2003－NL－155 (14)
2003－SLP－46 (3)
2003／5／27

音声自動要約を利用した講演速聞きシステムの検討

新中 庸介　　菊池 智紀　　岩野 公司　　古井 貞熙

東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
Email: {shinnaka, kikuchi, iwano, furui}@furui.cs.titech.ac.jp

本論文では，我々がこれまで提案してきた音声自動要約手法を用いて，要約音声を作成し，ユーザに提示する講演音声の「速聞きシステム」を提案する．要約音声は，テキストで出力された音声自動要約結果に対応する音声を，元の音声から切り出し，接続することにより作成し,「聞きやすさ」「要約内容」の 2 項目で評価する．評価には要約を行う際の削除・選択単位が影響する．そこで，要約音声にふさわしい削除・選択単位を検討するため，単語単位，文単位，文をさらにフィラーで区切った単位の 3 つの単位の削除・選択により作成した要約音声について，聴取実験を行った．その結果，要約率 50% では文単位要約が有効であることが確認された．

A Rapid Listening System for Presentations Using Automatic Speech Summarization Techniques

Yousuke Shinnaka, Tomonori Kikuchi, Koji Iwano, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: {shinnaka, kikuchi, iwano, furui}@furui.cs.titech.ac.jp

This paper proposes a ”rapid listening” system for presentations, which produces summarized speech using automatic speech summarization techniques. The summarized speech is produced by extracting speech units corresponding to textual summary from the original speech and concatenating them into a presentation. Three kinds of units, words, sentences and between-filler units, are investigated. Summarized speeches using these units are subjectively evaluated in terms of the ease of listening and the quality of summary. Experimental results show that the sentence units are most appropriate for summarization at 50% summarization ratio.

1　はじめに

近年，録音メディアの普及に伴い，講演・講義・会議等の大量の音声データが蓄積され，それらの参照機会が増加している．しかし，大量の音声データからの検索や参照には，多くの時間や労力が必要となる．このため，音声データから冗長な部分を除き，話し手が伝えようとした内容を抽出する音声自動要約の需要が高まっており，我々の研究室でも音声自動要約の研究 [1][2] が行われている．音声自動要約を利用することにより，音声データから抽出した内容をインデクス化して検索に用いたり，参照データを少量のデータにすることが可能である．

音声自動要約の要約結果の提示方法には，音声認識を基に作成したテキストで提示する方法と，音声で提示する方法があるが，我々の研究室で提案されている音声自動要約手法は，要約結果をテキストで提示している．しかし，テキストで提示する場合，認識誤りにより読みづらく，誤った情報をユーザに伝える可能性があり，かつ，音声にのみ存在する話者の意図や感情といった情報を伝えることが困難である．これに対し，音声で提示する場合，これらの悪影響を取り除くことができる．本論文では，講演音声について，音声自動要約手法を用いて要約音声を作成し，ユーザに提示する「速聞きシステム」の提案を行う．

本システムでは我々の研究室で提案されている音声自動要約手法 [1][2] を利用して要約音声を作成する．要約音声を提示する研究 [3][4] はこれまでにもなされているが，音声認識結果に含まれる単語重要度や連鎖確率といった言語的な情報を用いている研究はない．本システムではこれらの言語的な情報を用いて要約を行っている．

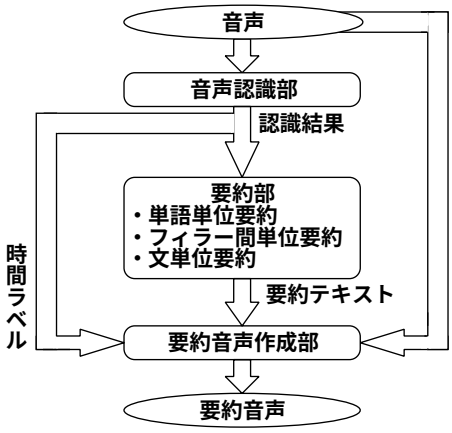

図 1: 速聞きシステムの構成

2 速聞きシステム

提案する速聞きシステムの構成を図 1 に示す．対象とした音声データは,「日本語話し言葉コーパス (CSJ)」の講演音声である．

CSJ のラベルに基づき，音声を 500 ミリ秒以上の無音区間で区切り，その内 1 秒以内の音声は隣り合う音声と接続，20 秒以上の音声はさらに分割して，1 秒以上 20 秒未満の音声を「文単位」とする．このように分けられた音声を音声認識部で認識し，その認識結果について要約部で要約を行い，要約テキストを出力する．要約部では単語単位要約方式，文単位要約方式，文単位をさらにフィラーで分けた単位で要約するフィラー間単位要約方式の 3 つの要約方式を用いることができる．単語単位要約方式は我々の研究室で提案されている単語抽出手法 [1]，文単位要約方式，フィラー間単位要約方式は我々の研究室で提案されている重要文抽出手法 [2] を利用して要約を行う．要約音声作成部では，音声と要約テキストの対応を認識結果の時間ラベルから取り，音声を切り出し，接続することにより要約音声を作成する．以下で，要約部と要約音声作成部について説明する．

2.1 要約部

3 つの単位の要約方式で用いる単語抽出手法 [1]，重要文抽出手法 [2] の 2 つの要約手法について始めに説明し，次に 3 つの単位の要約方式について説明する．

2.1.1 要約手法

単語抽出手法

単語抽出は，$V = v_1, v_2, \ldots, v_M$ を単語抽出後の部分単語列とすると，以下の式で表される要約スコア $S(V)$ を基に要約を行う．

$$S(V) = \sum_{i=1}^{M} \{ L(v_i) + \lambda_I I(v_i) + \lambda_C C(v_i) + \lambda_T T(v_i) \} \quad (1)$$

λ_I，λ_C，λ_T は各スコアのバランスを取るための重み係数である．L，I，C，T はそれぞれ言語スコア，単語重要度スコア，信頼度スコア，単語間遷移スコアであり，以下で説明する．

- 言語スコア

  言語スコア $L(v_i)$ は，要約文内の単語連鎖の適正度を示すスコアである．ここでは，統計的言語モデルである単語 trigram を用いる．

  $$L(v_i) = \log P(v_i | v_{i-2} v_{i-1}) \quad (2)$$

- 単語重要度スコア

  単語重要度スコア $I\ (v_i)$ は，文中における単語の重要度を示すスコアである．ここでは，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．話題語スコアには以下の式で表される単語の出現頻度に基づく情報量を適用する．

  $$I(v_i) = f_i \log \frac{F_A}{F_i} \quad (3)$$

  v_i : 音声認識結果に含まれる名詞
  f_i : 要約対象である音声中の名詞 v_i の出現頻度
  F_i : 大規模コーパス中での名詞 v_i の出現頻度
  F_A : 大規模コーパス中での総名詞数 $(= \sum_i F_i)$

- 信頼度スコア

  信頼度スコア $C\ (v_i)$ 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に対しペナルティを与える．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の対数値を，信頼度スコアとして用いる．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．単語仮説 v_i の信頼度は，単語グラフにおけるノード番号 k, l を用いて，次式のように forward 確率と backward 確率に基づく事後確率の対数値として求められる．

  $$C(v_{k,l}) = \log \frac{\alpha_k P_a(v_{k,l}) P_l(v_{k,l}) \beta_l}{\mathcal{G}} \quad (4)$$

k,l : 単語グラフにおけるノード番号 ($k < l$)

$v_{k,l}$: ノード k,l 間のリンクに対応する単語

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

$P_a(v_{k,l})$: 単語 $v_{k,l}$ の音響尤度

$P_l(v_{k,l})$: 単語 $v_{k,l}$ の言語尤度

$\mathcal{G}$: 始端 S から終端 T 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．

- 単語間遷移スコア [1]

  単語間遷移スコア $T(v_i)$ は，要約文内の単語連鎖が原文において係り受け関係にあるか否かを示す単語間遷移確率の対数値 $T(v_{i-1}, v_i)$ で定義され，係り受け関係にない単語連鎖にペナルティを与えるものである．

このように定義された式 (1) の要約スコアが最大となる単語の組み合わせを複数の文に分けられた認識結果から抽出する手法として， 2 段 DP を用いる．第一段階として，可能なすべての要約率で各文を要約し，第二段階として，全体の要約率が目的の要約率となるように各文の要約文を組み合わせ，その中から全体要約スコアが最大となる組み合わせを動的計画法により決定する．ただし，要約スコアを計算する前に，フィラーは取り除かれる．

重要文抽出手法

重要文抽出は，$W = w_1, w_2, \ldots, w_N$ を 1 文が N 個の単語からなる認識結果の単語列とすると，以下の式で表される要約スコア $S(W)$ を基に要約を行う．

$$S(W) = \frac{1}{N} \sum_{i=1}^{N} \{L(w_i) + \lambda_I I(w_i) + \lambda_C C(w_i)\} \quad (5)$$

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．L，I，C はそれぞれ言語スコア，重要度スコア，信頼度スコアであり，単語抽出で用いたスコアと同様である．ただし，重要度スコアは，名詞，動詞，形容詞，未知語の内容語に付与される．このように定義された (5) 式の要約スコアを各文ごとに計算し，フィラーを除いた後，目的の要約率まで要約スコアが高い文から選択していくことにより要約を行う．

2.1.2 要約方式

単語単位要約方式

単語抽出手法を用いて単語単位で要約を行う．重要部分を細かな単位で抽出することが可能であるが，短い単位の音声を多数接続することとなり，要約音声としては不連続性が生じやすく聞きづらくなるという可能性がある．

文単位要約方式

重要文抽出手法を用いて文単位で要約を行う．音声としては連続性が保存されやすく，聞きやすいものになるが，単位が大きいため不要な部分を含みやすいという問題がある．なお，フィラーは最終的に除かれるため,「文単位要約」であっても，要約音声を作成する際には，文単位内のフィラー位置で音声どうしの接続を行っている．

フィラー間単位要約方式

認識結果を，文境界とフィラー位置，または，フィラー位置とフィラー位置で区切った単位を「フィラー間単位」とし，これを 1 つの文とみなし，重要文抽出手法を用いて要約を行う．この単位は，単語単位と文単位の中間的な削除・選択単位として導入したもので,「細やかな重要部分の選択」「連続性の保持による音声の聞きやすさ」という両者の利点を併せた効果が期待される．

2.2 要約音声作成部

音声自動要約によって得られた要約テキストは，各単位の削除・選択やフィラーの除去によって作成されているため，要約テキストに相当する音声を作成する際には，元の音声からの切り出しと接続を行う．その際には，認識結果の時間ラベルを利用して，要約テキストと切り出す音声区間の対応を取っている．以下に，文の内部での音声接続と，文の内部で音声接続して作成した文どうしの接続について説明する．また，この要約音声作成の流れを図 2 に示す．なお，文の内部で音声接続してできた文を「要約音声文」，その音声区間に対応する要約テキストを「要約テキスト文」と呼ぶことにする．

文内部での音声接続

切り出した音声どうしをそのまま接続する場合，互いの音声のパワーの差から雑音が生じる可能性がある．これを防ぐため音声端のパワーを約 20 ミリ秒に渡って徐々に減衰させ，接続する．その際，接続部付近の音声の速度が速く，聞き取りにくくなるため，さらにポーズ区間を音声間に挿入する．挿入するポーズ長は手動で決定した．これを表 1 に示す．

要約音声文間の音声接続

文内部での音声接続と同様に，音声端のパワーを徐々に減衰させ，ポーズを挿入して接続する．ポー

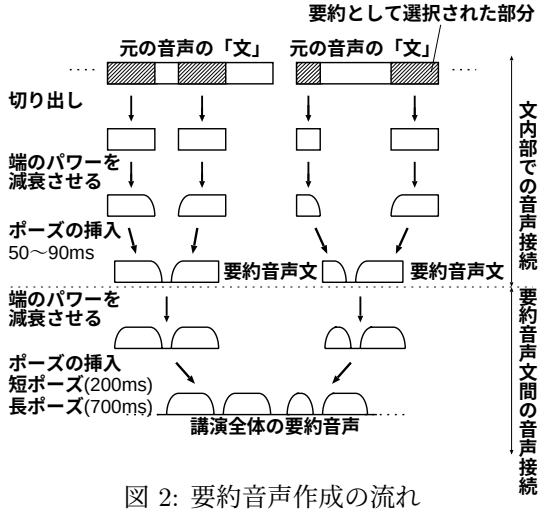

図 2: 要約音声作成の流れ

表 1: 文内部で挿入されるポーズ長 (ミリ秒)

	M74	M35	M31
単語単位要約	90	90	90
フィラー間単位要約	50	50	70
文単位要約	50	50	70

ズの長さは短ポーズ (200 ミリ秒)，長ポーズ (700 ミリ秒) の 2 種類とした．具体的には，要約テキスト文の末尾の表記が「した・です・ます・せん・ある・ない・ね」となる要約音声文に後続の音声文を接続する際には長ポーズを，それ以外は短ポーズを挿入する．これは，完結した文の後に長い間を取った方が聞きやすいと考えられるためである．ただし，単語単位要約方式を用いる場合については，要約テキスト文に「体言止め」が頻出するため，このような要約音声文に後続の音声文を接続するときには長ポーズを挿入した．

このようにして，文内部で音声接続を行って要約音声文を作成し，生成した要約音声文をさらに接続して，講演全体の要約音声を作成する．

3 評価実験

要約音声にふさわしい削除・選択単位を検討するため，講演音声について，2.1.2 節で説明した 3 つの要約方式を用いて要約テキストと要約音声を作成し，要約テキストに対する要約正解精度による評価と要約音声に対する聴取実験による主観評価を行った．

3.1 実験条件

講演音声に関するデータ

CSJ 中の男性話者 3 名による講演 A01M0074 (M74), A01M0035(M35), A01M0031(M31) を用いた．これらの講演時間，認識結果として得られる単語数，文数および単語正解精度を表 2 に示す．

表 2: 講演に関するデータ

	M74	M35	M31
講演時間 (分)	12	28	27
単語数	2,311	5,180	4,850
文数	86	212	208
単語正解精度 (%)	71.8	55.4	69.4

音声認識システムの条件

- 特徴抽出

  音声データを 16kHz，16bit でディジタル化し，フレーム長 25ms，フレーム周期 10ms で Δ 対数パワーと 12 次元の MFCC および Δ MFCC (計 25 次元) を抽出する．さらに文毎にケプストラム平均正規化を行う．

- 音響モデル

  CSJ 中の評価実験で用いた講演以外の男声話者による 94 時間 (455 講演) の音声データを用い，混合数 16 の不特定話者音素文脈依存 HMM(2000 状態) を作成した．

- 言語モデル

  単語 bigram，trigram を用いる．音響モデルを作成した際に用いた音声データの書き起こし文を，形態素解析システム JTAG により形態素に分割し，約 1.5M 形態素を用いて語彙 26k の学習を行った．ただし,「単語＋読み＋品詞」を形態素の単位とした．

- デコーダ

  単語グラフを中間表現とする 2 パスデコーダ，HMM-Toolbox を用いる [5]．第 1 パスでは HMM と bigram を用いてフレーム同期のビームサーチを行い，単語グラフを生成する．このとき，単語間の音素文脈依存も考慮する．

音声自動要約システムの条件

- 単語抽出手法

  − 言語スコア

  スコアの計算に使用する言語モデルは，要約文における単語連鎖をモデル化するものであるが，言語モデルを学習できる要約文の大規模なコーパスは存在していない．そのため，音声認識システムで用いた言語モデルの学習で使用した CSJ の講演の書き起こしを論説調の表現に変換したものと，60 講演の予稿集から作成した単語 trigram を用いてスコアを計算する．

表 3: 各講演に対する各方式の要約正解精度 (%)

	M74	M35	M31
単語単位	49.6	37.6	50.0
フィラー間単位	44.7	37.5	46.9
文単位	45.5	37.6	53.4

－ 単語重要度スコア

講演の書き起こし (約 1.5M 形態素)，60 講演の予稿集，WWW 上の講演録 (2.1M 形態素)，NHK のニュース原稿 (22M 形態素)，毎日新聞 (87M 形態素) および「音声情報処理」(51k 形態素) のテキストコーパスから，出現した全約 120k 種類の単語の各々の出現頻度を求め，スコアの計算に用いた．

－ 単語間遷移スコア

毎日新聞約 4 万文の構文解析済みの京大テキストコーパスを用いて，構文木制約付きの Inside-Outside アルゴリズムにより，係り受けパラメータの推定を行ったものから求めた．

• 重要文抽出手法

－ 言語スコア

単語抽出手法の言語スコアで用いた CSJ の講演書き起こしテキスト (約 1.5M 形態素) から作成した単語 trigram を用いた．ただし，文単位要約方式で用いる trigram は文単位で区切られているテキストで学習し，フィラー間単位要約方式で用いる trigram は文単位をさらにフィラーで区切ったテキストで学習を行った．

－ 重要度スコア

重要度スコアの計算には，上記の単語抽出手法の重要度スコアと同様のコーパスを用いた．

3.2 要約正解精度による評価

各講演音声を 2.1.2 節で説明した 3 つの単位の要約方式を用いて要約率 50% で要約し，それらの要約テキストについて，正解要約文単語ネットワーク [6] に基づく要約正解精度を求めた．これを表 3 に示す．

正解要約文単語ネットワークは，被験者が作成した正解要約文の単語連鎖をネットワークとしてまとめることにより，すべての可能性のある正解要約文の単語連鎖を近似的に網羅している．生成した要約に一番近い単語連鎖を正解要約文単語ネットワークから抽出し，これを正解として，式 (6) で表される要約正解精度により要約を評価する．正解要約文単語ネットワークは 9 人分の正解要約文から作成した．

$$Sum\_acc = \frac{Len - Sub - Ins - Del}{Len} \times 100(\%) \quad (6)$$

Sum_acc： 要約正解精度
Sub： 置換誤り
Ins： 挿入誤り
Del： 削除誤り
Len： 正解単語列の単語数

3.3 聴取実験による主観評価

3.3.1 要約音声の提示方法と評価方法

3.2 節で評価した各講演に対する 3 つの要約方式の要約テキストと，認識結果の時間ラベルから，2.2 節で説明した方法で要約音声を作成し，被験者に提示した．被験者には事前に講演の書き起こしを読んで内容を把握してもらってから，作成した 3 講演分の要約音声を提示した．ただし，講演 M74 と M35 に関しては講演の予稿も用意できたため，書き起こしとあわせて事前に読んでもらった．各講演は，内容として意味のある 3 つの段落に分けられており，段落の順序は入れ替えず，それぞれの段落について 1 種類の要約方式で作成した要約音声が被験者に提示される．1 講演分の要約音声で 3 種類の異なる要約方式で作成された要約音声が提示されることとし，その要約方式の並びは講演ごとにランダムに入れ替えた．

評価基準は「聞きやすさ」「要約音声としてのふさわしさ」の 2 つで,「1: 非常に悪い，2: 悪い，3: 普通，4: 良い，5: 非常に良い」の 5 段階で評価を行った．被験者は 11 名である．

3.3.2 主観評価結果

「聞きやすさ」「要約音声としてのふさわしさ」の評価結果を図 3，4 に示す．聞きやすさの評価結果は単語単位要約方式が全ての講演において評価が一番低く，講演 M74 と M35 については文単位，フィラー間単位要約方式がほぼ同等で，M31 については文単位要約方式の方が高かった．平均すると，文単位要約方式，フィラー間単位要約方式，単語単位要約方式の順に評価が高かった．要約音声としてのふさわしさの評価結果は，聞きやすさの評価結果とほぼ同様な結果であった．

単語単位要約方式については，表 3 のように，要約正解精度が他の要約方式より高いにもかかわらず「要約音声としてのふさわしさ」の評価は最も低くなった．これは，短い音声区間を多数接続したことによる「聞きやすさ」の劣化が原因であると考えられる．

文単位要約方式の評価は,「聞きやすさ」「要約音声としてのふさわしさ」ともに最も高かった．一方，フィラー間単位要約方式においては，意図した有効性を確認することができなかった．これは，要約率が 50% という条件では，文単位要約方式において不要部分が含まれることが少なかったこと，フィラーの頻度が多い話者では，フィラー間単位要約方式において，単語単位要約方式と同様の音声接続の頻発による「聞きやすさ」の劣化が生じるこ

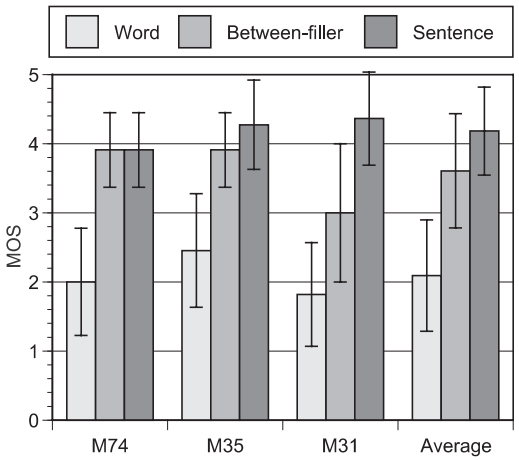

図 3: 「聞きやすさ」の評価結果

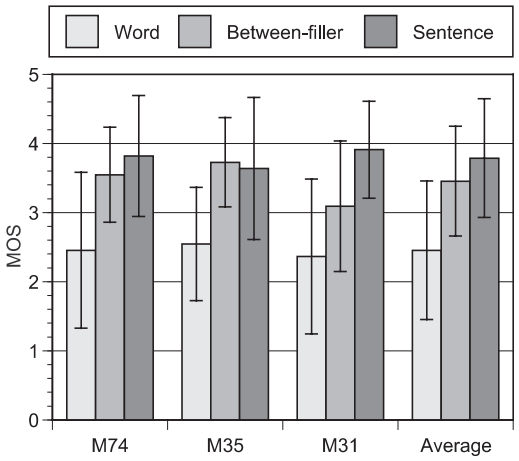

図 4: 「要約音声としてのふさわしさ」の評価結果

とが原因であると考えられる．図 3 の M31 でフィラー間単位要約方式での性能が，文単位要約方式に大きく劣っているのは，後者の原因によるものと考えられる．表 4 には，認識結果に基づく各講演におけるフィラーの割合を，表 5 には，認識結果に基づく各講演における各要約単位での単位数を示している．M31 は他の 2 講演と比べ，フィラーの割合が高く，それによりフィラー間単位としたときの単位数が大きく増えていることがわかる．一方，M74, M35 の結果をみると,「聞きやすさ」における文単位要約方式からの大きな性能劣化が見られない．要約率をより小さく設定した場合には，不要部分の影響によって文単位要約方式の「要約音声としてのふさわしさ」が劣化する恐れがあることから，このような講演に対しては，フィラー間単位要約方式の有効性が得られるものと期待される．

4 まとめ

本論文では，音声自動要約手法を用いて講演音声の自動要約を行い，出力された要約テキストに

表 4: 認識結果に基づく各講演におけるフィラーの割合

	M74	M35	M31
全単語数	2,311	5,180	4,850
フィラーの数	190	432	614
フィラーの割合 (%)	8.22	8.34	12.7

表 5: 認識結果に基づく各講演における単位数

	M74	M35	M31
単語単位	2,311	5,180	4,850
フィラー間単位	215	478	693
文単位	86	212	208

対応する音声を切り出し，接続することにより要約音声を作成し，これをユーザに提示する「速聞きシステム」を提案を行った．要約音声にふさわしい削除・選択を検討するため，単語単位・フィラー間単位・文単位の 3 の単位の要約方式により作成された要約音声について聴取実験を行ったところ，要約率 50% では，文単位要約方式が最も有効であることが確認された．

今後の課題としては，講演数の増加，様々な要約率での各要約方式の評価がある．要約率が小さい場合には，文単位要約方式の不要部分が影響して，フィラー間単位要約方式の有効性が大きくなると期待される．その他の課題としては，フィラーの多い講演に対応できるフィラー間単位要約方式の改良，ポーズ長の聞きやすさへの影響の検討，音声の切り出しと接続による韻律の不自然性の改善等がある．

参考文献

[1] 堀智織, 古井貞熙 : “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).

[2] 菊池智紀, 古井貞熙, 堀智織 : “重要文抽出と文圧縮による音声自動要約”, 電子情報通信学会技術研究報告, NLC2002-81 / SP2002-158, pp.61-66(2002-12).

[3] 木山次郎, 伊藤慶明, 岡隆一 : “Incremental Reference Interval-free 連続 DP を用いた任意話題音声の要約”, 電子情報通信学会技術研究報告, SP95-35, pp.81-88(1995-1).

[4] 小林聡, 吉川裕規, 中川聖一 : “表層情報と韻律情報を利用した講演音声の要約”, 情報処理学会研究報告, 2002-SLP-43, pp.41-46(2002).

[5] 堀貴明 : “大語彙連続音声認識の研究”, 平成 11 年度 博士論文 (1999).

[6] 堀智織, 古井貞熙 : “単語抽出による音声自動要約文生成方とその評価”, 電子情報学会論文誌 D-II, Vol.J85-D-II, No.2, pp.200-209 (2002-2).