Vol.2019-MUS-123 No.7
Vol.2019-SLP-127 No.7
2019/6/23

放送技術と音信号処理
～人にやさしい放送サービスを目指して～

小森智康[†1]　今井篤[†1]

概要：NHK では番組制作・伝送・再生機器の研究と併せて，放送サービスを充実させる人にやさしい放送技術の研究開発を行っている．本報告では，音技術を応用した音声認識を利用した字幕技術，合成音による音声サービス，ダイアログ制御技術に関する研究を紹介する

キーワード：解説放送，音声認識，字幕，ダイアログ制御

Broadcasting Technology and Audio Signal Processing
-Aiming for Realization of the Human-Friendly Broadcasting Services-

TOMOYASU KOMORI[†1]　ATSUSHI IMAI[†1]

Abstract: NHK is conducting researches and developments on the program production, transmission, and playback equipment. These researches include human-friendly broadcasting technology. In this report, we show the audio technology examples that the captioning services using speech recognition, the broadcasting sound services using synthetic speech processing and the program dialog and sound object control services using audio coding technology.

Keywords: Commentary broadcasting, Speech recognition, Closed caption, Dialogue control

1. はじめに

NHK では，番組制作・伝送・再生のための機器の研究と併せて，人にやさしい放送サービスを実現するための技術研究を進めてきた．1952 年にはステレオ放送を開始し，テレビの音声多重放送などの研究が開始され，1978 年には 2 か国語放送などが開始された．この音声多重放送技術により，解説放送を実施した．その後，2000 年開始のデジタル放送では文字放送サービスなど，障害のある方をはじめ，お年寄り，外国人なども対象とした，誰もが放送を楽しめる放送サービスの実現を目指した研究開発が進められてきた．アナログ放送の時代には，特別な受信機を必要とした文字放送は，音声認識技術を利用したリアルタイム字幕制作システムの技術も併せることで，一部の生放送番組で字幕制作が進められてきた．一般のテレビで字幕放送サービスを表示できるようになったことで，聴覚障害者，耳が不自由な方をはじめ，一般視聴者にも利用される放送サービスとなった．視覚障害者，目が不自由な方のためには，テレビ画面に表示される地震や津波，ニュース速報などの字幕を合成音声で自動的に読み上げる音声放送サービスも開発した[1]．デジタル放送で開始したデータ放送では，情報内容を音声や点字に変換して伝える受信システムを開発した．他にも，放送音声をゆっくり聞きやすくする話速変換技術を開発，2002 年以降ラジオやテレビに導入し[2]，背景音と音声の音量調整に加齢による聴力変化を取り入れる研究を進めた[3]．以下 2 章では近年取り組んでいる AI を音技術に応用した放送サービスの事例を紹介し，3 章では 2018 年から開始された SHV 放送や次世代の放送に応用できるダイアログを制御するサービスについて紹介し，4 章で総括する．

2. AI の音技術への応用

人にやさしい放送サービス実現に向けて取り組む AI を音技術に応用した放送サービスの最近の事例として，音声認識を利用した字幕サービスと，合成音を利用した音声サービスを紹介する．

2.1 音声認識を利用した字幕サービスの拡充

全国向けの定時の短いニュース番組では，番組音声を直接認識し，その際に認識誤りをした単語の修正などを経て番組の字幕を制作している．会場の騒がしい環境で実況する相撲中継や複数の話者が自由に発話する情報番組では，アナウンサーが番組の音声を再度読み上げて音声認識をするリスピーク方式と呼ばれる音声認識を採用している．一方で，発話の少し不明瞭なインタビューなどの音声を認識する試みを進めており，取材してきた音声を字起こしするために音声認識する課題を設定し[4]，認識性能の向上を図っている．多くの番組で字幕サービスできるようになってきたこともあり，視聴者からは地域放送局発の番組を含め，さらに多くの番組への字幕付与が求められている．ただし，音声認識誤りを修正する人手やコストをかけられないこと

が，多くの地域放送局の課題である．

そこで，字幕サービスの拡充に向けて，音声認識結果が視聴者の番組理解をどれぐらい支援できるかを評価するため，2019 年には，認識結果をそのままインターネット配信するサービスの試行に着手した[5]. 図 1 にそのトライアルのイメージを示す．福島・静岡・熊本にある放送局からクラウド上にある音声認識サーバで入力音声を認識し，各家庭にある PC やタブレット端末で認識結果をそのまま表示する．

人名漢字をカナ表記することや，オープンキャプションを使用している英語やあまり明瞭でないインタビューの区間などを自動で判別して，「...」などのように字幕を表示しない手法を試みている．

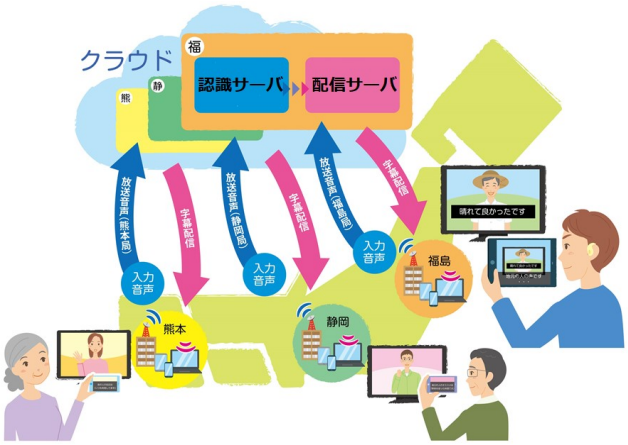

図 1　地域生字幕トライアルのイメージ

Figure 1　Trial image of complementing captioning services at local broadcasting stations

2.2 合成音による音声サービス

・自動解説放送

視覚に障害がある方にもテレビのスポーツ中継を楽しんでもらうために，生放送番組にも対応可能な自動解説放送サービス(図 2)の研究に取り組んでいる．生放送番組への適時の発話はアナウンサーでも難しいとされており，解説放送のサービス拡充が進んでいない現状がある．そこで，リアルタイムに配信される競技関連データを用いて，自動で試合の状況を説明するサービスの可能性について検討を行った．

国際的なスポーツ大会では，競技イベントに対応したデータ（いつ，誰が，何をした，など）を，インターネットに逐次配信するサービスが行われている．まず，このデータを用いて，競技イベントを説明する発話を自動生成する技術を開発した[6]．この技術により，同大会の 17 種目・1625 試合の動画に自動で実況のような発話を付与し，データの受信タイミングや情報の粒度などを検証した[7]．2018 年の国際スポーツ大会では，同技術を「ロボット実況」と命名し，独立したサービスとして実用化した．図 3 に同手法で自動生成された発話例を示す．

この「ロボット実況」技術を，自動解説放送の発話に活かす研究を進めている．図 3 の発話は実況調に組み立てられているが，解説放送用途には，「日本シュート」，「ゴール」のように，端的なスタイルで発話させる必要がある．さらに，放送音声との被りなどの問題もあり（現行の解説放送には基本的に音声の被りはない），好ましい情報の提示タイミングや提示方法，２つの音声の聞き取りやすさについても検討を進めている[8].

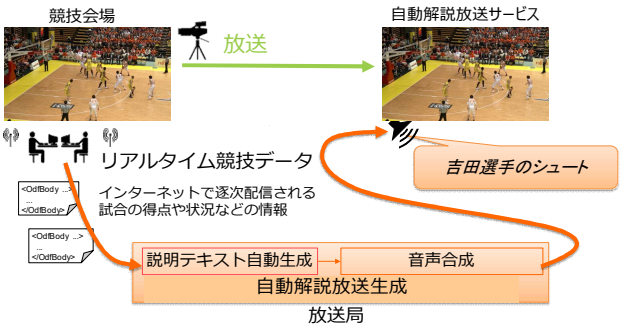

図 2　自動解説放送技術によるサービスの概要

Figure 2　Outline service using the automatic commentary broadcast technology

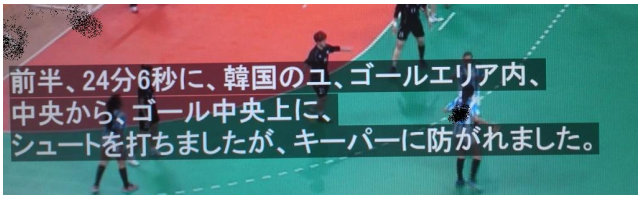

図 3　自動生成した発話例

Figure 3　Example of automatically generated comment

・「AI アナウンサー」

地域放送局のラジオ気象情報の一部を音声合成に担わせることを目的として，アナウンサー品質の発話を実現するシステムを開発した．気象庁から配信される気象データから自動で原稿を作成し，アナウンサーの話し方を学習した合成音声で発話する技術である．アナウンサーは，気象データから伝える内容の優先順位を考え，放送時間に収まるように原稿を構成しているが，この作業も自動化した．

2019 年 3 月に甲府放送局のラジオ県域放送でのテスト放送を実施した[9]. 図 4 に気象情報のための音声合成システムの構成を示す．

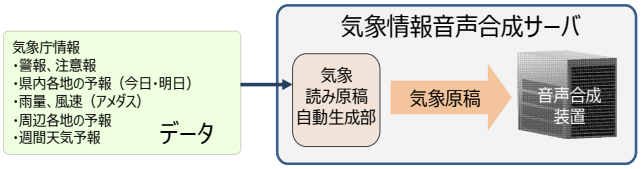

図 4　気象情報のための音声合成システムの構成

Figure 4　System configuration of speech synthesis system for weather information

3. SHV 放送や次世代放送向けの音響サービス

HDTV の 16 倍の画素数をもつ 7680×4320 の 8K 映像と 22.2 マルチチャンネル音響（以下 22.2ch 音響）による 8K スーパーハイビジョン（以下 SHV）は，その場にいるような高臨場感を実現できるメディアである．日本では 2016 年に試験放送を開始，国のロードマップに基づいて，2018 年 12 月に本放送を開始，2020 年の本格普及を目指している．NHK は対応する研究開発および設備整備や標準化活動を進めている．また，22.2ch 音響は上層・中層・下層の 3 層のレイヤーに合計 24 チャンネルを配置する（図 5）3 次元立体音響方式で，番組制作・伝送・再生に至る一連の技術開発を進めている[10] [11] [12].

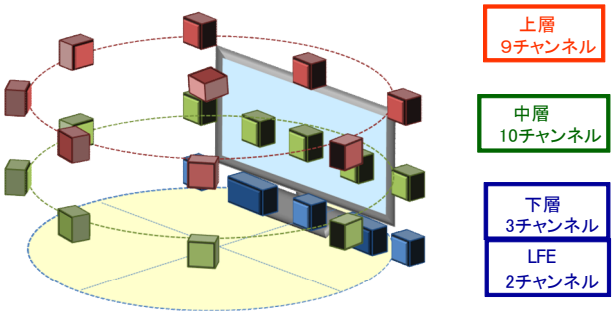

図 5　22.2ch 音響のスピーカー配置

Figure 5　Speaker layouts of 22.2ch sound systems

3.1 ダイアログを適切に制御するサービス

SHV 放送では，将来の放送サービス拡張のためにダイアログレベル調整およびダイアログの差替え機能が規格化され，22.2ch 音響の臨場感を保ちながらナレーションの音量を聞き取りやすくするレベル調整や，多言語の音声への切り替えを可能とした[13]．図 6 にダイアログ制御を使った放送サービスのイメージを示す．次世代の放送方式を想定した音声をオブジェクト化して符号化する方式についても研究を進めており[14]，ダイアログ以外の番組音声信号も含めたレベル調整や，受信側で設置したスピーカーの数や配置にあわせて，各チャンネルのスピーカーに信号を再配分するレンダリング[15]などにより再生する技術についての検討も進めている．こうした技術に加え，ダイアログと背景音楽，あるいは実況と解説などの異なる音声をスピーカー配置や発話タイミングを考慮して適切に再生することで，聞きとりやすくする研究[8] [16]も行っている．

4. 今後の課題とまとめ

解説放送や文字多重放送から開始された NHK の音技術を応用した“人にやさしい放送サービス”を目指した研究事例として，音声認識を利用した字幕技術，合成音による音声サービス，ダイアログ制御により番組音声を聞きやすくするサービスについての技術などを紹介した．こうした研究の実用化に向けては，多少時間のかかる技術もあるが，研究開発や実用化への取り組みにより，あまねく視聴者の方達に役立つ放送サービスの実現を目指していく．

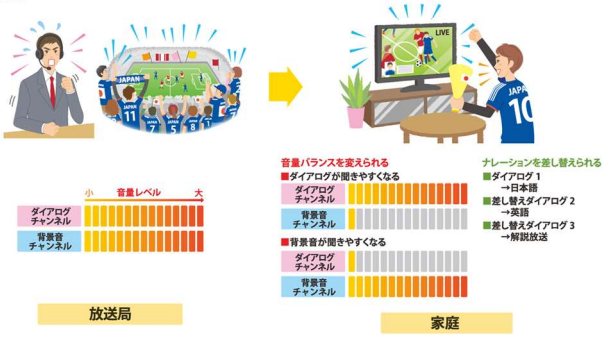

図 6　ダイアログ制御を使った放送サービスのイメージ

Figure 6　Image of broadcasting service using dialog control

参考文献

[1] 世木寛之，田高礼子，清山信正，都木徹，有森英明，松村欣司，清水俊宏.視覚障害者向け地震・津波緊急文字スーパーの自動読み上げ方式に関する一検討. 映情学冬季大, 2007, 2-2.

[2] 都木徹，今井篤，清山信正，世木寛之，田高礼子，田澤直幸，岩鼻幸男. 話速変換技術・音声変換技術の放送および関連ビジネスへの応用. 音声言語情報処理 (SLP), .2012, 1-6.

[3] 小森智康，今井篤，清山信正，田高礼子，都木徹，及川靖広. 高齢者に聞きやすい番組背景音レベル調整装置. 信学論 D, 2016, vol.99, no.9, p. 940-949.

[4] 三島剛，一木麻乃，萩原愛子，伊藤均，,小早川健，佐藤庄衛. 音声認識によるリアルタイム書き起こしシステムの開発. 映情学技報, 2018, 21D-4.

[5] 髙木康博. 音声認識技術とセカンドスクリーンを利用した字幕サービスの試み. 音講論(春), 2019, 1-3-12.

[6] 山田一郎，熊野正，佐藤庄衛，宮﨑太郎，今井篤，清山信正. オリンピックの競技状況を解説する音声ガイド自動生成. 映情学誌. 2017, Vol. 71， No. 1， p. 55-56.

[7] Kurihara, K. et al.. Automatic Generation of Audio Descriptions for Sports Programs. SMPTE J., 2019, p. 41-47.

[8] 一木麻乃，清水俊宏，今井篤，都木徹. スポーツ中継番組における自動解説音声の挿入タイミング決定法. 音講論(春)，2019, 2-P-39.

[9] “AI アナウンサーで気象情報を自動音声化，” http://www.nhk.or.jp/pr/keiei/shiryou/soukyoku/2019/01/004.pdf

[10] 佐々木陽，西口敏行，小野一穂. Development of multichannel single-unit microphone using shotgun microphone array. Proc. 22nd International Congress on Acoustics, 2016, ICA2016-0155.

[11] 杉本岳大，中山靖茂，大出訓史. 放送品質を満たす 22.2ch 音声信号のビットレート. 映情学技報, 2014, vol.38, no.35, BCT2014-76, p.17-20.

[12] 松井健太郎，伊藤敦郎，服部永雄，末永健明，岩内謙一. ラインアレイスピーカを用いた 22.2 マルチチャネル音響のトランスオーラル再生システムの開発. 信学技報, 2018, vol.118, no.234, EA2018-55, p. 7-12.

[13] 杉本岳大，中山靖茂，小森智康，知念徹，畠中光行. Dialogue Channel Control for 22.2 Multichannel Sound Broadcasting. J. Audio Eng. SOC., 2017, vol.65, no.6, p.507-516.

[14] 大出訓史. 放送における高臨場感オーディオの標準化の動向～8K SHV・22.2ch 音響からオブジェクトベース音響まで. 音響学音楽音響研資,2018, vol.37, no.5, MA2018-35, p.19-24.

[15] V. pulkki. Virtual Sound Source Positioning Using Vector Base Amplitude Panning. J. Audio Eng. SOC., 1997, vol.45, no.6, p. 456-466.

[16] 小森智康，都木徹，及川靖広. 空間的なマスキングリリースを利用した高齢者にも聞きとりやすい音響再生方法の検討. 映情学誌, 2017, vol.71, no.5, p. J172-J178.