字幕表示のための VCML プロジェクトの研究開発の現状

– 環境情景音表示と VCML 文書の木構造化 –

鈴木 隆広[1], 栗田 将史, 杉山 雅英

あらまし　本報告ではビデオ字幕表示システムのための VC プロジェクトの最新状況について述べる。”VCML (Video Caption Markup Language)” は映像データに付加する字幕を制御するための言語であり、VCML で記述された文書は ”VCML Player” によって処理され、映像と字幕がリアルタイムで合成再生される。本報告では VCML Player (v2.3) の新しい 2 つの機能を述べる。1) VCML 文書の木構造化機能、2) 環境情景音の表示機能。VCML 文書の木構造化機能は大規模な VCML 文書を分割し効率良く整備するために有効である。また環境情景音の表示機能は言語情報以外の環境情景音字幕表示の要望に答えるものである。さらに本報告では幾つかの種類のビデオに対する VCML 文書の整備状況について述べ、その文書の書き起し文の性質について述べる。さらに環境情景音の一つである「笑い声」の自動検出手法の研究の現状について述べる。

Latest Achievement of VC Project for Automatic Video Caption Generation

T. Suzuki, M. Kurita, M. Sugiyama

Abstract　Abstract This paper describes the latest achievement of our VC project for video caption displaying system. ”VCML (Video Caption Markup Language)” has been designed for controlling video captions and video data. The VCML document is processed by ”VCML Player”, and video data and captions are composed in real time. This paper describes two new functions of VCML Player(v2.3): 1) tree-structured VCML document, 2) displaying auditory scenes. The first function is useful for generation and maintenance of large size VCML documents. The second function corresponds to the desire of expansion to auditory scenes which express video scenes. This paper also describes our progress of the construction of VCML documents for several kinds of video data and analysis of sentences in VCML documents. Furthermore, performance of automatic laughter detection is described.

1　はじめに

現在、日本の聴覚障害者の数は 35 万人と報告されている。また近年は高齢化が進み、テレビ放映の視聴に何らかの不自由を感じる人が多くなってきた。平成 13 年に日本で放映された全番組の中で、字幕付きで放送されている番組は NHK 総合で 22.95%、民放では 6.3%に過ぎない。この数字は年々増加してはいるものの、まだ需要に応えるだけの数は無いというのが現状である。

NHK では数年前から字幕放送に関する研究を行っており、2000 年には「ニュース７」の放送を開始した [1, 2]。これはアナウンサーの声を音声認識を通じて字幕化している。また変換に誤りが生じた場合には修正が行えるようなシステムも開発された。また、リスピークによる雑音下での音声認識の性能改善も行っている [3]。また国際会議での発言をリアルタイムで字幕化し、変換誤りなどを人手で修正後投影するという試みも行われている [4]。

我々は 1997 年からビデオ字幕のための VC プロジェクトを行っている [5, 6, 7, 8, 9, 10]。これは字幕のシナリオを既知のものとして、音声を用いて字幕表示の時間情報を計測する。また生成された時間情報を記述する言語として VCML(Video Caption Markup Lamguage)、それを表示するためのシステムである VCML Player の開発も行った。字幕作成、表示の流

[1]会津大学大学院 コンピュータ理工学研究科, Graduate School of Computer Science and Engineering, The University of Aizu

れは図 1 のようになる。

VCML の設計には今後の拡張や発展性を考え XML を利用した。VCML では製作者やユーザーは映像に直接文字等を書き込むことなく、また特別な編集ソフトを使用することなく、容易に字幕付き映像を作成することができる。

VCML ではタグで文書を記述する。例えば文字のサイズや色、文字につける影の色などを指定できる。通常は表示位置を top,bottom などから選ぶが、座標を指定することでその位置に字幕を表示させることができる。また字幕の横及び縦表示、表示している文字列のスクロール機能などがあり、ユーザーはエフェクトを付けたり、思い通りの位置に字幕を表示させることができる。現在表示しているものとは別の映像を画面に重ねることもでき、そうして表示させた映像に別の字幕を付けることも可能である。

本報告では、VCML 文書の整備とそのために有効な文書の木構造化について、また環境情景音の検出についても述べる。環境情景音にも様々なものがあるが、本報告で取り上げたのは “観衆の笑い声” である。さらに環境情景音を表示するために行った VCML Player について述べる。

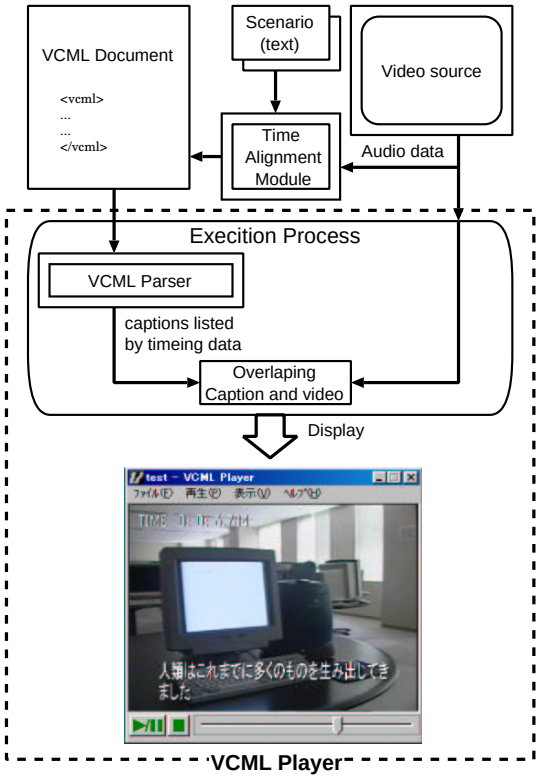

図 1: 字幕表示の流れ

2 環境情景音のシンボル表示

VCML Project では環境情景音を自動的に検出する試みを続けている。そしてそれを表示するための機能をプレイヤーに組み込んだ。それらの情景音は、VCML 文書中に記述する事で、指定したタイミングにその情景音を示すシンボルを表示する事が可能となる。

2.1 情景音シンボルの表示

VCML ではシンボルの表示を<symbol>というタグで行う。記述は図 2.1 のようになる。この例では再生後すぐに”laugh” という要素で指定されているシンボルを表示し、1 秒後に消去する。また、“align” で表示する位置を指定する。

```
<time begin=0.0 end=1.0>
  <symbol type="laugh" align="center"/>
</time>
```

図 2: Symbol の記述例

VCML Player では表 2 のように、幾つかのアトリビュートとそれに対応するシンボルが定義されている。

表 1: symbol のアトリビュートと効果

<symbol>	attribute	
	type	表示するシンボル
	align	表示位置の指定

表 2: 要素と対応するシンボル

	type	表示
笑い声	laugh	[笑い声]
拍手	handclap	[拍手]
ノックする音	knock	[ノック]
電話のベル	bell	[電話]
風の音	wind	[風の音]

図 3: 環境情景音「笑い声」と字幕の表示例

2.2 ユーザーによる情景音シンボルの定義

VCML Player ではいくつかのシンボルが事前に定義されているが、あらかじめ用意されていない情景音のシンボルを表示させようとしたとき、字幕の制作者がオリジナルのシンボルとそのアトリビュートを定義できるようにした。VCML では、<sbldef>というタグを<body>中に記述することで、新しいシンボルを指定させることができる。具体的な記述例は図 4 のようになる。この例では"rain"という名前の要素に対して、[雨音] というシンボルを記録させる。これによって新しいシンボルが定義され、symbol タグを使用してそれを表示させる事が出来る。

VCML Player ではあらかじめいくつかのシンボルが定義されていることは前述したが、このタグで再定義することにより、それらのシンボルをユーザーが置き換えることが可能となる。

```
<sbldef name="rain" str="[雨音]"/>
```

図 4: "sbldef"の記述例

今後、ビットマップで作成されたシンボルの表示にも対応させていく予定がある。その際表示するイメージに関する検討も行っていく。

3 VCML 文書の木構造化

長時間の映像にはそれに比例して大量の字幕テキストが必要となる。また VCML Player では字幕表示言語を切り替えることが可能なので、使用する言語の数に比例してテキスト量は増えることになる。これまで、一つの映像に字幕を付ける際には一つの VCML 文書ファイルで全てを表現していたため、ファイルが長大化し、後に編集や整備を行う際に困難を来すという問題があった。

もし長い文書ファイルを章や言語毎に幾つかに分割しておけば、文書をある程度区切って編集出来るようになる。そこで、VCML に文書の分割を行うのに必要な機能を付加した。ルートとなる VCML 文書ファイルから分割されたファイルを指定し、VCML Player に読み込ませることで、分割前と同じように処理させる。

VCML 文書ファイル中から別の文書ファイルを呼び出すために、<vcfile>というタグを用いる。このタグの書式は図 5 のようになっている。ここの src で指定されたファイルを VCML Player が読みとり、処理を行う。階層に制限はないが、再帰的にファイルを呼び出すことや、親となるファイルを呼び出すことは禁止される。また、呼び出す文書中の時間にバイアスを加えることができる。

```
<vcfile src="filename.vcml" bias="1.0"/>
```

図 5: vcfile の記述例

3.1 言語による分割

会津若松市の観光案内ビデオ (18 分 48 秒) に対して字幕作成を行った。字幕として表示されるテキストはビデオ中の音声から書き起こしされている。このテキストにはさらに英訳も行った。ここで、日本語と英語の二種類の字幕を表示させることを考える。

VCML Player では、表示の機能の一つとして字幕表示言語の切り替え機能がある。会津観光案内のビデオではこれを利用して、日本語、英語の字幕を表示させる。表示結果は図 6 のようになる。

ここで、字幕部分の記述を、日本語と英語に分割し、それぞれを一つのファイルとして記述した。その際の記述が図 7 になる。日本語と英語それぞれの VCML ファイルを、ルートになる文書ファイルから呼び出している。

3.2 重複部による分割

会津大学の入学案内ビデオは年度毎に作成されている。このビデオの前説の部分で大学の概要を説明

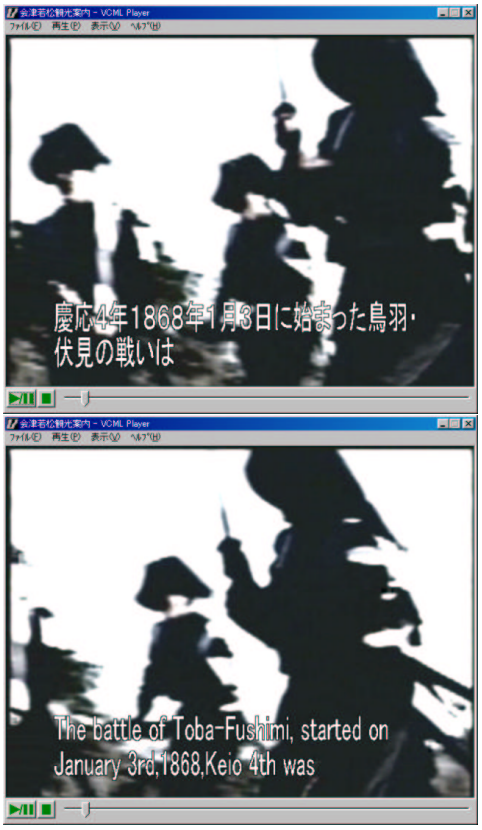

図 6: 二カ国語の字幕の切り替え (会津観光案内)

しているが、この部分だけは毎年同じ内容になっている。そこで、この共通部分の VCML 文書を一つのファイルに収め、各年度の VCML 文書の冒頭部分でそれを呼び出すことにした。

ただし、この入学案内ビデオのデータは VHS から取り込んでいるため、各年度の冒頭説明が始まるまでの時間に差が出てしまう。そのため、ただ共通部分を読み込むだけでは発話と字幕表示に誤差が生じてしまう。そこで、呼び出した VCML ファイル内で指定されている表示時間に適当なバイアスを加えてやることで、タイミングを調整する。こうして作成された VCML 文書が図 8 になる。

3.3 情景音の分割

落語ビデオなどでは言語音に対する字幕の他に、笑い声などに対して情景音を視覚的に表示する記述がある。この情景音の記述の部分を別のファイルに記述したものが図 9 になる。

観光案内ルートファイル

```
<vcml>
<head>
  <title>会津若松観光案内</title>
  <rootlayout width="640" height="480"/>
</head>
<body>
<time begin="0.0" end="18m48s">
  <video src="SightSeeing.avi"/>
</time>
<!--観光案内日本語ファイル-->
<vcfile src="kanko_jp.vcml"/>
<!--観光案内英語ファイル-->
<vcfile src="kanko_en.vcml"/>

</body>
</vcml>
```

観光案内日本語ファイル

```
<vcml>
<body>
<time begin="00m15s" end="00m21s">
  <caption language="Japanese">
    慶応４年１８６８年１月３日に始まった
    鳥羽・伏見の戦いは
  </caption>
</time>

...
</body>
</vcml>
```

観光案内英語ファイル

```
<vcml>
<body>
<time begin="00m15s" end="00m21s">
  <caption language="English">
    The battle of Toba-Fushimi, started on
    January 3rd,1868,Keio 4th was
  </caption>
</time>

...
</body>
</vcml>
```

図 7: 会津観光案内の VCML 文書-木構造化後

入学案内冒頭部分

```
<vcml>
<body>
<time begin="00:06" end="00:07">
  <caption>人類はこれまでに</caption>
</time>
...

</body>
</vcml>
```

2000 年度入学案内

```
<vcml>
<head>
  <title>2000 年入学案内</title>
  <rootlayout width="640" height="480"/>
</head>
<body>
<time begin="0.0" end="15:00.0">
  <video src="Aizu-2000.avi"/>
</time>

<vcfile src="Aizu-intro.vcml" bias="0.25"/>

<time begin="02:46" end="02:50">
  <caption>２１世紀の技術を先導する</caption>
</time>
...
</body>
</vcml>
```

2002 年度入学案内

```
<vcml>
<head>
  <title>2002 年入学案内</title>
  <rootlayout width="640" height="480"/>
</head>
<body>
<time begin="0.0" end="15:00">
  <video src="Aizu-2002.avi"/>
</time>

<vcfile src="Aizu-intro.vcml" bias="0.1"/>

<time begin="02:47" end="02:50">
  <caption>会津大学は、非常に特徴のある大学で
す</caption>
</time>
...
</body>
</vcml>
```

図 8: 入学案内の VCML 文書-木構造化後

米朝落語-言語音字幕

```
<vcml>
<head>
  <title>米朝落語</title>
  <rootlayout width="640" height="480"/>
</head>
<body>
<time begin="0.0" end="15:00.0">
  <video src="Beicho-01-01.avi"/>
</time>

<!--米朝落語-情景音表示-->
<vcfile src="Beicho-01-01.laugh.vcml"=/>

<time begin="00:39" end="00:40">
  <caption>私はこの落語をやる時に</caption>
</time>
...
</body>
</vcml>
```

米朝落語-情景音表示

```
<vcml>
<body>
<time begin="00:69.33" end="00:74.50">
  <symbol type="laugh"/>
</time>
...
</body>
</vcml>
```

図 9: 米朝落語の VCML 文書

4 VCML 文書について

4.1 VCML 文書の作成

作成を行なった vcml 文書は以下の 3 種類である。会津観光案内ビデオは複数人による演劇であり、米朝落語は寄せで行われた落語である。会津大学案内ビデオは主に朗読音声から構成されている。表 3 にそれらに含まれる文の数、継続長を示す。ここで文の数は書き起し者が字幕表示を念頭に文の区切りを与えたものであり、句点による文の区切りよりは読点による区切りに対応している。

- 会津観光案内ビデオ (SightSeeing)
- 米朝落語 (Beicho)
- 会津大学案内ビデオ (Aizu)

表 3: 映像及び VCML ファイルの規模

	文数	映像継続長	字幕継続長
会津観光案内	149	18:48	17:41
米朝落語			
帯久 (01-01)	895	45:43	45:24
足上り (01-02)	523	26:14	26:01
会津大学案内			
1997	156	14:01	13:30
1999	168	13:36	13:03
2000	146	13:36	13:03
2001	151	15:29	14:57
2002	161	16:11	15:43

VCML 文書作成の手順は以下の通りである。1. ビデオ、LD から AVI ファイルの作成, 2. 書き起しテキストの作成, 3. 書き起しテキストの検査, 4. 書き起しテキストの修正, 5.vcml 形式への変換, 6. 英訳文の付与、7. 音響情景シーン情報の付与。

4.2 作成の手順

4.2.1 書き起しテキストの作成

AVI ファイル化したビデオデータを WINDOWS 上の MediaPlayer で確認しながら発話内容のテキストの書き起し及び時間情報の付与を行う。図 10 の形式でそのテキスト (拡張子: txt) を作成する。MediaPlayer の制約で時間分解能が秒単位となっている。時間情報については書き起こしテキストに基づいて時間整合モジュールを用いて自動的に与えることができるが、ここで目視で与えた時間情報をモジュールの性能の評価に使用できる。

```
00:39 <私はこの落語をやる時に> 00:40
00:41 <あんまり袴はいたことがないねん> 00:43
00:44 <なんかこう嫌いでんねんこれ> 00:45
00:46 <正月だとかねまあのこじょうであるとか> 00:48
00:48 <しゃーない時にはつけますが、まあー、あとはまあ> 00:50
00:51 <婚礼か葬式かぐらいのものでございまして> 00:54
00:54 <まあるというところをみて頂くわけでございますが、
> 00:56
00:57 <まあ話によってはこんなもんつこうたほうが> 00:58
00:59 <ええ場合もございます。> 01:00
```

図 10: 書き起しテキストの例

4.2.2 英訳について

生成した VCML 文書中の和文を英訳した。英訳を行う際に以下のような原則に従った。現時点で英訳が完了しているのは会津観光案内ビデオと会津大学案内ビデオ (2001, 2002) である。

1. 複数の可能性がある場合

   分かりやすい表現、文字数の少なくなる表現を採用する。

2. 英文と和文との対応

   各文毎に英文をつける文毎の直訳とする。

3. 和歌などの英訳

   観光案内ビデオには和歌 (辞世の句)[2]や漢詩[3]が含まれている。これらには意訳をつけることとする。

4.2.3 音響情景シーン情報の付与

言語情報の書き起しだけでなく、ビデオ音声に含まれる様々な音響情景シーンの情報を字幕として表示することが望まれている。落語ビデオでは寄席の観衆の笑い声の情報を目視で与えた。その結果が図 11 となる。

```
 69.334602     1     74.508620
 78.393538     1     82.148292
194.850824     1    197.157138
208.203875     1    209.603678
```

図 11: 音響情景シーン付与の例

4.3 書き起こし字幕文の性質

図 12 に会津観光案内, 落語 (Beicho-01-01), 大学案内 (Aizu-2002) の文書中の各文の継続時間長の分布を示す。これらから継続時間長の分布は全て 2 秒に集中していて大きな差は見られないが、朗読調の大学案内は特に 2 秒に集中していること、会津観光案内の継続長は比較的広く分布していることがわかる。図 13 に観光案内, 落語 (Beicho-01-01), 大学案内 (Aizu-2002) の各文書中の文字数の分布の比較を示す。図 13 からは朗読と自由発話に対する文字数の

[2]蒲生氏郷の辞世の句「限りあれば　吹かねど花は散るものを心みじかき　春の山風」

[3]詩経の「采薇」の一節「昔われ往きしに楊柳は依依たり、今われ来れば雨雪霏霏たり」

分布に差があることがわかる。VCML Player では字幕表示に 1 行 20 文字で 2 行を用いるので最大 40 文字表示可能である。図 13 から、書き起こし字幕は全て 40 文字以内の制約を満たしていることがわかる。

図 14 に落語の話速を示す。ここで話速は 1 文の発話時間で文に含まれる文字数を割って計算した。文字は漢字及びかなまじりであり音節ではない。落語の平均話速は 7.42 文字/秒であり、最小値 1.91、最大値 28.57 である。ここで最大を与えるのは「はい？」の短い文である。

図 15 に落語 (01-01) の各文の文字数と話速との関係を示す。1 文に含まれる文字数が大きく変化するのに対して話速の変化はそれほど大きくないことがわかる。

会津大学案内に対して、同一内容の文において文の区切りの仕方の違いが見られた。これらについて統一化する必要がある。

5 笑い声の検出

聴覚障害の方々の意見として、映像中の人物の声以外の環境情景音も視覚的に表示してほしいという要望がある。環境情景音とは風の吹く音や電話のベルが鳴る音などのことである。

音響情景音を自動検出することでその表示タイミングの決定は容易になり、字幕作成が短時間で作成可能となる。本報告では区間単位での笑い声の検出 [11] の現状について述べる。

5.1 実験データ

笑い声の評価を行うために「古今亭志ん生名演集」の落語 CD の第 1 巻第 1 話及び第 2 話を用いた。また詳細な音響シーンの時間情報を記述したラベルデータを作成した [11, 12]。第 1 話には笑い声のシーンが 96 回、第 2 話には 99 回存在し、笑い声のシーン 50 回分を学習データ、残りを評価データとした。

5.2 区間単位での検出

一定の区間毎 (例えば 1 秒毎) に区切って検出する。第 n 番目の区間に対して以下の値を算出する。ここでは区間長を L_1 とした。

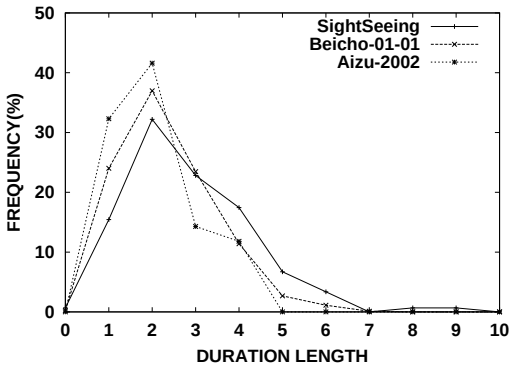

図 12: 文の継続時間長の分布の比較

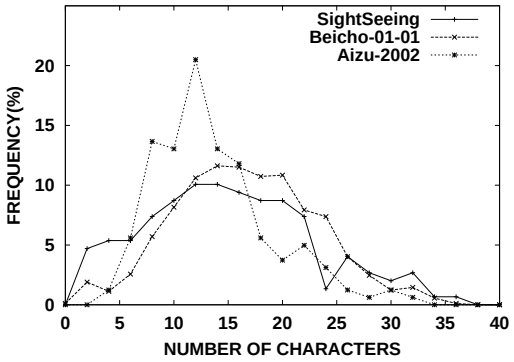

図 13: 文の文字数の分布の比較

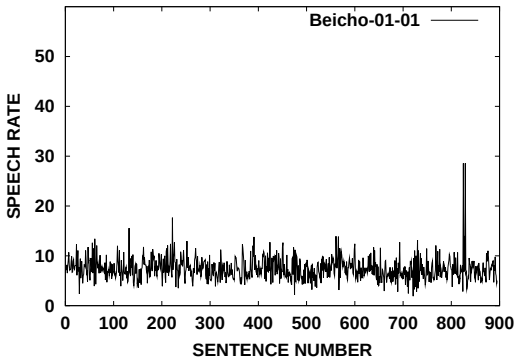

図 14: 文毎に見た話速 (落語 Beicho-01-01)

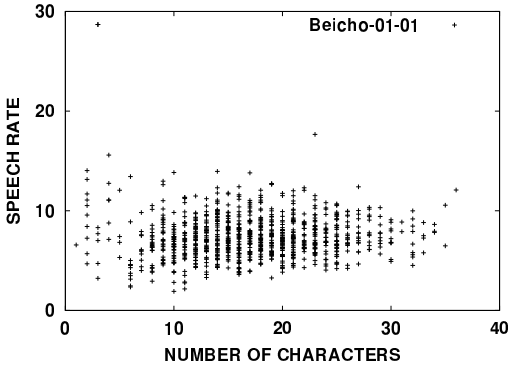

図 15: 文字長と話速との関係 (Beicho-01-01)

$$D_n^i = \sum_{l=1}^{L_1} \min_{1 \le m \le M} d(\boldsymbol{c}_{L_1 * n + l}, \boldsymbol{v}_m^i) \qquad (1)$$

これを各区間長ごとに行う。

5.3 区間単位での検出の評価

シーンラベルを用いてその区間が非笑い声より笑い声を多く含む場合には笑い声、そうではない場合は非笑い声とした。

また、実際の検出での誤りの多くが過渡区間 (一つの区間内でラベルが変化している区間) で起きていることが分かった。過渡区間を除外した場合の検出誤り率は第一話で 10.79%、第二話で 3.37%であり、過渡区間での検出誤りを少なくすることが今後の課題であることがわかった。

過渡区間以外での誤りで特に多かったのが笑い声とざわめきの判断が難しいものに関してだった。これより、パワーなどを考慮に入れて検出することが必要であることが示唆された。

6 むすび

本稿では、VCML Player2.3 の新しい機能である文書の木構造化と音響情景音表示について述べた。また VCML 文書の整備状況と情景音の一つとして「笑い声」の検出について述べた。今後の課題は以下の通りである。

表示用 Player に関しては、ネットワーク上におけるビデオファイルのストリーミングとその字幕表示が挙げられる。音声処理部分に関しては、「笑い」を含む音響情景音の自動検出、2 話者の重なり音声や BGM 下の音声のテキストと音声との自動対応法の開発が挙げられる。

また字幕作成に関連して、字幕書き起こしや検査をしやすくするためのオーサリングシステムの開発が挙げられる。

謝辞

会津観光ビデオを提供していただいた会津若松観光課、会津大学案内ビデオを提供していただいた会津大学学生課に感謝します。テキストの書き起こしおよび時間情報の付与を担当してくれた五十嵐理恵さんおよび橋本美穂さん、和文テキストの英語への翻訳作業を担当してくれた長谷川紀美子さんに感謝します。

参考文献

[1] NHK 技研 R&D No65, "ニュース字幕化特集", 日本放送出版協会, http://www.nhk.or.jp/publica/rd/rd65-j.html, (2001-06).

[2] 安藤, 他, "音声認識を利用した放送用ニュース字幕制作システム", 電子情報通信学会論文誌, Vol.J83-DII, No.6, pp.877-887 (2001-06).

[3] 本間, 他, 番組字幕放送のための音声認識 -システムの概要とリスピークの効果-, 電子情報通信学会/日本音響学会 音声研資, SP2002-50, pp.49-54(2002).

[4] 加藤, 他, 国際会議における聴覚障害者支援を目的とした音声字幕変換システムの設計, ヒューマンインタフェース学会, (2002-11).

[5] 深田, 他, 字幕表示用言語 VCML の設計とその表示システムの開発, 情報処理学会 ヒューマンインタフェース研究会, Vol., No.87-7, pp.37-42 (2000-01).

[6] K. Watanabe, N. Fukada, M. Sugiyama, Design of Video Caption Markup Language VCML and Development of VCML Player, Proc. of ICME2000, pp.163-166 (Aug. 2000).

[7] 鈴木, 杉山, 字幕表示用言語 VCML とプレーヤーの改良, 情報処理学会, 7Q-02 (2001-03).

[8] 鈴木, 他, 字幕表示システム VCML Player の新機能について, 情報処理学会 ヒューマンインタフェース研究会, 2001-HI-93-7, pp.39-45 (2001-05).

[9] T.Suzuki, K.Watanabe, M.Sugiyama, Enhancement of VCML Player, Proc. of MMSP2001, pp.365-370 (Oct. 2001).

[10] T.Suzuki, T.Kitazume, M.Sugiyama, Latest Achievement of VC Project for Automatic Video Caption Generation, Proc. of MMSP2002, (Dec. 2002).

[11] 金田, 杉山, 音響情景字幕表示のための笑い声の検出, 音楽講論, 3-P-3, pp.165-166 (2001-03).

[12] 栗田, 鈴木, 杉山, VCML Player 字幕生成のための笑い声の検出, 音研資, (2002-10).