考慮していない．

3.3 ROUGE-3

自動要約と正解要約との単語列の重なりを調べるため，単語 trigram の再現率 (ROUGE-3)[5] を求める．

4 評価実験

4.1 要約実験条件

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音声認識結果を得た．単語正解精度は約 69% (フィラーや言い誤り等を除けば約 70%) で，自動文区切りの再現率/適合率は約 72%/75%であった．得られた認識結果に対し，50%と 10%の要約率において評価実験を行った．

COMB では，各スコアについて平均値が 1 になるように線形変換を行った後，式 (3) の λ_C, λ_L は 5.0, 2.0, 1.0, 0.5, 0.2, 0.1, 0.0 のいずれかに設定し，各客観評価値が最もバランスよく改善されるものを採用する．参考のため，信頼度スコア C，言語スコア L を単独で用いたときの結果も示す．また，重要文抽出手法の有効性を検証するため，文をランダムに抽出する手法 (RDM) の評価も行う．

4.2 要約率 50%

表 1 に要約率が 50%のときの評価結果を示す．参考として，要約結果に対する音声認識精度 (単語正解精度: ACCY) についても示している．なお，COMB に関しては，$(\lambda_C, \lambda_L) = (5.0, 1.0)$ のときの結果を示す．

DIM では，RDM に比べて評価値がそれぞれ 4〜6%改善されている．さらに COMB では，F-measure 以外の評価値が DIM に比べ 2%程改善されている．C, L のみを利用した場合には，要約文中の認識誤りが少なくなり，ACCY の改善や F-measure 以外の要約評価値の改善がみられる．この C, L の特性が反映されたため，COMB では，特に F-measure 以外の改善が得られたものと考えられる．

4.3 要約率 10%

表 2 に要約率が 10%のときの評価結果を示す．なお，COMB + IC に関しては，$(\lambda_C, \lambda_L) = (2.0, 0.0)$ のときの結果を示す．

IC を考慮しない場合，DIM を用いると RDM に対し，SumACCY は約 10%の改善がみられ，他の評価値は 2〜4%の改善がみられた．DIM + IC では，DIM に比べ 3〜8%の改善がそれぞれの評価値で得られ，さらに C, L を考慮した COMB + IC によって，SumACCY-E を約 3%，F-measure, ROUGE-3 を約 1%改善できた．これは，C, L の認識誤りを削減する効果に加え，IC による位置の推定精度が向上したためであると考えられる．

5 まとめ

本稿では，特異値分解や位置情報といった講演の構成に関する特徴を利用した手法に対し，信頼度，言語スコアを導入し，種々の客観尺度を用いて要約結果の評価を行った．要約率 50%の条件下では，次元圧縮を用いた手法に，信頼度，言語スコアを導入することで，導入前に比べ要約中の認識誤り単語の数が減り，要約正解精度や ROUGE-3 に改善がみられた．また，要約率が 10%の場合，序論および結論からの文抽出により，F-measure をはじめとした各客観評価値に大きな改善がみられ，さらに信頼度スコアを利用することで，その特性を反映することができた．

今後の課題としては，単語抽出が組み込まれた際の検討があげられる．また，位置情報をより詳細に利用した話題セグメントからの文抽出手法 [7] に対して，信頼度，言語スコアの導入を検討していく必要がある．

謝辞 本研究の一部は，文部科学省 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の支援を受けて行われた．

参 考 文 献

[1] 広畑 他，音講論，2-6-17, pp.93-94 (2003-9)
[2] 広畑 他，音講論，2-1-4, pp.43-44 (2004-9)
[3] 菊池 他，信学技報，SP2002-158, pp.61-66 (2002-12)
[4] 北出 他，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)
[5] C.Y.Lin, Proc. NTCIR-4, vol.supl.2, pp.1-8 (2004-6)
[6] T. Kawahara, et al., Proc. SSPR2003, pp.135-138 (2003-4)
[7] 新中 他，音講論，2-1-3, pp.41-42 (2004-9)