音楽情報科学 34−11
(1999. 2. 18)

ビデオデータにおける音声とクローズドキャプションの同期手法の検討

山崎 博信　馬場口 登　北橋 忠宏

大阪大学 産業科学研究所
〒567-0047 大阪府茨木市美穂ヶ丘8番1号
{yamasaki,babaguti,kitahasi}@am.sanken.osaka-u.ac.jp

あらまし　マルチメディアの検索においては、言語(テキスト)から音あるいは画像を検索すること、あるいはその逆向きの検索、いわゆるクロスモーダル検索(cross-modal retrieval)が重要視されている。本報告では、ビデオデータに含まれる言語(テキスト)情報であるクローズドキャプション(CC)ストリームと音声ストリームの対応付けによる同期法について述べる。CCストリームは、音声ストリームの写し(transcript)であるが、実際には出現遅れや欠落などがあり、完全な対応付けは難しい。ここでは、音声ストリームから母音区間を抽出し、その音素系列を求め、一方、CCストリームからは単語辞書に基づき音素系列を構成し、DPマッチングにより対応付けを図る手法を提案し、実験的検討を加える。

キーワード インタモーダル協調、クロスモーダル検索、クローズドキャプション、母音検出、フォルマント周波数

Synchronizing Closed Caption Stream with Speech Stream in Video Data

Hironobu YAMASAKI, Noboru BABAGUCHI, Tadahiro KITAHASHI

Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka Ibaraki-City Osaka, 567-0047 JAPAN
{yamasaki,babaguti,kitahasi}@am.sanken.osaka-u.ac.jp

Abstract　In multimedia systems, cross-modal retrieval is required that searches video or audio segments by text queries, e.g. searching a scene where the Japanese national anthem was played. For this purpose, the correspondence between liguistic/textual and audio streams should be made. In this report, we present a method of temporally synchronizing a closed caption (CC) stream, which is a transcript of a speech stream, with a audio stream. CC streams in broadcasted live sports programs suffer from delay or lack of the elements. Our method attempts to extract phoneme sequences from both CC and audio streams, and to give temporal correspondences by means of DP matching. Finally, we show experimental results.

Key words　intermodal collaboration, crossmodal retrieval, closed caption, vowel detection, formant frequency

1 はじめに

映像 (ビデオ) を中心としたマルチメディアシステムの普及に伴い、膨大なビデオデータから本当に必要な情報を検索できるようなシステムの重要性が増してきている。そのためにコンテンツに基づいた検索を行うことが必要になる。

ビデオデータは、画像、音、言語 (テキスト) などのマルチモーダル情報ストリーム (MMIS:Multi Modal Infomation Stream) から構成され、画像は情景を、音は場面の雰囲気、言語はコンテンツの説明などを表し、各々が相互に時間的にも論理的にも密接に関連している。よってビデオデータのコンテンツを解析するためには、単一のストリームを利用するのではなく、各ストリームのインタモーダル協調 [1),2)] を利用することが不可欠である。

一方、ビデオの検索において、言語 (テキスト) から音あるいは画像を検索すること、あるいはその逆向きの検索、いわゆるクロスモーダル検索 (cross-modal retrieval)[†] を実現するには、ストリーム間のセグメント、オブジェクトの厳密な対応付けも必要となる。本稿で対象とする言語ストリームと音声ストリームの同期法はそのための要素技術となる。

ビデオデータに関連する言語・テキスト情報には、シナリオ (台本) やテロップ, クローズドキャプション (CC:Closed Caption) などがある。これらは画像, 音と違いテキストデータであるため非常に扱いやすく、顔によるインデキシング [3)]、イベント抽出 [1)]、アクション抽出 [4)]、およびシナリオに基づく映像編集 [5)] などに広く利用されている。また、音ストリームと言語・テキスト情報の対応付けに着目したものとして、シナリオを用いてセリフのモーラ数からビデオデータとの同期をする手法 [6) 7)] があるが、シナリオはビデオの付随的なデータであって、必ずしも容易に入手できるものではない。

そこで本稿では、ビデオデータに埋め込まれている言語情報である CC ストリームと音声ストリームとの同期法を提案する。CC ストリームとは音声ストリームの写し (transcript) として得られるテキスト (字幕) 情報のことを指す。

CC は、映画などでは予めセリフの発声時刻に同期するように入力されている。そのため、CC の表示時刻と実際の発声時刻に差はほとんど生じない。しかし、スポーツ中継などのような CC がリアルタイムで入力されるビデオデータでは、表示時刻と実際の発声時刻との間に大きな差が生じる。

従って、ビデオデータの CC ストリームに対して検索を行い、時間的に対応する部分をそのまま用いれば実際の音声よりも大きなずれを含んだ結果になる。そのため、CC ストリームとと他のストリームとの間のずれを修正する必要がある。ここでは CC ストリームが、音ストリーム内に含まれる音声ストリームをもとに生成されていることに注目し、それぞれの CC に対応する音声を見つけることによって同期させることを試みる。音声ストリームから母音区間を抽出し、その音素系列を求め、一方、CC ストリームからは単語辞書に基づき音素系列を構成し、DP マッチングにより対応付けを図る。

[†] 例えば、国歌を歌っているシーンを探せなど

2 ビデオデータにおける音, 言語ストリーム

2.1 音ストリーム

音ストリームはビデオのサウンドトラックから得られ、その音源は以下の 3 つに分類される。

(1) 会話音声 (speech) : 登場人物の声、ナレーションなど。

(2) 音楽 (music) : 声楽、器楽など。

(3) 音 (sound) : 歓声、拍手、笑い、ブーイングなど。

一般にこれら 3 つは重畳された音信号として得られ、各々を良好に分離するのは困難な処理である。

2.2 CC ストリーム

ビデオ特有のストリームとして言語ストリームがある。言語ストリームとは、クローズドキャプション (CC : Closed Caption) テキストからなるストリームを指す。これは音声ストリームの写し (transcript) として得られるテキスト (字幕) 情報であり、映像信号の中に埋め込まれたデータである。米国では聴覚障害者用に TV 番組やビデオソフトへの CC の組み込みが法律によって義務付けられている[‡]。CC は、TV 受像機や VTR の CC モードを ON にすると、画像フレーム中に出現し、出現レートは 2 文字／フレームである。

CC の文生成規則を以下に挙げる [2)]。ここに、“ ” 内の記号は CC 文に出現する記号、*NAME*、*WORD* は具体的な人名、文 (単語列) を表す。また、記号 [ ], { } は各々0 または 1 回の出現、0 または 1 回以上の出現を表す。

[‡] 現在のところ我が国においては導入されていないものの、21 世紀初頭には CC を組み込んだ放送が計画されている。

CC	:=	[Start][Speaker] {Speech | Sound | Music}
Start	:=	">>"|">>>"
Speaker	:=	*NAME* ":"
Speech	:=	Sentence
Sound	:=	"["Sound-Description"]"
Music	:=	"♪" {Lyrics}"♪"
Lyrics	:=	Sentence
Sound-Description	:=	{*WORD*}
Sentence	:=	{*WORD*}

上から明らかなように CC 文には話者 (speaker) が記録されることもあり、記号 ">>"、">>>" は各々話者、話題の変化時に出現する。CC に含まれる情報は、単に音ストリームを音声認識して得られる情報のみならず、音源同定、話者同定を経て得られるべき情報をも含んでいることに注意されたい。本稿で対象とするスポーツ中継のように、雑音を含む環境下での音声認識が十分な精度を上げられない現状では、その補完手段として CC を使うことは合理的と言える。

ところで実際のビデオデータでは、対応する音声ストリームと CC ストリームの出現に関し、時間差が生じ、完全な同期は取れていない。ライブのスポーツ放送では CC ストリームに時間遅れが生じ、映画では逆に CC ストリームが先行する場合もある。また、音声ストリームがすべて CC ストリームに写されるとも限らず、CC で欠落したり、誤ったりする文がある。

```
049193
>>PAT : BRETT FAVRE THROWS A PERFECT
049245
PASS.
049255
THAT WAS MAXIMUM PASS
049318
PROTECTION.
```

図 1: クローズドキャプション (CC) の例

ここで用いる CC データの例を図 1 に示す。CC は 6 桁の整数とそれに対応する文字列の組により構成されている。この数字は CC 列の最初の文字が出現する画像フレームに対応し、1/30 秒毎にインクリメントされる。

またコロン ":" 以前が話者名、">>" は話者の転換をあらわし、あらかじめその話者の音声の特徴を獲得していれば高精度の対応づけができる。

2.3 音, 言語ストリーム間の対応付けの問題点

スポーツ中継における CC はリアルタイムでタイプ入力されるため、音声の発生時刻と CC の出現時刻に大きな差が生じる。そのため最も時刻の近い音声と対応づけるような単純な手法では正しい対応づけができない。また、タイプスピードよりも速く話されたときに、発声内容を要約した CC が生成されるため、CC から発声時間を推定して対応づけることも困難である。また、発話された文 (主に短文か単語) が CC では欠落する場合もある。要するに、CC ストリームは音声ストリームの不完全な写しであり、単純に対応付けが可能となるわけではないことに注意されたい。

3 音声と CC の同期手法

本手法では、要約された CC の文が実際に発声された文に含まれていることに注目し、音ストリームからは雑音環境下においても比較的容易に検出できる母音のみを検出する。検出したそれぞれの母音区間からフォルマント周波数を求め、母音区間の音素を推定し、音素系列 (系列 A) を生成する。また、言語ストリームからは単語辞書を用いて音素系列 (系列 B) を生成する。最後に系列 A から系列 B を DP マッチングを用いた対応付けし、音, 言語ストリーム間の同期と取る。

以下に各音素系列の生成方法および音素系列間のマッチング手法について述べる。

3.1 母音区間検出

音声検出において子音区間は、雑音となる会場の歓声のある環境下での検出が困難である。そこで誤認識の起こりにくい高調波構造をもった母音区間の検出を行う。

大勢の歓声の声が重なるとホワイトノイズのようになり、調波構造をもたなくなる [8] ため母音区間検出法にはケプストラム法 [9] [10] を用いる。スペクトル微細構造から基本周波数を求め、

1. スペクトル成分が閾値以上
2. 基本周波数が人間の声の範囲内
3. 検出区間内での基本周波数の分散が閾値以下

の条件を満たすものを、母音区間であるとして検出する。但し、検出区間内での基本周波数の分散が小さいものは、音楽区間であると認め、除去する。

3.2 音素推定

テレビのスポーツ中継では観客の歓声が約 700Hz を中心に分布しており、ケプストラム法ではフォルマント周波数が埋もれてしまう問題点がある。そのため、よりスペクトルのピークを重視した LPC ケプストラム法 [9] [10] を用いて、スペクトル包絡を求める。

またCCから話者が特定されているため、あらかじめ各話者について各母音の第1・第2フォルマント周波数 (F_1, F_2) をモデルとして獲得しておき、各音素との類似度により音素の推定を行う。

本手法での母音は、/a/(AA), /æ/(AE), /ʌ/(AH), /ɔ/(AO), /ɛ/(EH), /ɝ/(ER), /i/(IY), /I/(IH), /U/(UH), /u/(UW) の10種類を用いる。ここではフレーム毎に音素を推定するため、二重母音は対象としない。

検出された母音区間の各フレームに対して、LPCケプストラム法を用いてスペクトル包絡を求める。さらに、第1・第2フォルマント周波数 (f_1, f_2) を求め、母音のフォルマント周波数のモデルとのF1-F2平面上での距離

$$D(a, X) = (F_{X1} - f_1)^2 + (F_{X2} - f_2)^2 \quad (X\text{は各母音})$$

を求める。ここで a は音素を表す。図2にこの様子を示す。

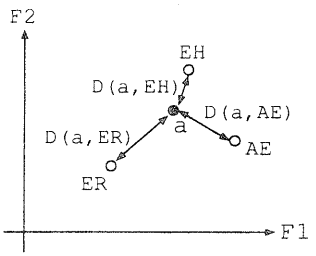

図 2: 音素の推定

音素 a は $D(a, X)$ を最小とする母音 X であると推定する。この距離 $D(a, X)$ を対応付けの際の類似度に反映させる。

3.3 CCの音素系列作成

CCは一文毎に改行コードが含まれており、その改行コードまでを一つのセグメントとする。図1を例にとると、">> PAT:...PERFECT", "PASS", "THAT WAS MAXIMUM PASS", "PROTECTION" がそれぞれ一つのセグメントとなる。

セグメント内のそれぞれの単語毎に単語辞書を参照して母音のみの音素系列(系列 B)を生成する。図3に音素系列生成の流れを示す。

単語辞書には「The CMU Pronouncing Dictionary [11]」を用いた。

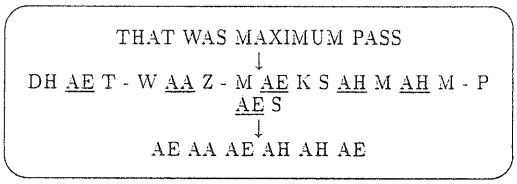

図 3: CCからの音素系列

3.4 DPマッチングによる音素系列間の対応付け

異なる時間軸をもつ二つの時系列を、時間軸を非線形に伸縮し対応付けを行う手法として一般にDPマッチングが知られている。音声認識の分野においてDPマッチングが用いられる場合は、始端と終端が対応している場合が通常である。ところが、超長配列からある配列の近似パターンを照合するときにDPマッチングを適用する報告もある [12]ため、ここではマッチングにDPマッチングを用いる。

一般にCCは実際に話されてから約3～5秒遅れて表示される。従ってここでは、CCが表示される10秒前からCCが表示されるまでの10秒間の音声から生成された音素系列を用いる。

音声から生成された音素系列 A, 並びに1セグメント分のCCから生成された音素系列 B を

$$A = \{a_1, a_2, ..., a_n\}$$

$$B = \{b_1, b_2, ..., b_m\}$$

として、B との類似度 $score$ を最大とする $A' = \{a_i, a_{i+1}, ..., a_j\}(0 \leq i < j \leq n)$ をDPマッチングを用いて求める。

CCから生成される一音素の時間に対して音の1フレームは非常に短いため、系列 B の複数の要素を系列 A の要素に対応付けることは許さない。また系列 A' が長くなりすぎ複数の文にわたることにも注意が必要である。

以上を考慮し、類似度 $score$ を求める初期値及び漸化式を以下のように定める。ここで Δ は正定数である。$score(j, m)$ を最大とする j を求め、バックトラックにより $score(i, 0) = 0$ とする i を求める。これによって、音素系列 B と対応する部分系列 $A' = \{a_i, a_{i+1}, ..., a_j\}$ が得られる。

- 初期値

  $score(p, 0) = 0 \ (0 \leq p \leq n)$

- 漸化式

  $score(p, q)$

  $$= max \begin{cases} score(p-1, q) - \Delta \\ score(p-1, q-1) + (1/D(a_{p-1}, b_{q-1})) \end{cases}$$

$(0 \leq p \leq n, 1 \leq q \leq m)$

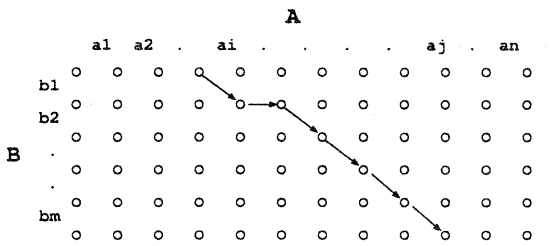

図 4: 部分系列との DP マッチング

4 実験および評価

本手法の性能を評価するため、アメリカンフットボールのテレビ中継のビデオデータ (英語音声) を用いて実験を行った。使用したビデオデータは CM の終わりから次の CM の始まりまでの約 2 分間であり、音は、サンプリング周波数 8kHz, 量子化ビット数 16bit, モノラルである。また、フレーム長は 512, フレーム周期は 128 とした。CC は、セグメント数 52, 単語数 220, 音素数 768 であった。

まず、母音区間検出結果により、音ストリームからの音声区間の検出効率を評価する。次に、推定した音素系列が CC から生成した音素系列を含むか調べることにより、音素推定の効率を評価する。最後に、音素推定した音素系列と CC から作成した音素系列間の対応付けを DP マッチングによって行った結果を示し、本手法の有効性について評価する。

4.1 母音区間検出結果

全フレーム (7535 フレーム) 中、連続音声として聞き取れる音声区間 (6677 フレーム : 約 106.8 秒) を正答とした。提案手法による検出結果を表 1に示す。ここで適合率、再現率は以下のように定義される。

$$\text{適合率} = \frac{\text{正検出数}}{\text{検出数}}, \quad \text{再現率} = \frac{\text{正検出数}}{\text{正答数}}$$

表 1: 母音区間検出結果

	フレーム数 (秒)
検出数	6078(97.2)
正検出数	5339(85.4)
誤検出数	739(11.8)
未検出数	1305(20.8)
適合率	87.8 %
再現率	80.3 %

再現率は、歓声の有無にかかわらず 80 % 前後であった。これは連続音声として聞き取れる範囲を音声区間としたため、正解フレームには子音区間が多く含まれ未検出が多くなったことが原因である。母音区間のみの検出では、子音の検出までは困難であるため再現率は低い値となった。

適合率は歓声のない箇所では比較的良好な値であった。しかし歓声のある区間では 歓声を誤って音声区間として検出したことがあったため、全体的には 87.8 % となった。

歓声のある箇所においても、歓声のない箇所と同等の再現率を得られた。これは本手法が歓声の中での母音区間検出として有効であることを示している。

4.2 音素推定結果

実験データ内では二人の話者 (アナウンサーとコメンテーター) が登場した。本実験データとは別の歓声のないシーン約 1 分間を用いて手動で与えた母音部分を対象にし、予め求めた二人の各母音の第 1・第 2 フォルマント周波数のモデルを表 2 に示す。

表 2: 母音の第 1 ・第 2 フォルマント周波数

母音	話者 A	話者 B
AA	701.7Hz, 1075.0Hz	775.9Hz, 1155.4Hz
AE	627.0Hz, 1685.6Hz	693.0Hz, 1828.1Hz
AH	494.0Hz, 1128.8Hz	547.5Hz, 1259.7Hz
AO	541.5Hz, 798.0Hz	601.8Hz, 882.0Hz
EH	508.8Hz, 1840.0Hz	557.3Hz, 1950.4Hz
ER	480.2Hz, 1282.5Hz	519.4Hz, 1429.1Hz
IH	382.2Hz, 1910.4Hz	411.2Hz, 2103.7Hz
IY	261.9Hz, 2267.1Hz	286.6Hz, 2430.7Hz
UH	432.2Hz, 1002.8Hz	465.1Hz, 1075.4Hz
UW	294.0Hz, 852.6Hz	318.9Hz, 917.2Hz

検出した全フレームに対する正誤を決定するのは困難なため、CC の 1 セグメントごとに手動により音声区間を対応付け、その範囲内で、距離 $D(a, X)$ を最小とする音素 X(第 1 候補) による系列に、CC からの音素系列 B の順に現れているものを正答とした。同様に、距離 $D(a, X)$ を二番目に小さい値にする音素 X(第 2 候補) までを許したものも検証した。表 3 にその結果を示す。

表 3: 音素推定結果

総セグメント数	55
第 1 候補	54 (98.1 %)
第 2 候補	55 (100.0 %)

第 1 候補のみを許したときには、全ての CC において正しく音素推定できかったが、第 2 候補まで許すことにより全ての CC において正しく音素推定できている。

4.3 マッチング結果

各 CC セグメントについて、マッチングによって得られた開始時間, 終了時間と実際に対応する音声区間の開始時間, 終了時間ともに、0.5 秒,1.0 秒,2.0 秒以内の誤差のものを正解とした結果を表 4 に示す。

表 4: 音素推定結果

総セグメント数	55
0.5 s	21 (38.1 %)
1.0 s	32 (58.1 %)
2.0 s	37 (67.2 %)

実際に対応する音声区間以外の音素との類似度が高くなり、結果が大きくずれた例が見られた。対処法として他の CC セグメントのマッチング結果を用いて修正する手法が考えられる。

5 まとめ

音声からはフォルマント周波数による音素推定し、CC からは単語辞書を用いた音素決定し、両音素列を対応付けることによって、音声と CC の同期手法を提案した。これにより音声認識を行うことなく容易に発話内容を得ることができ、また CC において他のストリームとの時間的同期性を高めることが可能となった。

今後、マッチング法の改良による対応付け精度の向上、及び音素系列間のマッチングに計算時間を多く費すため、マッチングに用いる音声から生成された音素系列を適当な箇所に限定して効率をあげることなどを検討する。

謝辞

CC からの音素決定部分には、CMU によって作成された “The CMU Pronouncing Dictionary” を使用した。作成に携わった関係各位に感謝する。

本研究の一部は、日本学術振興会科学研究費・基盤 (B)(代表：馬場口) の補助を受けている。

参考文献

[1] N. Babaguchi, S. Sasamori, T. Kitahashi and R. Jain, “Detecting Events from Continuous Media by Intermodal Collaboration and Knowledge Use,” Proc. IEEE ICMCS'99, Vol.1, p.p.782-786, June, 1999.

[2] 馬場口登, “メディア理解による映像メディアの構造化”, 電子情報通信学会技術研究報告, PRMU99-42/IE99-18/MVE99-38, pp.39-46, July, 1999.

[3] S.Satoh, Y.Nakamura and T.Kanade : "Name-It: Naming and Detecting Faces in News Videos", IEEE Multimedia, pp.22-35, Jan–March 1999.

[4] 新田 直子, 馬場口 登, 北橋 忠宏, “言語情報による連続メディアからの人物オブジェクトとアクションの抽出”, 電子情報通信学会 情報・システム総合大会, D-12-192, 1999.

[5] 柳沼良和, 和泉直樹, 坂内正夫, "同期されたシナリオ文書を用いた映像編集方式の一提案", 電子情報通信学会論文誌 (D-II), Vol.J79-D-II, No.4, p.p.547-558, 1996.

[6] 谷村 正剛, 中川 裕志, “ドラマのビデオ音声トラックとシナリオのセリフの時刻同期法”, 情報処理学会研究報告 知能と複雑系 118-4, 1999.

[7] 谷村 正剛, 中川 裕志, “テレビドラマにおけるシナリオのセリフと音声トラックの同期システム” 人工知能学会 第 13 回全国大会, p.p.205-208, 1999.

[8] 梶田将司, 小林大祐, 武田一哉, 板倉文忠, “ヒューマンスピーチライクノイズに含まれる音声的特徴の検討”, 日本音響学会誌 Vol.53, No.5, p.p.337-345, 1997.

[9] 古井 貞熙, “音響・音声工学”, 近代科学社, 1992.

[10] L.Rabiner, B-H,Juang 著, 古井 貞熙 訳, “音声認識の基礎 (上・下)”, NTT アドバンステクノロジ, 1995.

[11] The CMU Pronouncing Dictionary, Copyright 1993, 1994, and 1995 by Carnegie Mellon University, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[12] 伊藤実, 清水邦保, 中西通雄, 橘本昭洋, "文字列集合における識別文字列を求めるための多項式時間手続き", 電子情報通信学会論文誌 (D-I), Vol.J77-D-I, No.8, p.p.531-538 1994.