Vol.2018-SLP-124 No.5
2018/10/10

番組制作支援のための音声認識を用いた書き起こしシステム

萩原 愛子[1,a)]　伊藤 均[1]　小早川 健[1]　三島 剛[1]　佐藤 庄衛[1]

概要：放送局などの報道機関は番組制作業務において，事実確認や編集のために取材映像の音声を文字に書き起こしている．近年，取材機器や映像伝送技術の進展により，より多くの取材映像が得られるようになった．この書き起こし作業の負担増加は，多くの報道機関にとって課題となっている．NHK は音声認識を用いることで，書き起こしを高速化・省力化する技術の研究開発に取り組んでおり，開発技術を実験的に番組制作に利用している．本稿では，書き起こし制作支援のための音声認識精度の向上に加えて，番組制作のワークフローを改善するために開発した，音声認識インターフェースを報告する．

1. はじめに

放送局をはじめとする報道機関は，正確な情報をいち早く視聴者に届けることを使命としている．この情報の源となるのが取材映像であり，番組制作者は長時間にわたる取材映像を適切に編集して，わかりやすく伝えている．この番組制作の過程では，取材した映像や音声素材の発話内容を文字化した書き起こしが不可欠である．一方，この書き起こし作業は多くの労力を要し，番組制作の効率性や迅速性の妨げになっている．さらに，要人による討論番組など生放送番組や記者会見など生伝送素材は，直後のニュース番組などで取り上げられるケースもあり，書き起こし制作に時間をかけられないため，人海戦術で書き起こすなどより多くの労力が求められている．本稿では，音声認識を用いて書き起こしを支援し，番組制作の効率を改善するシステムを新たに開発したので報告する．

2. 放送局における書き起こし制作の背景

取材によって得られる映像メディアは，時間方向にリニアなメディアであり，一覧性に欠けることに加え，内容の比較が容易ではない．そのため，次の 2 つの課題を解決するために書き起こしが必須となる．

- 記者やディレクターといった番組制作者が取材内容を正確に把握し，その理解に誤りがないか複数人で情報を共有して確認するため
- 放送で利用するために切り貼りして編集された結果が，元々の内容を正確に伝えていることを確認するため

さらに，ジャーナリストである番組制作者は，放送前の内容漏防止やインタビューを受ける人のプライバシー保護，取材源の秘匿に十分に配慮しなければならない．そのため取材を担当した番組制作者が自ら書き起こし作業を行うことが多く，これが番組制作時間の増加と常習的な長時間労働の一因にもなっている．

また，放送用の取材機器の特殊性により，番組制作者が書き起こしを制作するにしても，取材内容を参照できる環境が限定されていることも課題になっている．取材に用いるカメラは XDCAM など専門機材であり，メディア自体に加えて映像ファイル形式も特殊であるため，汎用的なパソコンや DVD プレーヤーでは編集どころか再生もできない．書き起こしを制作するためには，高価な専門の再生・編集機材が必要になるが，放送局といえどこれらの機材は需要に対して数が少ない．他の利用者の空きを待つということが多発している．こうした事情から，書き起こし作業に時間をとられ，本来番組制作者が十分議論すべき番組構成の議論や確認作業への時間が圧迫されているという現状がある．

筆者らが開発した音声認識を用いた書き起こしシステムは高精度な認識システムと，番組制作のワークフローに沿って利便性を追求したインターフェースによって，これらの課題を網羅的に解決した．

3. 番組制作に用いる音源への音声認識

前章で述べたとおり，書き起こしは番組制作過程の中で必須でありながら負担の大きい作業である．この省力化のため，音声認識の活用が番組制作者から強く求められている．

これまで NHK は，生放送番組にリアルタイムで字幕を

[1] NHK 放送技術研究所
NHK STRL, 1-10-11 Kinuta, Setagaya-ku, Tokyo 157–8510, Japan
[a)] hagiwara.a-iy@nhk.or.jp

付与することを目的として音声認識技術の研究開発を進めてきた．定時ニュース番組では認識単語誤り率が 3%程度の認識精度を実現している．その理由は，アナウンサーの発話が明瞭であることに加え，周囲の環境音などの雑音が少なく，発話者の音声を適切なマイクで集音されているためである．

字幕付与のための音声認識は，より難易度の高い情報番組の認識にも挑戦してきた．屋外の周辺音やスタジオの BGM などが混在し，発話者も一般人の方へのインタビューや対話といった砕けた発話が含まれる番組の音声を認識する技術の開発である．これらの課題を解決するため，筆者らは Bi-directional long short-term memory 構造の Deep Nural Network を設計し，これを 4500 時間分の音声で学習し，従来十分な認識精度がえられなかった環境でも，頑健に認識できるシステムを構築した．この 4500 時間の学習データは，過去に放送された NHK の番組や記者会見の膨大の音声と放送番組に付与された字幕など必ずしも一字一句書き起こされていない字幕を対応づけ，ディープラーニングに適するように，適応率を考慮して選別されたデータである．この音響モデルにより，雑音や砕けた発話による対話も含む情報番組の音声認識単語誤り率は 11% と，十分実用可能な水準に達した．

この認識対象範囲の拡大により，番組制作者からは番組制作の過程で用いられるインタビューや会見などの取材映像の音声の認識への期待が高まった．インタビューや会見は情報番組以上に認識の難易度が高い．一般人へのインタビューではマイクと話者の距離が一定ではないため S/N が悪く，その上マイク自体も IC レコーダーのように簡易的な場合も含まれている．こうした厳しい集音条件でも精度よく認識する技術は，今後も研究課題として取り組んで行くが，現時点でも，20%程度の音声認識単語誤り率が得られる取材映像は少なくない．この程度の認識誤り率であれば，十分に書き起こし作業を支援することが可能である．書き起こし作業を人手で行う場合は，音声を全て人手で聞いて逐一文字を記していく必要がある．これに対して音声認識を用いた場合，認識文字列の中から誤り部分を修正するだけで書き起こしが完成するため，大幅な省力化に繋がる．

4. 書き起こしのワークフローを考慮したインターフェースの開発

番組制作者が書き起こし作業に音声認識システムを利用するためには，十分な認識精度に加えて，ワークフローに適したインターフェースが必要である．筆者らは番組制作者の書き起こしにおける作業工程を精査し，容易かつ高速な作業を可能とするインターフェースを開発した [1][2]．図 1 はインターフェースの基本操作画面である．

このインターフェースの開発には，次の 4 つの番組制作のワークフローを考慮した．

図 1 インターフェース操作画面

4.1 着目コメントへの容易なアクセス

番組制作者は取材映像の音声を一字一句すべてを書き起こしてはいない．実際に取材に立ち会っている時点で，取材内容の概要は把握できているため，編集や事実関係検証のために重点的に書き起こすべき部分を理解している．そこでインターフェースの開発にあたっては，素早く目的の取材コメントにアクセスして再生できるように設計した．取材者が目的のコメントの時刻を把握できていれば，容易に実現可能であるが，多くの場合，文脈上の流れで記憶されていることが多く，「あの話題の最後の方」程度の曖昧な記憶しかない．そこで，目的のコメントに容易にアクセスできるように，取材映像を自動的に項目に分割し，それぞれの項目を特徴付けるキーワードを自動付与した．

この項目の分割には映像処理技術を導入して映像のカット点などを検出し，音声を認識した結果から長時間発話がなかった部分を検出して，項目の分割点を定めた．これにより，長時間にわたる取材映像を，目的のコメントが含まれる適切な長さの項目に分割することができる．

分割された項目に対しては，項目内の発話を特徴づけるキーワードを付与した．話題の推移を把握するのに十分なキーワードを得るため，対象となった取材映像の音声を認識して得られた単語の名詞を対象とし，項目ごとに TF-IDF を求めて，項目内の上位 3 単語をキーワードとして付与した．対象とする項目内における出現頻度が高く，なおかつ他の項目内における出現頻度が低い単語がその項目の特徴を表すキーワードになっている．また，それぞれの項目の映像の先頭からサムネイル画像を選び出した．無音区間を基準に項目が分割されることの多い取材映像では，概ね発話の開始部分をサムネイル画像として選び出すことができる．

提案するインターフェースの初期画面では，このサムネイルとキーワードのみが提示されており，これらを頼りに目的のコメントが含まれる項目に効率よくアクセスすることができる．

4.2 動画再生・停止と文字編集操作の一体化

キーワードとサムネイルから目的のコメントが含まれる項目を特定した後，対応する区間の動画と認識結果の単語列を表示して，番組制作者の必要に応じて，ば認識誤り単語を修正する．この作業では，音声を聞き直すことと文字を編集するという異なる二つの作業を同時に行なわなければならない．聴取した音声を記憶できる量に限界があることに加え，文字を修正していく作業に要する時間は単語を発話するのに要する時間よりもはるかに長いため，ある程度の区切りで音声の再生と停止を繰り返しながら，文字を編集していくことになる．多くの音声認識インターフェースでは，動画・音声の再生／停止を行うボタンと，文字を編集するパネルが別々になっているため，これらの作業は煩雑になりやすい．画面中の再生／停止ボタンを用いる方法では，これらのボタンと編集文字へのカーソル移動のためにマウスを操作している時間が無視できない．

提案するインターフェースでは，映像の再生・停止動作と文字の編集を一体化させた．具体的には，インターフェースに表示される認識単語をクリックすると，その単語に対応する時刻から映像が再生され，カーソル移動などの認識誤り単語を修正するためのアクションによって映像の再生が停止するインターフェースを設計した．これにより，ユーザーは任意の箇所から自由に映像を再生し，文字列の編集をシームレスに行えるようになった．このインターフェースは認識処理の中で，認識単語列には時刻情報が紐付けられているため実装可能となった．加えて，キーボード上のショートカットコマンドも実装し，マウス操作を不要としたストレスのない高速な再生・停止・文字列編集が可能となった．

4.3 専用機器の占有が不要な書き起こし制作

前述のとおり，取材映像は XDCAM など放送用の特殊なメディアに収録されているため，書き起こし制作のために専用機器を長時間占有しなければならない課題があった．提案システムでは，PC に接続可能な簡易的な XDCAM ドライブを介して，動画ファイルをサーバーに転送して，サーバー側で動画ファイルの音声を認識する．音声認識結果を動画とともに提供するインターフェースは Web アプリケーションとして実装されており，書き起こし制作者は，動画ファイルのアップロード時だけ短時間専用機器を占有し，もっとも時間を要する試写と認識結果の修正には，自席の Web ブラウザがあれば良い．専用機を用いることなく，Web アプリケーションの動作する汎用的な PC で映像を再生できるという点は，専用のソフトをインストールしなければならないという利用者の心理的負担を軽減し，書き起こし作業という目的に限らず，番組制作の中で編集前の映像の確認などに役立っている．

放送用の XDCAM には，放送品質の高画質・高音質な映像と音声が記録されているが，書き起こしインターエースには，簡易な品質の映像があればよい．帯域に制限のあるイントラネットワークを介して映像ファイルをストレスなく伝送するため，映像伝送には，撮影時に撮影動画を簡易に確認するためにカメラ内で生成されるプロキシと呼ばれる低品質・高圧縮素材をサーバー伝送する．このプロキシ映像に収録されている音声のサンプリング周波数は 8kHz であり，筆者らの音声認識が学習してきたサンプリング周波数 16kHz の音声とは不整合がある．したがって，プロキシ映像に加えて，高品質素材ファイルから抜き出した音声を 16kHz に変換した音声も同時に伝送する．このように映像・音声ともに必要最低限の容量を高速で伝送することにより，アップロードに要する時間を短縮できたほか，サーバーの負荷も軽減された．

4.4 セキュリティ

現在実証実験として運用しているシステムであるが，取材源の秘匿など社会的責任を負う番組制作者が利用することを考慮したセキュリティを確保した．映像素材は本人だけがアクセスできるメールアドレスとともにアップロードされ，映像素材にユニークなハッシュ文字列を含んだ素材専用の URL がメールアドレスに通知される．この URL を介して，番組制作者は書き起こしインターフェースにアクセスするほか，認識誤りを修正した結果をメールで取得できる．よって，アップロードした本人のみが映像の確認・文字列の編集が可能となる．

今後，イントラ内に限定しない実運用にあたっては，さらなるセキュリティが必要になると思われ，セキュリティに関しては更なる改善が求められている．

5. オフラインのシステム利用状況

音声認識システムは入力音声を逐次確定していき形態素単位で認識仮説を出力する「リアルタイム方式」と，終点まで入力された音声を認識する「オフライン方式」に分けることができる．この 2 つを比較すると，一般的にはリアルタイム方式の方が音声を入力してから認識結果が表示されるまでの待ち時間が短いが，その分認識精度はオフライン方式の方が高い．

NHK 内での音声認識システムの現場利用への要望は非常に高く，オフライン方式のシステムを実験的に設置した．オフライン方式は字幕制作のような生放送中の番組内ではなく，取材から放送までの間の時間に利用される．具体的な利用方法としては，番組制作者が取材してきた複数

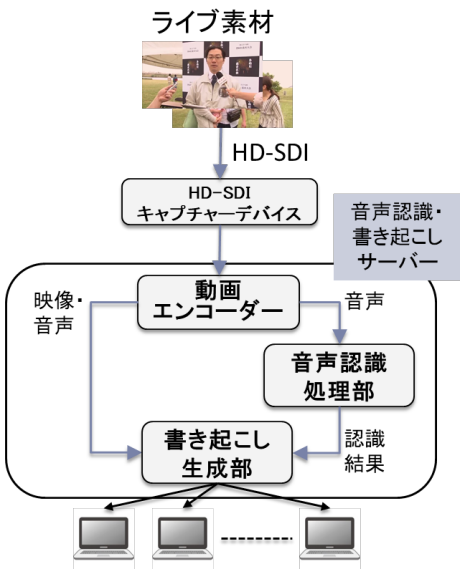

図 2 システム系統図

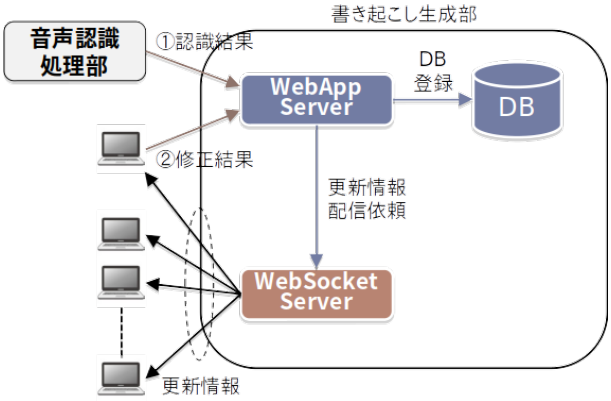

図 3 認識結果および修正結果の反映フロー

のインタビュー等の音源をまとめてシステムにアップロードし，認識が終了次第，番組構成も検討しながら必要な箇所だけ詳細な書き起こしを作成していく．

本システムはこれまでに多様な部署で利用されている．例えば，定時ニュースなどの報道部門，あさイチといった情報番組の制作部門，加えて全国各地のローカルな話題を中心とした番組を制作している地域放送局などである．2018 年 9 月現在，これまで約 1 万素材の利用があり，一日平均約 60 素材程度利用されている．

6. リアルタイムに対応したシステムの構築

NHK では国会中継や討論番組，緊急記者会見などでの発言内容を直後のニュース番組で放送するために，時間に制約がある中での書き起こしが日々求められている．放送現場ではこれに対応するため，番組毎に 10 名前後の担当者が書き起こし作業を分担している．しかし短時間での作業は作業負荷や書き起こしミスを増大させ，かつ発言内容を全て書き起こすことを困難にしている．また人的な連携は取っているが，書き起こし内容が重複するなど，書き起こし結果をまとめる際の弊害も発生している．

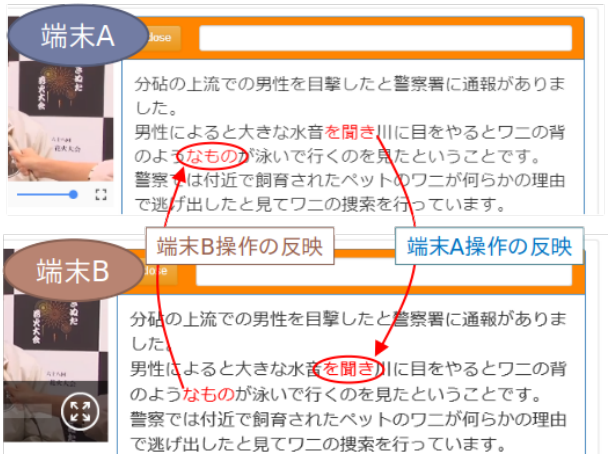

図 4 インターフェース上での修正状況の共有

これまで筆者らは作業効率を改善するために収録素材を対象とした書き起こし支援システムの研究・開発を進めてきた．これに加えて，既存システムの利点を活かしながらライブ素材の逐次書き起こしに最適なシステムを構築した [4]．リアルタイム利用に必要な要件を以下のように整理した．

(1) 直近の発話の認識結果をアプリケーション上に反映させ，逐次書き起こしを可能とする．
(2) 複数人での協調作業を可能にするため，修正状況の共有機能を持たせる．
(3) ライブ素材を収録しながら，認識誤りの確認・修正に必要な任意箇所の再生を可能とする．

上記要件の具体的な実現方法を下記に述べる．

リアルタイム書き起こしシステムの系統図を図 2 に示す．ライブ素材の入力には様々な形態が想定されるが，放送局での汎用性を考慮して HD-SDI 信号を用いることにした．この HD-SDI 信号を音声認識・書き起こしサーバーに取り込み，音声を認識して書き起こしコンテンツを生成する．書き起こし担当者は局内のイントラネットワークに接続された PC から Web ブラウザを介して書き起こしコンテンツにアクセスする．各要件への対応内容を以下に記す．

(1) ライブ素材の認識結果をリアルタイムに確認・修正するためには，認識結果が即座に作業者のアプリケーション画面上に提示されることが望ましい．筆者らは以前より生放送番組における字幕制作に音声認識技術を活用しており，この技術で使用している認識結果の逐次確定方式 [3] を本システムでも利用することにした．逐次確定された認識単語を複数作業者の端末に反映させる手段を図 3 に示す．「認識結果」の系統から形態素と時間情報を WebApp Server が受け取り，データベース (DB) に登録すると同時に WebSocket Server に形態素の配信を依頼する．追加された形態素は該当ページを表示している全ての端末にブロードキャスト

されるため，形態素を修アプリケーション画面上に同時に反映させることが可能になる．

(2) ライブ素材から迅速に書き起こしを作成するためには，複数人が協調して確認・修正できることが望ましい．そのため本システムでは各担当者の修正状況を全端末に逐次通知し確認箇所を明示することで，修正作業の競合を回避している．修正状況の明示は，修正箇所の文字色を変化させる事で実現している (図 4)．修正内容の通知は図 3 に示した「修正結果」を WebApp Server に送信する系統を使用し，以降は (1) と同様の情報反映を実施する．

(3) ライブ素材の場合，素材の終端が未確定であるため，単体の映像ファイルを再生する形態はとれない．任意の箇所を再生可能にするため，本システムでは HLS(HTTP Live Streaming) 方式での素材再生を採用した．Web ブラウザ上での再生には m3u8 ファイルを使用するが，修正インターフェースは認識結果を一定以上の音声ポーズ区間を境に複数の項目に自動に分割しているため，項目ごとに m3u8 を作成して再生を制御している．TS ファイルは 10 秒前後の比較的短いファイルで構成されているため，任意箇所の再生に加えリアルタイム付近の再生も可能である．

放送現場では定常的にライブ素材の書き起こしを実施しているため，本システムを実際の業務に活用したいという要望が多く寄せられている．まずは利用頻度が高い報道現場への導入を計画しており，実際の運用を通じて本システムの有用性や課題に関する情報を収集する予定である．

7. 利用実態と付加機能実装

NHK 内では本システムの実用化への要望が強く，リアルタイムに未対応のオフライン版であっても前述の通り多数利用されている．その実態を検証していくと，開発当初の想定になかった利用方法が発見された．それを踏まえて付加機能を実装し，2018 年 5 月に実施した技研公開において新機能を発表した [6]．

特徴的な利用実態には話者識別と映像要約の 2 点が挙げられる．

7.1 話者情報の付与

まず話者識別について説明する．ユーザーは文字列の編集において話者が切り替わる冒頭部分に取材を受ける人の名前や質問と回答を示す「Q/A」といった話者の情報を追記している例が多く見られた．こうした話者情報が番組を編集する上で目印になると推測される．そこで，自動話者識別の機能を実装した．解析した話者の特徴を色情報として画面上に表示するため，話者が切り替わると単語列の文字色が変化し，視認性が向上している．これは技術的には音声認識に用いる特徴量である i-vector を元に解析を行っている．

i-vector は話者の特性や周辺環境音といった特徴を強く表し，話者識別にも利用されている [7]．この i-vector が話者情報として扱い，これを文字色として表示する方法を提案した．大量の学習データから高次元の i-vector 空間を構築し，それを 1 次元へと圧縮する．この 1 次元の数値と色の情報を対応させることで，各 i-vector と対応色のデータベースを構築できる．未知の音声データが入力された際は，入力音声から作成された i-vector とデータベース上の i-vector との類似度を算出し，最も類似した i-vector と対応する色情報を画面上に表示している．現在，話者の大まかな傾向は分類できているが，個人の識別の機能はない．インタビューを受ける人と質問する記者を分離するといった識別に対する要望があるため，それは今後進めていく考えである．同時に，現在は文字色で示している話者情報の最適な表示方法についても検討していきたい．

7.2 映像要約

続いての特徴的な利用実態は，映像要約の記述である．自動で付与された各項目のキーワードは，表示だけでなく編集も可能である．このキーワード欄を，元のキーワードを全て削除し，項目の映像要約を記入している例があった．具体的には，「1shot」といった人物数の記述や，「料理ズーム」といった被写体やカメラ動作などの記述である．これは汎用的な PC からも映像を確認できるというメリットから，収録した内容の編集前のメモを記録していると考えられる．こうした実情を踏まえて，音声認識に限らず映像内容を自動で記述する機能を実装した．キーワード欄とは別に，項目のサムネイル付近に映像内容の記述が表示される．これは技術的には，画像認識技術の DenseCap[5] を用いてサムネイル画像を対象に認識結果を記述している．DenseCap の特徴としては，一文にまとめるのではなく画像の中心や背景など複数の領域ごとに記述する．よって具体的には「男の人が喋っています/窓がある部屋です」といった文が生成される．

生成された要約文の精度は悪くないが，一般物体認識の汎用的なデータで学習されたモデルであるため，番組制作者からの要望と若干の乖離がある．具体的には，「1shot/2shot」といったような被写体の人数や有名人の名称なども求められているため，今後は映像全体だけでなく，人物に注目した映像要約についても調査を進めて機能の拡充を図りたい．

こうした付加機能を実装することで，利用実態に合わせた利便性の追求を続けていく．

8. おわりに

書き起こし業務全般を支援することで，視聴者に正確な情報をより迅速に伝えることに寄与できると考えている．今後は書き起こしシステムの有用性を高めるとともに，書

き起こしに限定しない音声認識技術の展開によって放送業務全般の支援を進めていく．

今後の展開として最も重要な点は従来の番組制作工程との融合である．従来の番組制作工程の中では取材が終了した時点で放送用の素材を放送局の収録機へ伝送している．そして現在の実験システムの仕様上，書き起こしシステムを利用するためには別途音声認識サーバーへ収録素材をアップロードする必要があり，これは従来の工程から一つ手間を増加させてしまっている．そこで，放送局の収録機と書き起こしサーバーを連携させることで，取材終了時に伝送された時点で自動的に書き起こしも付加させることを考えている．音声認識による書き起こしは制作過程の中で可能な限り上流時点で付加されていることが望ましい．何故ならば，重要なニュースなどは同一の素材が異なる番組に繰り返し利用されることがあるためである．上流で認識単語列のデータを付与することで，同じ素材の繰り返し認識といった無駄を省略できる．

また，テレビに限らずラジオ番組への活用も進んでいる．ラジオ番組の聴き逃しへのフォローとして，放送終了後の番組音源をインターネットを介して一定期間再生できるサービスがある．これに加えて，放送の発話内容を文字に起こしそれを番組ホームページ等で公開するサービスも進んでいる．これまでに本システムは，気象災害報道の情報をとりまとめた web ページへ掲載する書き起こし作成などに活用されている．あらゆるデバイスや通信方式の発達に伴い，視聴形態が変化していく現代では，多面的な展開を進めていくことで番組接触率を向上させることが重要である．書き起こしを閲覧することで，番組の概要を把握することができる．短時間で把握できるというメリットだけでなく，インターネットを利用することで書き起こしのテキスト上からピンポイントで聴き逃しストリーミング配信へ誘導するといった発展も考えられる．従来のラジオ放送に加えて，長い時間の番組のうち自分の聴きたいところだけを手短に聴くといった視聴形態も広まっていく可能性がある．

このように放送現場とも連携できる強みを活かして，効率的な番組制作を支援していきたい．

参考文献

[1] 三島 剛, 他：取材映像の書き起こしインターフェースの開発, 2017 映情学年大 , 23D-3, 2017.
[2] 三島 剛, 他：音声認識技術による書き起こしインターフェースの検証実験 , 2017 映情学冬大, 12C-6, 2017.
[3] 佐藤 庄衛：音声認識を用いた生放送番組への字幕付与, メディア教育研究, 第 9 巻, 第 1 号, S9-S18, 2012.
[4] 三島 剛, 他：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, 2018 映情学年大, 21D-4, 2018.
[5] Johnson *et al.*:DenseCap: Fully Convolutional Localization Networks for Dense Captioning, Proc. CVPR, pp.4565-4574, 2016.
[6] https://www.nhk.or.jp/strl/open2018
[7] 小川 哲司, 他：i-vector を用いた話者認識, 日本音響学会誌, 第 70 巻, 第 6 号, pp.332-339, 2014.