Vol.2016-SLP-113 No.6
2016/10/6

スポーツ中継のリアルタイムデータからの解説音声自動生成実験

佐藤 庄衛[1,a)]　熊野 正[1]　清山 信正[1]　今井 篤[1]　山田 一郎[1]

概要：スポーツ競技のラジオ中継放送では，競技の進行状況が音声だけで理解できるように実況されている．一方，テレビの実況音声は，得点など映像上に表示されている情報や映像を見ただけで容易に理解できる情報の一部はコメントされていないため，音声だけで競技の進行を理解することが難しい．NHK では，視覚に障害のある方など映像を伴わずに視聴する方々が，音声だけで楽しめるスポーツ中継の実現を目指して，テレビのスポーツ実況を補完する音声ガイドの自動生成に取り組んでいる．本稿では，この自動生成システムの大規模な検証実験を紹介する．

A System Verification for Automatic Speech Guidance for Sports Broadcasts

SATO Shoei[1,a)]　KUMANO Tadashi[1]　SEIYAMA Nobumasa[1]　IMAI Atsushi[1]　YAMADA Ichiro[1]

1. はじめに

テレビ放送は，映像・映像中の文字と音声を通して，世界中の様々な出来事を伝えるメディアである．テレビの放送方式は，アナログ放送の (640×480 画素) から，デジタル放送 (1,920×1,080 画素) への移行に伴って高精細化され，今後，2020 年に向けて 4K や 8K と呼ばれる超高精細映像へと進みつつある．この高精細化に伴って，画面上にスーパーインポーズされるオープンキャプションの文字数も増加している．また，受信機側で必要に応じて画面上に文字情報を表示する仕組みも整備された．2000 年に開始されたデジタル放送でデータ放送が導入され，2013 年には，ハイブリッドキャストが開始され，放送波に多重された文字情報だけでなく，通信経由で取得した文字情報も表示できるようになった．**図 1** に，ニュース番組中のオープンキャプションの文字数を年代ごとに示した．1980 年代には一番組あたり 1,000 文字以下であったものが，近年では 3,000 文字程度まで増加している．

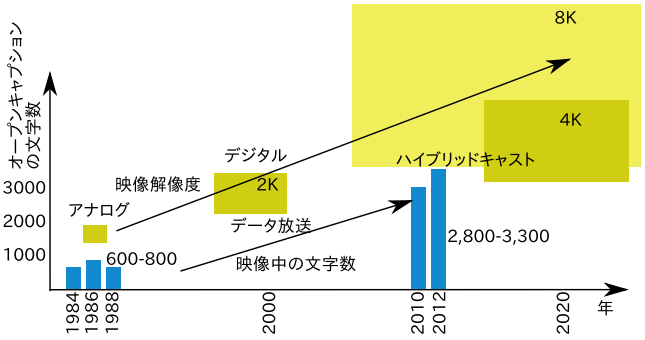

図 1　各放送システムの解像度と，ニュース番組中のオープンキャプションの文字数

一方，テレビ放送の音声で伝えられる単位時間あたりの言語的な情報の量は，年代や放送システムを問わずおおよそ一定であるため，増加した文字情報の全てを音声で伝えることが難しくなり，映像や映像中の文字情報を補足するコメントへと推移している．

このように，テレビの映像と音声の情報量の推移より，テレビを音声だけで楽しむことが難しくなってきていると推測される．さらに，多様なテレビの視聴スタイルが生まれたため，映像・音声の両方でテレビを楽しんでいる人と

[1] NHK 放送技術研究所
NHK Science & Technology Research Laboratories
[a)] satou.s-gu@nhk.or.jp

同じ受信機で映像なしにテレビを楽しむという局面は少なくない．視覚に障害を有する方をはじめ，子供とともにテレビを楽しみながら食事の準備をする人，車載テレビで家族が番組を楽しんでいる傍ら運転をする人，ワンセグ端末などで歩きながらテレビを楽しみたい人などにとっては，音声だけで楽しめて，テレビを見ている家族と情報を共有できる放送が望ましい．筆者らは，映像中の出来事や文字情報を合成音声で伝え，放送番組の音声コメントの不足を補完するシステムを目指している．

本稿では，大規模なスポーツイベントを対象として，実況音声だけでは番組を楽しめない現状を明らかにし，実況音声を補完する解説音声を自動生成するために構築したシステムを紹介し，その検証実験の結果を紹介する．

2. 解説付与の課題

2.1 解説放送制作の難しさ

解説放送は，番組中の人の動作や場面など音声だけではわからない情報を，副音声で提供するサービスである．解説音声は限られた収録済みの番組に提供されており，ラジオドラマと同等のコストと時間をかけて制作しているため，番組全体の約 10%の番組にしか付与されていない [1]．解説音声の制作が容易ではない理由は次の 2 点である [2]．

- 番組の演出意図に沿って提供すべき内容を決めることが難しい．
- 番組の主音声と重ならないように解説音声を挿入することが難しい．

2.2 人手による解説放送

生放送番組に付与することが難しい解説音声であるが，スポーツ中継を解説放送を利用して楽しみたいという要望は高い．NHK では，スポーツイベントのダイジェスト番組などで，リアルタイムに準じた方法で人手による解説音声の付与を試みている．NHK が 2014/3/10〜3/17 に放送したスポーツイベントのダイジェストに，準リアルタイムに付与した解説コメントの分類が**表 1** である [3]．調査対象は 30 分番組 6 回分の 180 分であり，アルペン大回転や車いすカーリングなど，約 20 競技が取り上げられている．付与された解説は 113 コメントであり，平均すると 4 分に 1 回しかコメントがない．また，映像の内容に関する 84 のコメントの内，約 7 割が実況音声のコメントと同内容であり有効な解説ではなかった．重複した内容は，競技者の転倒など注目すべきプレイに関するもので，旗門の通過などの競技の進行の理解に必要な情報の多くは，実況と解説のどちらからも得られなかった．それ以外の試合後の選手のインタビューの話者の解説や，外国語のインタビューに付けられた日本語訳の読み上げは，映像中の文字情報を音声で伝えたものであり，有効な解説であったと思われる．しかし，これら以外の映像中の文字情報については，分量が多く全てを読み上げることが困難であり，事前に伝えるべき情報の精査が必要であり，人手でリアルタイムに解説音声を付与することは容易では無いことがわかった．

表 1 準リアルタイムに人手で付与された解説コメントの分類

コメントの分類	コメント数
有効な映像解説	26
実況音声と同内容の映像解説	58
インタビュー話者名	20
日本語訳字幕の読み上げ	9
合計	113

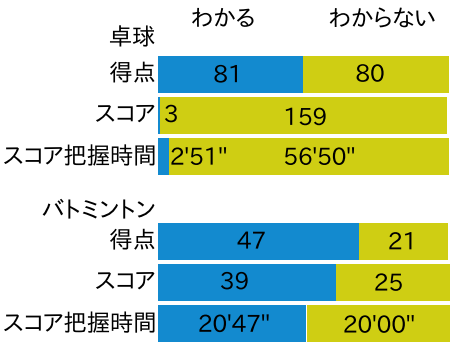

図 2 実況音声で発話される得点イベント，スコアと正しいスコアを把握できる時間

2.3 実況音声だけでは伝わらない情報

筆者らは，競技の勝敗に関わる情報がどの程度実況音声で発話されているかを調査している [3]．**図 2** はこの調査結果の抜粋であり，2012/8/5〜8/8 に放送された卓球団体戦とバトミントンダブルスの調査結果を示す．それぞれで，得点イベントを伝える発話の有無と，得点により更新されたスコア（何対何）を伝える発話の有無の割合を調査した．また，競技中に正しいスコアを把握できる時間を示している．[3] で行われた調査では，卓球団体戦は最もスコアの把握時間率が低かった競技で，バトミントンダブルスは把握時間率が最も高かった競技である．

バトミントンのように比較的スコアに関する実況が多い競技でも，得点に関するイベントを視聴者が数え上げない限り，半分の時間帯は実際と異なるスコアで競技を楽しむ事になる．一方，映像を伴って視聴した場合には，スコアは常時画面上にスーパーインポーズされており，得点イベントの有無も競技を見ていれば明白である．

また，映像上に表示された文字情報を実況音声で伝えないケースは，水泳などの順位やタイムを競う競技でも起こる．このような競技では，各トラックで同時に競技する選手のリストや，競技途中経過，競技結果の順位やタイムが表形式で映像上に表示される．2012/7/28〜8/1 に放送された 32 の水泳競技について，選手名，所属，タイムなどの表示された表中の要素を実況が伝えたかどうか調べた所，35%の要素しか伝えていなかった．

スポーツニュースやダイジェストのように最終結果が明確になれば良い番組とは異なり，スポーツ中継では，競技の進行を逐次把握するための情報が無いと競技を楽しむことができない．

2.4 実況音声のオーバーラップ

従来の解説放送では，放送の主音声である実況音声にオーバーラップしないという制約があった．ドラマなどを対象とした解説放送では，セリフがきちんと聞き取れることが必須となるが，スポーツ中継では，実況コメントよりも競技進行を把握できることが重要となる場合もある．今井らは，スポーツ中継を対象として，競技進行を伝える音声のオーバーラップの可否を調査している [4]．テニスとバスケットボールの競技に，得点，攻守の別，選手，ファールなどの客観的情報に基づいて，人手で用意した解説文を合成音声で再生した．ここで制作された解説音声は，テニスで 20%，バスケットボールで 70%が主音声の実況にオーバーラップしていた．視覚障害者 6 名によりサービスの必要性を主観的に評価した所，このオーバーラップに関わらず，評価者全員が「このサービスをぜひともほしい」と最高点の回答をした．この結果より，必要があれば解説音声をオーバーラップしてもよいと結論づけている．

3. スポーツ競技のリアルタイムデータ

近年，通信サービスの進展とともに，スポーツ競技の結果や進行をリアルタイムに配信するサービスが利用できるようになった [5], [6]．テレビのデータ放送などで提供される競技関連情報は，このデータを利用している．[5] では xml 形式で，競技のスコアやサーブ決定率などの統計情報，競技で起こったイベントなどを配信している．**図 3** は，配信された xml エレメントの例である．この例では，第 1 セット試合開始 9 分 21 秒に，日本の背番号 3 番の選手がサービスエースを決めて，スコアが 11 対 11 の同点になった事を伝えている．

これらのリアルタイムの配信データから解説音声を生成すれば，実況音声が競技の進行の全てを伝えていなくても，競技の進行が音声だけで理解可能になると期待される．しかし，この配信データは競技ごとに決められた公式記録の蓄積を目的とするもので，解説音声の生成を目的としたものではない．したがって，有効な解説音声を生成するためには，配信データが下記の要件を満たしているかを検証しなければならない．

- 競技の進行情報が配信されていて，それを容易に読み解くことができる．
- 配信データに付与された時刻情報が，十分な精度を持っている．
- 遅延なくデータが配信されている．

```
<UnitAction Type="UAC" Code="S1" Pos="61"
    Result="ACE" Value="SRV" Time="09:21"
    LeadH="0" LeadA="0" ScoreH="11" ScoreA="
    11" Rally="22" Win="H" Speed="59" Line="1
    ">
  <Competitor Code="VOW400JPN01" Type="T"
      Order="1" Organisation="JPN">
    <Composition>
      <Athlete Code="1084952" Order="1" Bib="
          3">
        <Description GivenName="Saori"
            FamilyName="Kimura" Gender="W"
            Organisation="JPN" BirthDate="
            1986-08-19" />
      </Athlete>
    </Composition>
  </Competitor>
</UnitAction>
```

図 3 配信データの例

4. 解説音声自動生成検証実験

4.1 システム概要

2016 年 8 月に開催された大規模スポーツイベントで配信された [5] を用いて，競技映像に解説音声を付与する実験を行った．**図 4** は，検証実験システムの概要である．本実験の主な検証項目は，解説音声の自動生成を目的として，配信データの詳細さと配信データのリアルタイム性を確認することにある．そこで，全ての競技の配信データを受信してそれらの受信時刻を調べるとともに，全ての競技の映像も同時に受信して適切なタイミングで解説音声を付与できることを確認した．45 チャンネルの全ての競技の国際映像 [*1] を標準画質に変換し，現地のネットワークと国内のネットワークを専用回線で接続して，配信データと映像データのデータベースを構築した．映像は MPEG-2 トランスポートストリームのマルチキャストで配信され，競技データは，http プロトコルの push, pull 機能により配信される．

また，解説音声のタイミングや解説内容の適切さを放送コンテンツ制作に携わる多くの人の評価を得るため，競技映像に解説音声を付したコンテンツを，NHK のイントラで閲覧できるシステムを構築した．

4.2 競技データからの解説文の生成

配信データのメッセージは，シュートや得点など競技のイベントを伝えるものと，対戦スコアや累積ペナルティなど競技の状態を伝えるメッセージに分類される．これらのメッセージは，競技記録を残すことを目的としているため，メッセージの一部はそのまま言語化しても，適切な解説音

[*1] 実況音声が付与されていない競技映像

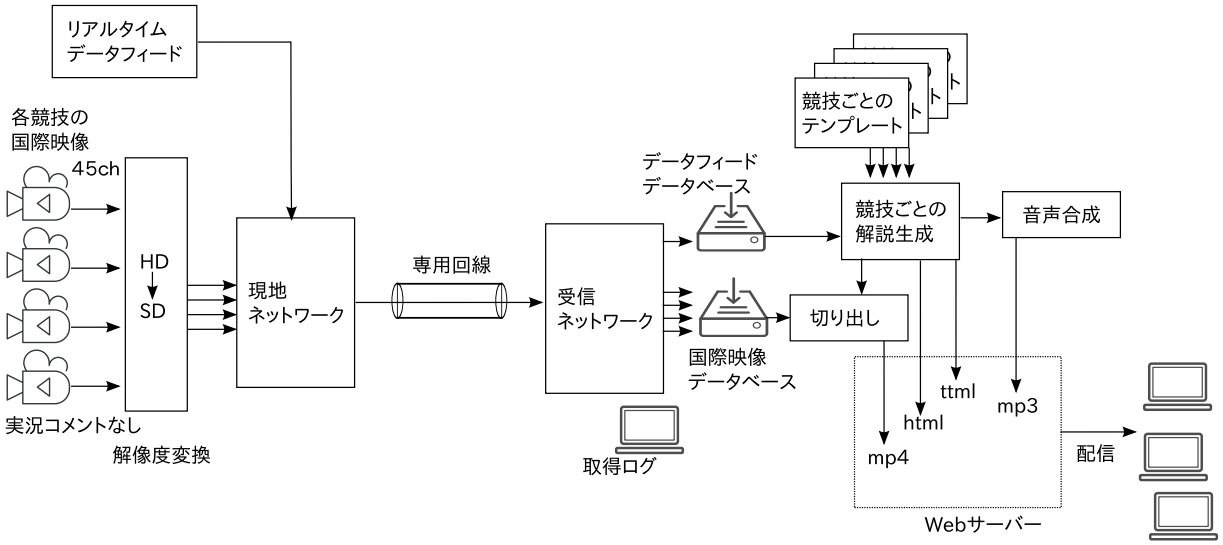

図 4 検証実験システム概要

声にはならない.

イベントを伝えるメッセージは，概ねそのまま言語化すれば良いが，競技の状態を伝えるメッセージと重複する情報が含まれる．一方，競技の状態を伝えるメッセージは，そのまま言語化できない場合が多く，状態の変更を検出して，イベントとして言語化しなければならない．また，競技映像に常時スーパーインポーズされている対戦スコアと同等の情報を提供するため，対戦の説明や競技者の名前，対戦スコアなどは，イベントの有無に関わらず繰り返し伝達しなければならない．さらに，解説音声を再生している間に，複数回のイベントが受信されたり，状態が複数回更新される場合もある．そのため，伝達済みの情報を管理した上で，未伝達の情報から必要な事項を適切に選択して伝達しなければならない.

このような解説音声を生成するために用意した解説生成部は，配信データを分類・解析する機能に加えて，状態を伝えるメッセージからイベントを抽出する機能と，伝達済みの競技状態と未伝達の状態を管理する機能，一定期間伝達されていない状態を再伝達する機能を有する.

配信されるメッセージは，競技ごとに定められており [5]，一部の競技ではイベントメッセージが配信されず，状態メッセージだけが配信されるものもある．この検証実験では，それぞれのメッセージを受信した際に生成する解説文と，一定期間ごとに生成する解説文のテンプレートを競技ごとに人手で作成しておき，選手名やスコア，プレーの名前などを，受信メッセージにしたがって置き換えて解説文を自動生成した.

4.3 解説音声の合成

上記解説文から合成音を生成して競技映像に重畳して，解説音声とする．音声合成にはエーアイ社の合成器 [7] を利用した．さらに，短時間に競技の情報をもれなく伝えるために，話速を速くしても明瞭度が低下しない話速変換 [8] を用いて約 1.5 倍速の音声を生成した.

音声を合成するにあたり解説文の各単語の読みを適切に付与する必要がある．国名や配信データで伝えられる競技のプレーの名前や競技用語には，事前に読みを与えることができるが，選手名は，人数が多く国によって読み方が異なるため，適切な読みを与えるのは容易ではない．配信データからは選手名のアルファベット表記が得られるため，このアルファベット表記から各言語での選手の読みを推定する技術（音訳技術）利用し [9]，選手名の読みを与えた．この読み推定は，機械翻訳技術を応用したもので，選手名のアルファベット表記と選手の国籍を入力として，日本語の読みを出力する再帰型ニューラルネットワークを学習して実現している．この推定により，“Peter” という選手に対して，国籍がアイルランドの場合には「ピーター」を，ハンガリーの選手には「ペーテル」という読みを与えることができる．本実験では，事前に人手で読みを与えることができなかった選手名の読みをこの技術で推定した..

4.4 解説音声の重畳

国際映像から競技区間を切り出して，解説文の字幕と解説音声を付与して，評価用のコンテンツを制作した．字幕と解説音声の重畳には，電波産業会が標準化した字幕フォーマット arib-ttml を用いた．この規格は，次世代の放送用字幕として標準化された字幕フォーマットであり，通常の字幕ファイル形式 ttml を拡張して，音声ファイルによる解説音声も重畳できるフォーマットである．動画の再エンコードが不要で，インターネット配信動画にも放送用途にも利用可能である．本検証実験では，競技映像の mp4，字幕の重畳と解説音声の重畳タイミングを制御する ttml，解説音声の mp3 と，これらを表示するための html を競技ごとに配信した.

表 2 解説音声を自動生成した競技

競技名	試合数	競技名	試合数
柔道	414	バレーボール	60
レスリング	273	サッカー	56
競泳	196	フェンシング	46
テニス	124	カヌー	16
卓球	112	飛び込み	12
アーチェリー	91	陸上	10
ビーチバレー	86	シンクロ	5
バスケットボール	60	トランポリン	4
ハンドボール	60		

バレーボール女子予選リーググループ A。
日本対アルゼンチン。
マラカナンジーニョで行われます。
第 1 セット開始。第 1 セット、現在、0 対 0。同点。

....

○第 1 セット、現在、2 対 1。日本リード。

△アルゼンチンのタニアアコスタのサーブ。
△日本の石井優希がスパイク成功。日本のポイント。
△アルゼンチンのカスティグリオネがレシーブ失敗。

○第 1 セット、現在、2 対 2。同点。

△日本の荒木絵里香のサーブ。
△日本の長岡がスパイク失敗。
△アルゼンチンのエミルセソサがブロック成功。
△アルゼンチンのポイント。

図 5 自動生成されたバレーボールの解説の例

カヌー・スラローム・スタジアムで開催される、
カヌースラロームカナディアンペア男子準決勝。
競技チームを紹介します。第 1 走者ポーランド、
ピオトルシュテパインスキー選手とマルチンポチワラ選手のペア。
第 2 走者アメリカ、デビンマキュアン選手とケイシーエイクフェルド
選手のペア。

...

ロシアチームの競技。
...
第 5 ゲート。
第 6 ゲート。
第 6 ゲート、ゲート接触で 2 ポイントのペナルティ。第 7 ゲート。
第 8 ゲート。
第 8 ゲート、ゲート接触で 2 ポイントのペナルティ。第 9 ゲート。
第 10 ゲート。ロシア、第 1 中間地点のスコアは 37.38 ポイント。
4 ポイントのペナルティが加算されています。
現在のリーダーを、1.71 ポイント下回っています。

...

ロシアは、112.39 ポイントで、現在、第 2 位。
第 6 ゲートでゲート接触。2 ポイントのペナルティがありました。
第 8 ゲートでゲート接触。2 ポイントのペナルティがありました。
第 20 ゲートでゲート接触。2 ポイントのペナルティがありました。

図 6 自動生成されたカヌースラロームの解説の例

5. 実験結果

2016 年 8 月に開催された大規模スポーツイベントで，本検証システムを用いて解説音声の自動生成実験と評価実験を行った．実験では**表 2** に示す 17 競技 1,625 試合に解説音声が付与された．

5.1 生成された解説

図 5 は，バレーボールで生成された解説の例である．文頭に○を記した文が，状態を伝える文である．状態の変更タイミングや一定期間伝達されていない場合に繰り返し生成される文である．文頭に△を記した文が，イベントを伝える文である．この競技では得点が確定した後に，サーブ打者と得点につながったプレーのデータが配信される．これを国際映像に重畳した場合，得点に関わったプレーのスロー再生に合わせて，これらの文が再生されるタイミングになった．

図 6 は，カヌースラロームで生成された解説の例である．最初に競技出場選手の紹介があり，各選手の競技中には，通過ゲートと各ゲートでのペナルティ，中間地点でのスコアを伝達した．そして，各選手の競技後にその時点での暫定順位が伝えられた．

この競技は各ゲートのペナルティの値が更新されるだけで，イベントを伝えるメッセージは配信されない．各ゲートのペナルティが確定するのは，次のゲートに侵入する直前であるため，ペナルティ確定時に次のゲートを告げることにより，選手が通過を試みているゲートを適切なタイミングで伝えた．

ここに示した 2 例を含め今回の実験の対象となった競技では，十分な精度の時刻情報が配信データに付与されていることが確認された．

一方，この 2 例は競技の進行とともにテンポ良く解説音声が付与されているが，今回の実験の対象競技の中には，付与できる解説が少ないものもある．柔道がその例である．この競技では，試合の経過時間，技あり・警告などの有無と最後の決まり手が主な解説内容であった．しかし，経過時間のみが解説される時間帯が長かった．解説に十分な情報が配信されている競技があることが確認されたが，情報が少ない競技については，今後の評価実験を通して有効性を検証していく必要がある．

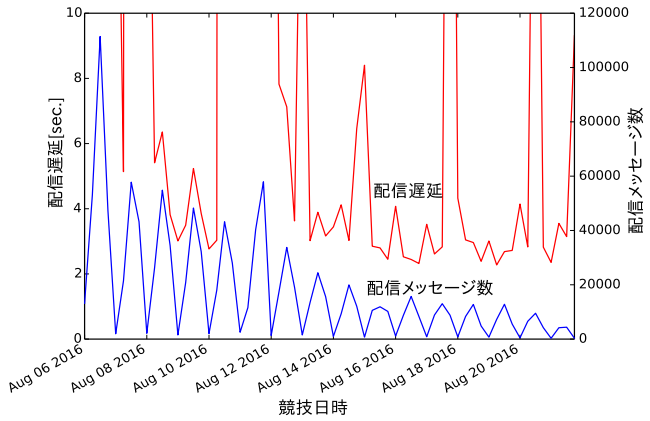

図 7 データ受信遅延

5.2 競技データ配信タイミング

上記の解説生成実験では，配信データに付与されている現地の時刻に基づいて解説音声を付与している．システムが有効に機能するためには，このメッセージを大きな遅延なく受信しなければならない．一般的に映像・音声の伝送には複数段の符号化と復号化に伴う遅延があるため，テキストデータの配信データの伝送遅延の方が短いと期待される．本実験で受信した国際映像の伝送遅延は 8.6 秒であった．配信データの伝送遅延がこの値より大きい場合には，解説文もこの遅延分を見込んで，過去形で表現したり解説するメッセージを選択するなどの工夫をしたテンプレートを用いて生成しなければならない．

図 7 は，スポーツイベント開催中の 6 時間ごとの配信データ数と平均のデータ受信遅延である．競技データの内容は開催ごとに充実しており，これまでにない数のデータが配信された．この配信量は，当初想定していたメッセージ数を大きく上回ったため，配信遅延が大きくなっている期間がある．この配信データは，本実験だけでなく NHK のデータ放送などでも利用していたため，8 月 11 日以降，データ放送で利用するメッセージに制限してデータを受信することとなった．そのため，配信遅延を正確に見積もることは今回できなかったが，条件が良ければ配信遅延は平均で 3 秒程度になることが確認された．また，今後同様のシステムを構築する際には，配信データの回線容量を十分に確保するとともに，データのリレー段数や配信方式を見なおさなければならないことがわかった．

6. おわりに

大規模なスポーツイベントを対象として, 映像なしでスポーツ中継を楽しむ人に向けた解説音声を自動生成するシステムの検証実験を報告した．リアルタイムに配信される競技データを利用することにより，人手では実現不可能な大量の競技に解説音声を付与できる事を示した．

今回の実験の主な検証項目は，リアルタイム配信データから得られる競技の進行情報の量と，リアルタイム配信データに付与されている時刻情報の精度，データの配信遅延の測定であった．カバレッジと時刻情報の精度については，概ね十分であると結論付けられたが，配信遅延については，配信に関わるトラブルのため十分に検証できなかった．今回のトラブルも含め，データ配信方法に今後の課題があると結論づけられる．

自動生成された解説の品質の評価は，スポーツ中継の演出という側面もあるため，自動生成されたコンテンツを様々な番組制作者に公開して意見を収集している．一方で，視覚障害者をはじめ解説音声を必要とする人々による評価実験も進めていく予定である．これまで番組制作者から寄せられた反響は，「想像をはるかに超えたもの．障害者との連携のある部署なので障害者にも見てもらいたい.」「興味を持っている．健常者へのサービスへの発展の可能性もある.」など，この技術の実現に向けた協力とともに，新たなサービスに繋がる技術としても期待されている．

参考文献

[1] 総務省：“平成 26 年度の字幕放送などの実績” (online), 入手先 ⟨http://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000126.html⟩ (2015.11.11)
[2] 加藤，清水：“番組台本を利用した解説放送用原稿作成支援システム”, FIT 2010, K-060, No. 3, pp. 753-754 (2010)
[3] 佐藤，宮崎，熊野，今井，山田：“スポーツ中継における音声ガイドの有効性の調査”, 音講論, 1-4-3 (2016 春)
[4] 今井，田高，尾上，清山，佐藤，宮，熊野，山田，岩城：“テレビ音声へのオーバーラップを許容した音声補助情報サービスの検討”, 信学総大, H-4-11, pp. 322 (2016)
[5] IOC“Olympic Data Feed” (online), 入手先 ⟨http://odf.olympictech.org⟩ (2016.9.4)
[6] Data Studium: “Data Studium” (online), 入手先 ⟨http://www.datastadium.co.jp⟩ (2016.9.4)
[7] 株式会社 AI: “音声合成エンジン AITalk とは？”, 入手先 ⟨http://www.ai-j.jp/about⟩ (2016.9.4)
[8] Imai, Tazawa, Takagi, Tanaka, and Ifukube: “A New Touchscreen Application to Retrieve Speech Information Efficiently” IEEE Transactions on Consumer Electronics, vol. 59, No. 1 (2013)
[9] 宮崎，熊野，今井：“国籍情報を用いた人名の音訳”, FIT 2016 (第 15 回情報科学技術フォーラム), E-018, pp. 145–146 (2016)
[10] 電波産業会: “デジタル放送におけるマルチメディア符号化方式” (online), 入手先 ⟨http://www.arib.or.jp/english/html/overview/doc/2-STD-B62v1_0-1p2.pdf⟩ (2016.9.4)