信頼尺度を用いた音声自動要約の改善 *

◎堀 智織　　古井 貞熙　　(東工大)

1. はじめに

これまで我々は，単語重要度と言語尤度を用いて，放送ニュース音声を発話毎に要約する音声自動要約手法を提案し [1][2]，さらに，提案手法により生成された自動要約文に対し，人間の作成した正解要約文に基づき定量的に評価する方法を検討してきた [3]．本稿では，音声認識結果に含まれる認識誤りによる誤要約を防ぐため，単語グラフに基づく信頼尺度を考慮した要約スコアを用いることにより，音声自動要約手法の改善を試みた．NHK のニュース音声 50 発話を大語彙連続音声認識システムを用いて音声認識し，5 段階の要約率の文字数に自動要約した．生成された自動要約文を，25 人の被験者が作成した正解要約文に基づき評価した．実験結果に基づき，信頼尺度による自動要約文の改善の効果を報告する．

2. 信頼尺度を考慮した音声自動要約手法

我々が提案している音声要約の枠組では，認識された各発話文から相対的に重要な単語を，原文の文字数に対する要約文の文字数の割合 (要約率) の特定の範囲で抽出し，それらを接合することにより要約文を生成する．これまで提案してきた音声自動要約手法では，要約文に抽出された各単語の単語重要度 (重要度スコア) と言語尤度 (言語スコア) の累積スコアを，要約文の尤もらしさを示す要約スコアと定義し，要約スコアを最大とする部分単語列を最適な要約文として動的計画法を用いて求める [1][2]．この手法は，テキスト要約への応用も可能であるが，音声要約では認識結果に含まれる誤りを考慮しなければならない．認識誤りによる誤要約を防ぐため，本稿では，認識された各単語が正解である信頼度を音響的，言語的に検証することにより，音声自動要約文の誤要約の改善を検討する．

要約スコアは，単語重要度スコア I と言語スコア L，および信頼度スコア C に基づき，次式のように定義する．N 個の単語からなる認識単語列 $W = w_1, w_2, \ldots, w_N$ から要約文として M $(M < N)$ 個の単語を抽出し接合した単語列 $V = v_1, v_2, \ldots, v_M$ の要約スコアは次式によって示される．

$$S(V) = \sum_{m=1}^{M} \{L(v_m) + \lambda_I I(v_m) + \lambda_C C(v_m)\} \qquad (1)$$

但し，λ_I，λ_C は各スコアのバランスをとるための重み係数である．

単語重要度スコア

単語重要度スコア I (v_m) は，文中における単語の重要度を示すスコアである．本研究では，名詞の単語重要度スコアとして話題語らしさを示す話題語スコアを適用する．話題語スコアには重み付き TF・IDF 尺度を用いる [2]．

言語スコア

言語スコア L (v_m) は，単語連鎖の適正度を示すスコアである．本研究では，統計的言語モデルである単語 trigram を用いる．

信頼度スコア

信頼度スコア C (v_m) は，認識結果に含まれる認識誤りを要約文に抽出しないよう，音響的，言語的に信頼度の低い単語を含む要約文候補に対しペナルティを与えるものである．デコーダから出力された単語グラフに付与された音響尤度および言語尤度に基づく各単語に対する事後確率を，信頼尺度として用いる [4][5]．図 1 で示すように，単語グラフは文頭ノード S から文末ノード T に至る各ノードとノード間を接続するリンクによって表される．単語間境界を示すノードには時間情報が格納され，単語を示すリンクには各単語の音響尤度と言語尤度が格納されている．

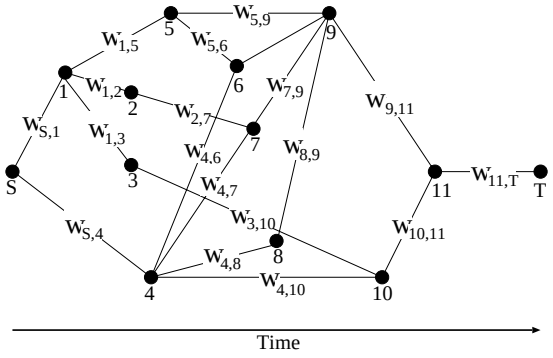

図 1．単語グラフの例

ノード間 k，l に単語仮説 $w_{k,l}$ が出現する事後確率は，次式のように forward 確率と backward 確率によって求められ，これを信頼度スコアとする．

$$C(w_{k,l}) = \log \frac{\alpha_k P_a(w_{k,l}) P_l(w_{k,l}) \beta_l}{\mathcal{G}} \qquad (2)$$

k,l	:単語グラフにおけるノード番号 $(k < l)$
$w_{k,l}$	:ノード k,l 間に生成した単語
$C(w_{k,l})$	:単語 $(w_{k,l})$ の事後確率
α_k	:始端 S からノード k までの forward 確率
β_l	:ノード l から終端 T までの backward 確率
$P_a(w_{k,l})$	:単語 $w_{k,l}$ の音響尤度
$P_l(w_{k,l})$	:単語 $w_{k,l}$ の言語尤度
$\mathcal{G}$	:単語グラフ始端から終端までの forward 確率

この信頼度スコアは，認識された各単語と単語グラフにおける対立候補の尤度比を示す値であり，値が大きいほど高い信頼度で認識されたとみなすことができる．

認識された単語列より抽出された部分単語列を $V = v_1, v_2, \ldots, v_M (M < N)$ とするとき，要約処理は (1) 式で表される要約スコアを最大にする $\hat{V}$ を求める問題となり，動的計画法を用いて解くことができる．さらに，同一の認識結果から生成された要約率の異なる要約文を比較するため，単語数に基づく正規化要約スコアを定義した [1][2]．

3. 自動生成要約文に対する評価尺度

人間によって作成された正解要約文に基づき，原文の重要な情報の抽出，日本語としての適正度，原文の文意の保持という 3 点から，自動要約文を定量的に評価する [3]．

---

* An Improvement of Automatic Speech Summarization Using a Confidence Measure
By Chiori Hori and Sadaoki Furui (Tokyo Institute of Technology)

重要単語の抽出率

原文の情報の保持という点から要約文を評価するため，正解要約文における各単語の被験者による選択率を単語の重要度とし，自動要約文へ抽出された重要単語の抽出率を評価する．

単語連鎖の適合率

日本語としての尤もらしさと原文の文意の保持を評価するため，自動要約文中に含まれる特定の長さの単語連鎖が正解要約文に含まれるか否かを評価する．自動要約文中に含まれる特定の長さの各単語連鎖が，被験者の作成した正解要約文のいずれかに含まれる割合を単語連鎖適合率とする．但し，原文の異なる位置に出現する単語は，たとえ表記が同一であっても別の単語として扱う．

4. 音声自動要約実験と評価

実験および評価

放送ニュース音声を音声認識し，単語正解精度が 90%以上の 50 発話について提案する音声自動要約手法を用い，20%,40%,60%,70%,80%の 5 段階の要約率で要約文を生成した．要約用言語モデルとして 3 年分の毎日新聞 (1996–1998) を用いた．音声認識結果と書き起こし文 (TRS) に対し要約文を自動生成した．音声認識結果を自動要約する際，信頼度スコアを用いた要約文 (CM) と用いない要約文 (REC) の 2 種類を生成し，信頼度スコアの効果を検討した．各自動要約文は，25 人の被験者によって作成された正解要約文に基づき，重要単語抽出率と単語連鎖適合率により評価した．さらに，テストセットで 50 文のうち認識率の比較的低い 10 文を用い，低認識率の際の信頼度スコアの効果を検討した．

さらに，各被験者の要約文 (SUB) をその他の 24 人の被験者の要約文を正解として評価し，自動要約文の目標値とした．提案手法の有効性を示すため，無作為に単語抽出した要約文 (RDM) を生成し，手法間の比較を行った．

評価結果

重要単語抽出率による評価結果を表 1 に示す．

表 1．重要単語抽出率による評価結果

要約率	20%	40%	60%	70%	80%
RDM	0.17	0.35	0.54	0.66	0.75
REC	0.20	0.38	0.58	0.69	0.77
CM	0.27	0.41	0.57	0.68	0.75
TRS	0.31	0.43	0.60	0.71	0.78
SUB	0.45	0.56	0.68	0.75	0.80

要約率 60%以上では要約文に抽出される単語数が多いため，正解要約文の重要単語抽出率に手法間の有意な差はない．しかし，要約率 40%以下では，抽出する単語の選択肢が増えるため，手法により重要単語抽出率に有意な差がある．信頼度スコアを用いた音声自動要約文 CM は，信頼度スコアを用いない REC に比較し抽出率が改善されている．

要約率 70%における，単語連鎖適合率と単語連鎖数の関係を図 2 に示す．単語連鎖が長くなるに従い，単語間の文法的，意味的制約が厳密になるため，全ての要約文で単語連鎖適合率の低下が起こる．無作為に単語を選択した要約文 RDM の適合率の急速な低下は，RDM 中の文の連接が文法的に日本語として不適切かつ原文の文意に適合しない単語連鎖が多く含まれていることを示す．RDM に比較して，提案手法により自動生成された要約文 TRS，REC，CM は，長い単語連鎖においても適合率が維持されている．しかしながら，目標である人間の作成した要約文の適合率には達していない．これは，意味的なまとまりに関する情報の欠落によるものと考えられる．なお，70%以外の要約率においても，要約手法間における傾向は要約率によらず一定であった．要約率間で比較すると，低い要約率ほど単語の選択肢が多くなるため，単語連鎖の長さに伴う単語連鎖適合率の急速な低下がみられた．

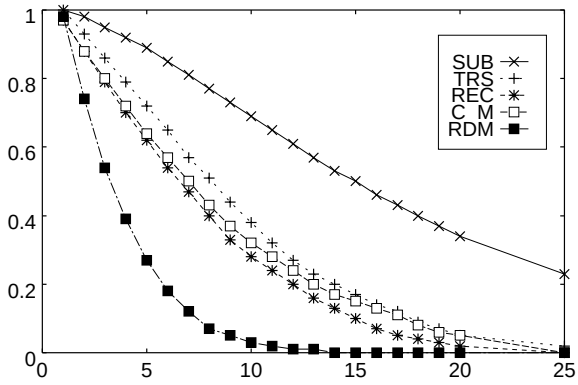

図 2．要約率 70%時の全テストセットにおける単語連鎖適合率と単語連鎖長の関係

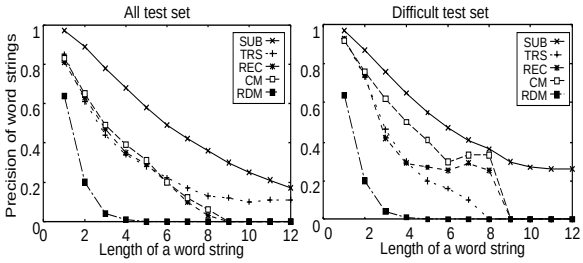

図 3．要約率 20%時の全テストセットと低認識率テストセットにおける単語連鎖適合率と単語連鎖長の関係

さらに，認識率が低い場合での信頼度スコアの有効性を示すため，要約率 20%で要約したテストセット 50 文 (all test set) と 50 文中認識率の低い 10 文 (difficult test set) について評価を行った結果を図 3 に示す．信頼度スコアを用いることにより，認識率の低い認識結果から作成した要約文が改善されていることが分かる．

5. まとめ

音声認識結果に含まれる認識誤りによる誤要約を改善するため，単語重要度と言語尤度に基づく要約スコアに信頼尺度を加え，自動要約文を生成し評価を行った．重要単語抽出率，単語連鎖適合率のどちらにおいても信頼尺度による改善がみられた．正解要約文の単語連鎖適合率に到達するために，さらに意味のまとまりを考慮した自動要約手法を検討する必要がある．

謝辞

放送ニュースのデータベースを提供して下さった NHK 放送技研に感謝致します．

参考文献

[1] 堀 他: 信学技報, SP99-110,pp.103–108(1999).

[2] Hori et al.: Proc. ICASSP2000, Istanbul, Vol.3, pp.1579-1582(2000).

[3] 堀 他: 音講論, Vol.1, 2-8-18, P63-64(2000).

[4] Kemp et al.: Proc. 5th Eurospeech '97 , Rhodes, Vol.2, pp.827-830 (1997).

[5] Valtchev et al.: Speech Communication Vol.22, pp.303-314 (1997).