DEWS2004 1-B-02

音声と映像の一貫性を考慮した要約動画の生成

伊藤　一成†　　酒井　康旭†　　斎藤　博昭†

† 慶應義塾大学 大学院理工学研究科　〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
E-mail: {k_ito,yasu,hxs}@nak.ics.keio.ac.jp

あらまし　本稿は自然言語処理を核とした新たな動画要約手法を提案する．動画内容はすべてメタデータを用いて表現できると仮定すると，音声と映像を分離して要約することが可能となる．すなわち，ユーザが指定する任意の要約率で音声テキストを要約した後に，対応する映像の重要区間を決定する．要約結果の提示の際には映像の重要区間を再生するのと同時に，日本語スピーチエンジンを利用して要約テキストを音声に変換することで，音声と映像の一貫性を考慮した要約生成が実現できる．ニュース報道番組の動画要約システムを試作し，提案手法の有効性を確認した．

キーワード　マルチメディア情報処理，動画要約，自然言語処理，メタデータ解析

Movie Summarization in Consideration of the Consistency of Voice and Video Image

Kazunari ITO†, Yasuaki SAKAI†, and Hiroaki SAITO†

† Department of Science and Technology, Keio University
Hiyoshi 3–14–1, Kouhoku–ku, Yokohama, 223–8522 Japan
E-mail: {k_ito,yasu,hxs}@nak.ics.keio.ac.jp

Abstract　This paper proposes a novel movie summarization method based on meta data analysis and text processing. Since all the contents of a movie can be described in a meta data format, it becomes possible to summarize the movie in two layers: voice and video image. Namely, the speech contents are firstly abridged at an arbitrary condense rate using natural language techniques. Then important video sections are determined corresponding to the selected speech parts. When the summarized result is presented, the video sections are reproduced along with the synthesized speech of the abridged text. This summarization method assures the consistency of sound and video. We have implemented a news summarization system and confirmed the effectiveness of our approach.

Key words　multimedia information processing, movie summarization, natural language processing, meta data analysis

1. はじめに

記憶装置の大容量化やネットワーク技術の発展などによって，音声や映像といったマルチメディアデータを容易に取り扱うことができるようになった．また，ディジタル放送やホームサーバの標準化並びに実用化の動向を見ると，番組コンテンツとして大量の映像データが一般視聴者に提供されることが近い将来に現実となることが考えられる．しかしその一方で，一般視聴者には，膨大な量の映像データの中から得たい情報を選択することを強いる結果となり，限られた時間の中で視聴者が必要とする情報だけを選択的に視聴することが困難な状況になることが予想される．これを解決するために，これらコンテンツを何らかの形で整理し，ユーザの要求を適切な

形で提供できる，より高度なハンドリング手法が望まれており，そのアプローチの一つとして動画要約技術が挙げられる．これまでの動画要約の研究は，スポーツのような動画を扱う場合では映像の要約だけを行えば良かったが，ニュース報道番組の動画などは，音声と映像の両方の一貫性を考慮する必要があり，これを同時に満たすのは困難であった. そこで，本稿ではニュース報道番組の動画に焦点を当て，動画内容を予め記述したメタデータを用いて，そこに自然言語処理の手法を適用する．これにより，動画中に含まれる映像と音声を分離してそれぞれを要約し，再合成することが可能となり，両方の意味的一貫性を考慮した新たな要約手法の実現が期待できる．

2. 動画内容の記述

動画などのマルチメディアデータは，テキストに比べて内容に基づく処理が極めて困難である．よって近年ではこの問題を解決するために，メタデータと呼ばれるマルチメディアデータの外部情報を利用した試みがある [1] [2]. 本稿では，メタデータの記述形式として，我々が提唱している MAML [3] と，橋田が提案する GDA [4] を利用し，ニュース動画の内容を記述する．図 1 に本手法で用いたメタデータの例を示し，以降で，それらメタデータの特徴と形式について述べる．

2.1 MAML(Maltimedia Annotation Markup Language)

MAML は人間が理解及び記述しやすい表現構造を念頭においた，XML 形式の汎用アノテーション記述言語である．メディアの種類やフォーマットに非依存な統一記述仕様であり，タグによるデータの構造化は最小限で，自然文章中心の構造となっている．MAML ではデータ形式やファイルフォーマットによるファイルの区分は行わず，すべてのファイルは音声情報（audio），映像情報（visual），内容情報（contents）を有しうるメディアとみなす．ここでいう音声情報とは人の聴覚から得られる情報，映像情報とは人の視覚から得られる情報，内容情報とはコンテンツやテキストデータに対して人の知識，主観，推論によって導かれる情報をいう．

2.2 GDA(Global Document Annotation)

GDA は橋田が提案する，多言語間に共通の統語・意味などに関する XML タグの標準を作って普及させようというプロジェクトである．GDA では，文法機能（主語，目的語，間接目的語），主題役割（動作主，非動作主，受益者など），修辞関係（理由，結果など）や照応関係を表すことができ，既に検索，要約，翻訳，対話処理，質問応答システムをはじめとした自然言語処理の分野で GDA を活用した研究報告がなされている [5] [6] [7]. 人手によって GDA タグを付与する場合，かなりの労力や専門知識を要するが，タグ付けが比較的容易な，文章の形態素情報及び構文情報などの表層的な情報であれば機械的処理によりタギング可能であり，それを MAML に埋め込むことができる．図 1 は，日本語の係り受け解析器 CaboCha [8] の出力結果から得られる形態素情報と

```
<?xml version="1.0" encoding="UTF-8"?>
<maml>
 <media type="movie"
  maml-location="http://www.xxx.jp/abc.maml"
  media-location="http://www.xxx.jp/abc.mpg"
  duration="01:42.562">
  <element id="1" begin="6.7" end="9.2">
   <audio>
    <utterance>
     <su>
      <adp>
       <adp>
        <adp bfm="この" prn="コノ">この</adp>
       </adp>
       <ad>
        <np bfm="地震" prn="ジシン">地震</np>
        <ad bfm="の" prn="ノ" sem="連体化">の</ad>
       </ad>
       <ad>
        <np bfm="影響" prn="エイキョウ">影響</np>
        <ad bfm="で" prn="デ" sem="格助詞">で</ad>
        <x>，</x>
       </ad>
      </adp>
      <adp>
       <adp>
        <np bfm="都内" prn="トナイ">都内</np>
        <ad bfm="の" prn="ノ" sem="連体化">の</ad>
       </adp>
       <ad>
        <np bfm="家屋" prn="カオク">家屋</np>
        <ad bfm="が" prn="ガ" sem="格助詞">が</ad>
       </ad>
      </adp>
      <n>
       <np bfm="倒壊" prn="トウカイ">倒壊</np>
       <x>．</x>
      </n>
     </su>
    </utterance>
   </audio>
  </element>
      :
 </media>
      :
</maml>
```

図 1　MAML と GDA の記述例

構文情報をから GDA タグを付与し，これを MAML に埋め込んだ例である．図 1 では，動画中に含まれる発話を転記した文章に対し GDA 文章を作成した．<su> は文を示す．<n>，<np> は名詞および名詞を主辞とする語句．<n> は他の語句の係り受け対象となることができ，<np> はならない．他についても同様である．<v>，<vp> は動詞，助動詞，終助詞，またはそれらを主辞とする語句．<ad>，<adp> は，終助詞以外の助詞，副詞，連体詞，接続詞，およびこれらの投射である．以上のようなメタデータ形式によりニュース動画の内容を記述し，実際に要約動画を生成する際には，このメタデータを解析することで音声と映像の重要部分をそれぞれ決定する．

3. テキスト要約処理

本手法では，要約結果の動画に含まれる音声情報についての一貫性も考慮する．そのため，前節で説明したメタデータ中に含まれるニュースキャスターの発話転記テキストについて要約処理を施す．テキスト要約処理の主なアプローチには以下のようなものが挙げられる．

- 重要文抽出による要約
- 抽象化，言い換えによる要約
- 文中の不要箇所削除による要約

重要文抽出による要約とは，テキスト中の文，あるいは形式段落を 1 つの単位とし，それらに何らかの情報を基に重要度を付与し，その重要度で順序付け，重要な文（形式段落）を選択し，それらを寄せ集めることをいう．抽象化，言い換えによる要約とは，テキスト中の表現を他の簡潔な表現で言い換えたり合成することで原文の内容を表現し直し要約を生成すること．文中の不要箇所削除による要約とは，一文ごとに文中の重要でない句や文字列を削り，テキストを短く表現し直して要約を生成することを意味する．このうち本稿では，一段階目の要約処理として重要文抽出による要約を行い，二段階目の要約処理として文内の不要箇所削除による要約を行う．以下に本稿で用いるテキスト要約手法について説明する．

3.1 重要文抽出による要約

重要文抽出の手法としては吉見らの手法 [9] を採用している．以下に吉見らの重要文抽出手法について説明する．

吉見らの手法は，文の重要度に関して次の 2 つの仮定に基づいている．

（1） 表題はテキスト中で最も重要な文である．
（2） 重要な文とのつながりが強ければ強いほど，その文は重要である．

最初に表題文 S_1 中にに含まれる重要語を抽出して重みを付与し，それぞれの語の重要度を求める．ここで重要語とは形態素解析の結果から得られる品詞が，名詞，人称代名詞，動詞，形容詞，副詞のいずれかである辞書見出し語を指す．次に，表題文 S_1 の重要度を次式で求める．

$$S_i\text{の重要度} = \frac{S_1\text{中の重要語の重み和}}{S_1\text{の重要語の数}}$$

この表題文の重要度を元に，後に続く文との関連度を算出してそれぞれの文の重要度を決定する．文 S_j の先行文 S_i へのつながりの強さ（関連度）を求める式を以下に示す．

$$S_i\text{と } S_j\text{の関連度} = \frac{S_j\text{中の重要語のうち } S_i\text{の題述中の重要語につながるものの重みの和}}{S_i\text{の題述中の重要語の数}}$$

ここで，文 S_j の主題は，S_j 中の重要語のうち S_j の関連文中の重要語につながるものから構成され，文 S_j の題述は，つながらない重要語から構成される．ただし，関連文を持たない冒頭文 S_1 では，それに含まれる重要語すべてが題述を構成する．重要文選択の手順を図 2 にまとめる．本稿で扱うメタデータ内には，文章に対する表題が存在しないため，映像中に含まれるクローズドキャプション文字列を表題に代用した．

ステップ1 GDA文書を入力とする。

ステップ2 表題への重み付け処理を行う。

ステップ3 冒頭文S_1の重要度を次式で求める。

$$S_1\text{の重要度} = \frac{S_1\text{中の重要語の重み和}}{S_1\text{の重要語の数}}$$

ステップ4 各文S_j(j=2,3,・・・,n)について、S_jから五文前までの先行文 S_iの範囲(j-5 ≦i<j)で重要度を求める。

ステップ5 あらかじめ定められた数だけ文を重要度の順に選択し、それらをテキストでの出現順に出力する。

図 2 重要文選択手順

3.2 文内の不要箇所削除による要約

重要文抽出による要約処理により，テキスト全体からユーザが指定する要約率を指定して重要な文の集合を得ることができる．その結果から得られた重要文の各文をさらに要約することで要約率を高めるために，我々が既に提案した手法 [7] を基本にして，各文内で重要度が低いと思われる節を削除する．以下に基本的な概念を示す．この手法では，係り受け情報に加えて，照応・代用・省略

といった詳しい情報があらかじめ付与された GDA 文書を処理対象としている．基本的には GDA タグの文法機能（主語，目的語，間接目的語），主題役割（動作主，非動作主，受益者など），修辞関係（理由，結果など）の情報を利用して得られる文のテキスト構造から各節に非重要度のスコアを付け，スコアの小さい語のみを抽出することで行なう．まず，文の必須語（主辞，主語，目的語）を抽出する．次に，図 3 のように，各節に対して GDA タグの文法の種類と係り受けに応じて非重要度を付与し，ある閾値以上の語を抽出する．この際，閾値は要約率と各文の非重要度の最大値から決定する．

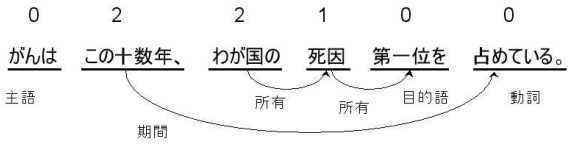

図 3　非重要度付与の例

本稿では，係り受け構造から非重要度を決定するという基本概念のみを採用し，非重要度付与と重要箇所抽出の方法は独自に決定する．各節への非重要度の決定には，GDA 文書構造から得られる係り受け情報を用いる．GDA 文書は XML 形式の記述であるため，係り受け情報を得るためには，その木構造を解析することが必要である．この場合まず自動生成された GDA 文書の su 節を抽出して，同じ親節を持つ同じ階層の葉節の集合を一つの処理対象とする．図 4 に，図 1 内の GDA 文書に対して重要度を付与する場合の処理対象の単位と付与した非重要度を例示する．

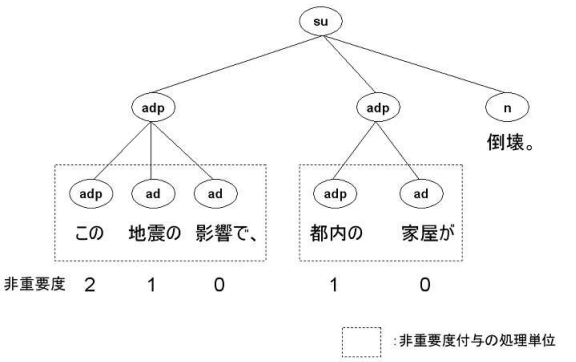

図 4　本手法における非重要度付与の例

非重要度を付与する際には，図 4 に示した各処理単位内において，最後の節の非重要度を 0 とし，次に，その他の節の非重要度については非重要度が 0 の節に係るまでの距離をその節の非重要度とする．以上の処理により決定した各節の非重要度が 2 以上の場合，その節を削除の候補とし，経験的に定めた独自ルールにより削除の可否を決定する．以上のテキスト要約処理により，ユーザが指定した要約率に応じて適切な長さの文章を返すことができる．この時，一秒間に発話可能な文字数を決定しておき，元となる動画の総時間と要約率から求められる要約結果の秒数から，重要文抽出の結果に含める文字数を決定するようにする．発話文字数を数えるには，GDA 文書内において prn 属性の文字列数を用いる．図 5 にテキスト要約結果の例を示す．

原文:

自民党総裁選は8日告示され、20日の投開票に向けて選挙戦がスタートする。再選を目指す小泉純一郎首相に亀井静香前政調会長、藤井孝男元運輸相、高村正彦元外相の3氏が挑む構図だが、毎日新聞が党都道府県連幹事長らに地方票の情勢を聞いたところ、25都道府県が「小泉首相が大勝する」とみていることが分かった。また11県が「小泉優勢」とした上で、亀井、藤井両氏のどちらかが2番手に迫っているとの見方を示した。青森、岐阜、広島など7県では亀井、藤井、高村各氏のいずれかが優勢または首相に拮抗(きっこう)、と回答。京都など4府県は「分からない」と答えた。首相の対立候補が正式に決まっていない8月末時点で毎日新聞が実施した調査と比較すると、前回「小泉優勢」と答えていた27道県のうち、25道県が「小泉大勝」または「優勢」との見方を示した。一方、「対立候補が優勢」とみていた7府県のうち、高知が「小泉大勝」、千葉、徳島が「小泉優勢」に転じた。「分からない」「五分五分」などと答えていた13都府県中、8都府県が「小泉優勢」に転じた。総裁選は、国会議員票357票と地方票300票の計657票で争われれる。地方票は各都道府県に対し、選挙人数に応じて4票から10票配分する。各県の持ち票は「ドント方式」によって各候補の得票率に応じ比例配分する。県連幹部が「小泉大勝」「小泉優勢」とみている36都道府県の 持ち票合計は224票。県連幹部の見方通りの情勢が続けば、このうちの多くが小泉首相に流れ、国会議員票で優位な首相が地方票でも大幅に上積みすることになる。今回は「小泉大勝」との見方を示した。逆に、小泉首相が勝利した41都道府県のうち、新潟が「亀井優勢」、宮崎が「亀井、藤井両氏が拮抗」と回答した。

重要文(要約率 50%) 抽出された文章:

再選を目指す小泉純一郎首相に亀井静香前政調会長、藤井孝男元運輸相、高村正彦元外相の3氏が挑む構図だが、毎日新聞が党都道府県連幹事長らに地方票の情勢を聞いたところ、25都道府県が「小泉首相が大勝する」とみていることが分かった。また11県が「小泉優勢」とした上で、亀井、藤井両氏のどちらかが2番手に迫っているとの見方を示した。青森、岐阜、広島など7県では亀井、藤井、高村各氏のいずれかが優勢または首相に拮抗(きっこう)、と回答。京都など4府県は「分からない」と答えた。首相の対立候補が正式に決まっていない8月末時点で毎日新聞が実施した調査と比較すると、前回「小泉優勢」と答えていた27道県のうち、25道県が「小泉大勝」または「優勢」との見方を示した。「分からない」「五分五分」などと答えていた13都府県中、8都府県が「小泉優勢」に転じた。

文内要約された文章:

再選を目指す小泉純一郎首相に3氏が 挑む構図だが、毎日新聞が情勢を聞いたところ、25都道府県が「小泉首相が大勝する」とみていることが分かった。また11県が「小泉優勢」とした亀井、藤井両氏のどちらかが2番手に迫っているとの見方を示した。広島など7県では亀井、藤井、高村各氏のいずれかが 優勢または 首相に 拮抗(きっこう)、と回答。4府県は「分からない」と答えた。対立候補が 正式に決まっていない8月末毎日新聞が 実施した調査と比較すると、「小泉優勢」と答えていた27道県のうち、25道県が「小泉大勝」または「優勢」との見方を示した。「分からない」「五分五分」などと答えていた13都府県中、8都府県が「小泉優勢」に転じた。

図 5　テキスト要約結果の例

4.　映像重要区間の決定

4.1　映像重要度の付与

ニュースでは，キャスターの発話内容は同時刻の映像に何らかの関係があると考えられるため，重要文抽出の結果から得られる各文の重要度を用いて，時間的に重なりのある映像区間に対し，各ショットごとに映像重要度を付与する．図 6 に，映像重要度の付与方法を示す．

ここでは時間的に重なりがあるショットに対して，重

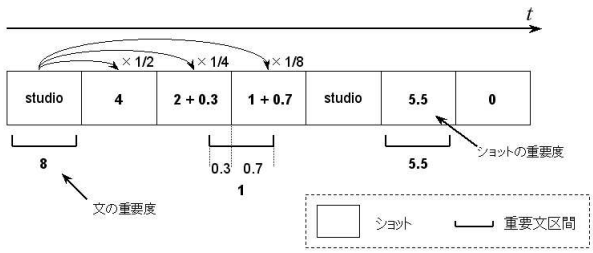

図 6　映像重要度の付与方法

なる時間の割合に応じて文の重要度を割り当てる．この際，スタジオのショットは候補から除外する．もし重要文区間がスタジオのショットと重なった場合は，その重要度を後方のショットに分配する．この時，重用度を分配するのはスタジオ以降の 3 つのショットまでとし，図 6 のとおり，それぞれ ×1/2，×1/4，×1/8 とする．ここで重要度を後方に分配する理由は，スタジオのショット中でなされるキャスターの発話内容を補完する内容が，それに続くショット中にあることが多いという経験則に基づいている．以上のルールで各ショットの重要度を決定しておき，ユーザが指定する要約率に応じて映像の要約候補を決定する．

4.2　要約映像区間の決定

映像重要区間の決定に際しては，図 7，図 8 に示す二つの方法を試みる．

図 7　要約映像区間決定方法 1

図 8　要約映像区間決定方法 2

方法 1 では，各ショットに割り当てられた重要度を用いずに，トピック中のスタジオを除く全ショットから，時間的に等しく映像重要区間を抽出する．方法 2 では，元となる動画の総時間とユーザによって指定された要約率から算出された要約結果時間を，ショットの重要度に応じて割り当てることで最終的な重要映像区間を決定している．これら二つの方法を試みる理由として，要約結果時間の長さによって，ユーザに結果を提示した際の，内容の理解度に差が出ると考えられるためである．例えば，要約結果時間が非常に短い場合に，方法 1 の様に全てのショットを等間隔で抽出すると，一つのショットの再生時間が非常に短くなり，映像の内容を理解するのに不十分である考えられるからである．また逆に，要約結果時間が長めの場合には，方法 2 によりショットの選択をすると，情報が激しく欠落する恐れがある．よって評価実験を行う際には，これら二つの方法で要約映像を生成し，比較することにする．また，ニュースに含まれる映像は，スポーツなどの映像と違いカメラワークが少ないため，ショットのどの部分を抽出しても重要度に変化がないことが考えられる．よって本手法では，割り当てられた時間に応じて各ショットの先頭から抽出し，それらを先頭から繋ぎ合わせたものを要約結果の映像とする．

5.　要約生成方法と評価実験

5.1　本手法の概要

ここまで説明した動画要約の手法を用いて要約動画の生成を行う．図 9 に本手法の概念図を示す．

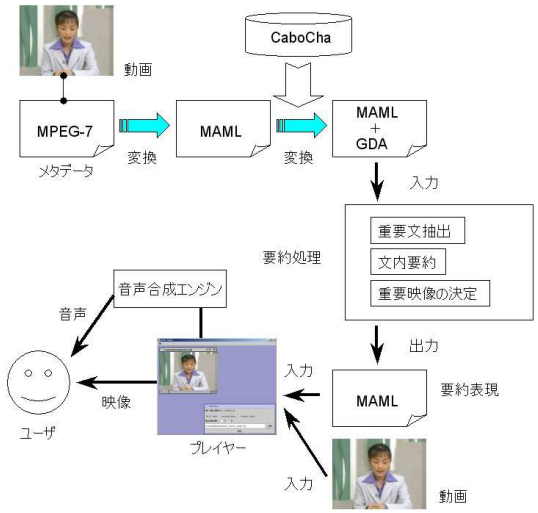

図 9　本手法の概念図

テストデータとしては国立情報学研究所が配布している評価用映像メディア DB [10] を用いた．このメディア

DB はニュース，ドラマ等を題材とした全長十数分の動画データ 10 本およびこれに付随する MPEG-7 形式のメタデータにより構成されている．これに含まれる約 15 分間のニュース動画 2 本を利用し，約 20 トピックの中からいくつかのトピックを選択し，これらに対して本手法を適用した．動画内容を記述した MAML ファイルは，動画データに付随している MPEG-7 形式のメタデータから変換し，さらにその中に含まれる発話転記を日本語係り受け解析器 CaboCha の出力結果によって GDA 文書化することで生成した．今回の実験ように利用できるメタデータが予め用意されていない場合，これを作成する際の人的コストについて懸念されるところであるが，今回必要とする情報は，シーンやショット毎の時刻情報，映像中のクローズキャプション文字列，ショットの簡易な映像情報，ニュースキャスターの発話転記内容のみであり，その大半は既存の画像・音声認識技術を用いることで自動化することが可能である．

次に，要約の際には，MAML を処理対象にして本手法による解析を実行し，要約された情報を新たな MAML ファイルとして出力した．要約された MAML ファイルには，各トピック内で音声として読み上げる文章と，再生する映像区間情報が含まれる．この時ユーザは任意の要約率を指定することが可能である．

5.2 アプリケーション

次に，本手法により生成した要約情報を含む MAML ファイルを読み込んで動画を再生するためのアプリケーション画面を図 10 に示す．

図 10　アプリケーションのスクリーンショット

このアプリケーションでは，画面右の内部フレームの参照ボタンから要約表現を記述した MAML ファイルを指定し，再生ボタンを押すと，読み込んだ MAML ファイルに基づいて元の動画から映像の重要区間だけを間引いて順再生する．この時，元の動画に含まれる音声は消音にして使用せず，代わりに日本語のスピーチエンジンを利用し，要約されたテキストから音声を合成し，映像と共に再生する．今回の実検ではマイクロソフト社が提供している日本語スピーチエンジンを用い，話速度を最大に設定して男性の声で発声させた．

5.3 実験方法

ここでは実験用動画に対して人間が予め作成した要約動画と，本手法で生成した要約動画を再生し，被験者に視聴・比較してもらうという主観評価形式を採用する．人間が生成した動画と比較するという評価方法は，いくつかの先行研究で採用されている．　実験用の動画としては，本章で述べた国立情報学研究所が配布している評価用映像メディア DB に含まれるニュース番組の動画 2 本中から選択した 8 トピックを用いる．このトピックの選択にあたっては，政治，科学，天気予報などできるだけ様々な内容を網羅するようにし，時間的にも短いものでは 40 秒程度のものから，長いものでは 2 分 30 秒程度のものまで様々な長さを選択した．これらの動画に対し人間の要約者 3 人が要約動画を作成した．まず，人間の要約者が作成した要約動画は以下の 2 種類である．

動画 1. 映像情報を重要視した要約
　要約者は，トピック内に含まれる映像情報を重視し，より情報を多く含んでいる（クローズドキャプション文字列が映っている等）動画区間を決定する．可能であれば音声情報も考慮する．

動画 2. 音声情報を重要視した要約
　要約者は，トピック内に含まれる音声情報をを重視しその繋ぎ目などを考慮しながら動画区間を決定する．可能であれば映像情報も考慮する．

ここで 2 種類の人手による要約を用意する理由としては，まず，人間が実際に動画を要約する際，重要と思われる区間を決定するにあたっては，要約者によって様々な決定方法が考えられるからである．そのため，“動画 1. 映像情報を重要視した要約” と “動画 2. 音声情報を重視した要約” を行った．これにより，既存の手法で正解とされてきた要約結果との比較を行うことができる．

以上に示した人手による 2 つの要約と，本手法により生成する要約を比較する．本手法により生成する要約は，全章で述べたとおり，2 通りの映像区間決定方法を試す．これを以下に示す．

動画 3. スタジオ以外の全ショットから等しく重要区間を抽出する要約

動画 4. ショットの重要度に応じて重要区間を抽出する要

約

また，要約結果の再生時に本手法で実装した専用プレイヤーを使用することで，音声情報と映像情報を完全に分離した要約と再生が可能となる．つまりこれは，全く新しい結果提示方法の提案であり，こうした結果提示の良否についても検討する必要がある．よって，更にここで人間の要約者に “5. 映像情報と音声情報のそれぞれを要約” も行ってもらい，これとその他の要約を比較する．

動画 5. 映像情報と音声情報のそれぞれを要約

要約者は，トピック内に含まれる映像情報と音声情報それぞれについて要約を行う．ここでいう音声情報とは発話転記テキストを要約することに相当し，要約結果を再生する際には，本手法で実装した専用プレイヤーを利用する．

以上の 1〜5 の要約方法で作成した 10 秒と 30 秒の動画をそれぞれ用意し，これらを 14 人の被験者に視聴してもらい，動画要約としての良否ついてそれぞれのアンケートにより 5 段階で良否を評価してもらった．5 段階の評価基準は表 1 のとおりである．

表 1　5 段階の評価基準

点	基準
5	十分適切である
4	まあまあ適切である
3	普通
2	あまり適切でない
1	全く適切でない

それぞれの方法で作成された要約結果が適切であるか評価するにあたり，被験者には，

- 重要な箇所が選択されているか
- 音声と映像の情報量は適切であるか
- 音声内容と映像内容の時間的同期がとれているか

という点に考慮して評価を行ってもらった．

5.4　結果と考察

評価実験の結果を表 2 に示す．

表 2　各要約動画の評価の平均値

	動画 1	動画 2	動画 3	動画 4	動画 5
10 秒要約の結果平均	1.75	2.43	2.41	2.30	3.81
30 秒要約の結果平均	2.72	3.60	3.06	3.02	4.07

まずこの結果から，従来手法における正解とされてきた動画（動画 1，動画 2）と，本手法により生成した動画（動画 3，動画 4）を比較すると，10 秒に要約した場合では動画 3，動画 4 の結果は動画 1 の結果を大きく上回っており，動画 2 と比べてもほぼ同等の結果であることがわかることから．また，30 秒に要約した場合では，動画 2 との比較ではやや劣るものの，動画 1 よりは良い結果を得ることができた．さらに，概して動画 1 よりも動画 2 の方が良い結果であることから，少なくともニュース番組の動画については，人間は映像情報よりも音声情報を重視するということが言える．つまり，本手法のように音声情報を重要視し，その処理結果に基づいて動画作成を行なう方法は，アプローチは正しいと考えられる．また，本手法の結果を下げている要因としては，言語解析時の形態素解析誤りや係り受け解析誤りなどから生じる重要な節の削除によって不可解な要約テキストが生成されることや，スピーチエンジンのイントネーションの不自然さなどが考えられるが，これらは今後各研究分野において精度が向上していくことで解決され，本手法を用いた場合でもさらに良い結果が得られると思われる．

次に，本手法で実装した専用プレイヤーを用いて要約結果を再生した場合，従来のように単に動画区間を再生するのではなく，動画中の映像情報と共に，スピーチエンジンを用いることで任意の音声を再生することが可能となる．動画 5 は，映像情報と音声情報を独立に再生することができるという本手法のメリットを用いた場合の最も良い要約結果として，人間の要約者が作成したものである．すなわちこれは，本手法を採用した場合の正解と考えることができ，本手法を用いて自動的に生成した動画 3，動画 4 を比較すると，結果から，全ての場合において動画 5 は動画 3，動画 4 の結果以上となっているのがわかる．また，動画 3，動画 4 はトピックによって結果の違いが激しいが，動画 5 はどのトピックでも安定した好結果を残しており，10 秒に要約した場合と 30 秒に要約した場合の結果の差が小さい．

最後に，従来手法における正解である動画 1，動画 2 と，本手法を採用した場合の正解である動画 5 を比較すると，動画 5 の結果の方が良好であることがわかる．これはすなわち，結果の提示方法にという点については，従来のように単に動画区間を再生する方法よりも，本手法の方が優れているということであり，視聴者が内容をこれまでよりも深く理解できる可能性を広げたことを示している．しかし，もともと短い動画を要約し，元動画の長さと要約結果の長さが大きく変わらない場合は，動画 5 の評価が低くなることがあった．この理由としては，それに含まれる要約情報自体には大きな差が出ず，元の音声をそのまま使用する従来手法の方が，音声合成エン

ジンによる音声よりも，より自然に聞こえるからであると考えられる．よって，現時点では，本手法は時間的制約が厳しい方が高い評価を得やすい傾向にあるが，今後，音声合成の分野で研究成果が上がっていくことでその他の場合でも良い結果を得られるようになると考えられる．

6. ま と め

本稿ではニュース動画を対象にして，動画内容を予め記述したメタデータを用い，自然言語処理の手法を適用した．これにより，動画中に含まれる映像と音声を分離してそれぞれを要約し，再合成することが可能となり，両方の意味的一貫性を考慮した自然な要約動画を生成することが可能となった．今回提案した手法では，予め与えられたアルゴリズムでのみ重要部分を選択したが，今後の方向性としては，ユーザの嗜好や要求に応じたパーソナライズ要約や，スポーツやドラマといったニュース以外の動画も扱えるようにしていくことで，より汎用的かつ実用的なシステムになっていくと考えられる．

文 献

[1] 橋本隆子, 白田由香利, 飯沢篤志, 北川博之: ターニングポイント解析に基づくダイジェスト作成方法, 情報処理学会論文誌（データベース）, Vol.43, No.SIG 5(TOD 14), pp.1–11, 2002.

[2] Alejandro Jaimes, Tomio Echigo, Masayoshi Teraguchi, Fumiko Satoh: LEARNING PERSONALIZED VIDEO HIGHLIGHTS FROM DETAILED MPEG-7 METADATA, IEEE International Conference on Image Processing, ICIP 2002, Rochester, NY, USA, September, pp.22–25, 2002.

[3] 伊藤一成, 斎藤博昭: メディアデータに対するアノテーション記述言語（MAML）の策定とその応用, 情報処理学会研究報告, FI70–4, pp.19–26, 2003.

[4] 橋田浩一: GDA 意味的修飾に基づく多用途の知的コンテンツ, 人工知能学会論文誌, Vol.13, No.4, pp.528–535, 1999.

[5] 鈴木潤, 橋田浩一: GDA タグを利用した回答抽出システムの提案, 言語処理学会第 7 回年次大会, 2001.

[6] 伊藤一成, 斎藤博昭: メタデータ解析に基づくメディア検索システム, 情報処理学会研究報告, DBS131–69, 2003.

[7] 野村雄司, 伊藤一成, 斎藤博昭: GDA タグを用いたテキスト自動要約, 言語処理学会第 9 回年次大会, 2003.

[8] http://cl.aist-nara.ac.jp/ katu-ku/software/cabocha/

[9] 吉見毅彦, 奥西稔幸, 山路孝浩, 福持陽士: 表題へのつながりに基づく文の重要度評価, 自然言語処理, Vol.6, No.1, pp.43–57, 1999.

[10] 馬場口登, 栄藤稔, 佐藤真一, 安達淳, 阿久津明人, 有木康雄, 越後富夫, 柴田正啓, 全炳東, 中村裕一, 美濃導彦, 松山隆司: 映像処理評価用映像データベースについて, 電子情報通信学会技術研究報告, PRMU2002-30, pp. 69–74, 2002.