音声認識を用いた生放送番組への字幕付与

Realtime captioning for live broadcast by using automatic speech recognition

佐藤 庄衛†

Shoei SATO†

† NHK 放送技術研究所　　† Science & Technology Research Labs, NHK

E-mail:　† satou.s-gu@nhk.or.jp

1. はじめに

字幕放送は,テレビの音声で伝えられている情報を文字で伝えるメディアであり，聴覚障害者やテレビの音を聞き取りにくいと感じている高齢者などのテレビ視聴を支援している．番組映像に部分的にスーパーインポーズされるオープンキャプションと比較すると，字幕放送は視聴者が必要に応じて表示／非表示を選択でき，番組音声のすべてが文字で表示される点が特長である．

公共放送である NHK は，あまねくすべての視聴者に情報を届けることを使命とし，できるだけ多くの番組に字幕を付与するように努めてきた．NHK 以外の放送局においても，総務省が平成 19 年に定めた「視聴覚障害者向け放送普及及び行政の指針」にある字幕放送，解説放送，手話放送の普及目標を目指して字幕放送の拡充を進めている．この指針は東日本大震災などを踏まえた平成 24 年の改正[1]を経て 10 年間継続した．その後，障害者の権利に関する条約，障害者基本法，障害を理由とする差別の解消の推進に関する法律などの外部動向を鑑みて，平成 30 年 2 月に「放送分野における情報アクセシビリティに関する指針」[2]が定められ，特に字幕放送に関しては，県域局にも目標が与えられた．また，この指針の策定にあたり障害者，放送事業者，メーカなどを構成員とした「視聴覚障害者等向け放送に関する研究会」では時代に即した情報保障のありかたが議論された．この研究会の報告書[3]では，インターネットやセカンドスクリーンなどの新たな技術の活用や，民間事業者などの第三者のサービスを取り込んだサービスの実績値などが言及され，放送の情報保障はその範囲だけでなくその形態の多様化も求められている．

NHK では，字幕放送を拡充するための技術として音声認識技術の研究開発を進め，2000 年には世界に先駆けて音声認識を用いて生放送のニュース番組に字幕を付与した．その後，様々な生放送番組の字幕を拡充するための音声認識技術と字幕付与技術が開発されてきた．本稿ではこれらの技術とその技術が解決してきた課題を解説する．

2. 生放送番組への字幕付与

NHK が総合テレビで放送する番組の概ね 1/4 は事前収録の番組であり，これらの番組の字幕は人手で時間をかけて制作されている．残りの番組は生放送番組である．字幕放送を拡充するにあたっては，生放送番組にどのようにして字幕を付与するかが問題となる．番組の音声を遅延なく文字に変換して字幕を制作しなければならないためである．

NHK では番組の特徴に応じて，人手で字幕を入力したり，音声認識を用いたりして生放送番組の字幕を制作している．これまで，次の 5 つの字幕付与方式が採用されている．

(1) パソコンなどに用いられている一般的なキーボードを利用するリレー方式．複数の入力者が短い発話単位の文字をリレー方式で入力する（図 1）．

(2) 複数のキーを同時に押下する特殊なキーボードを利用する高速キーボード方式．入力者と校正者のペアが数組で，短い発話単位をリレー方式で入力する．

(3) 音声認識しやすいように，字幕制作専用の話者が番組音声を復唱した音声を認識するリスピーク方式（図 2）．音声認識の誤りはオペレータが修正する（図 3）．

(4) アナウンサの発話など，高い認識精度が見込まれる部分では番組音声を直接認識し，それ以外の部分ではリスピーク方式を併用するハイブリッド方式．音声認識の誤りはオペレータが修正する．

(5) 番組音声を認識した結果と，番組を制作するために用意された原稿を比較して，原稿の読まれている部分を推定して，原稿を字幕とする方式（図 4）．

スポーツ中継の有無や季節ごとの番組編成により割合は変動するが，生放送番組の半分が(1)-(2)の方式で人手により字幕が付与されている．残りの半分が(3)-(5)の音声認識を利用して字幕が付与されている．音声認識を利用した番組のうち 8 割がリスピーク方式

図 1 一般キーボードを用いたリレー方式

図 2 番組の音声を復唱する字幕キャスター

図 3 認識誤りを修正するオペレータ

(3)による字幕であり，2 割が番組音声を直接認識する(4)の方式である．(5)の方式は地域局のみで運用されている．

人手で文字を入力する(1)-(2)の方式は，番組の話題や発話スタイル，背景雑音の有無による制約を受けることがないが，オペレータに熟練した専門技術を要し，リレー方式であるため多くのオペレータの確保しなければならないことが課題となる．一般のキーボードを用いる(1)の方式は，入力可能な文字の量に制限があるため，歌謡番組などの字幕の分量が少ない番組で利用されている．一方，高速キーボードを用いる(2)の方式は，入力可能な文字の分量の制約が小さいため，様々なジャンルの番組に字幕を付与できる．この方式のオペレータには，高度な専門性が必要であり，オペレータを育成している外部専門業者に字幕制作を委託している．災害時の緊急報道にも対応できるため，ニュースなどの報道番組の字幕付与に利用されている．

図 4 原稿推定インターフェース

(3)-(5)は音声認識を利用した字幕制作システムである．現在の技術では，100%の精度で音声を認識できないため，音声認識の誤りを修正する手段が必要になる．認識誤りをオペレータが修正するシステムを想定した場合，音声認識の精度はオペレータが修正可能な精度以上であることが求められる．例えば，ニュース番組の場合，アナウンサは平均 40 単語から成るニュース文を平均 12 秒で読み上げるため，単語の認識精度が 95%である場合，6 秒に 1 単語の認識誤りを修正することになる．仮に，認識精度が 90%に低下すると，オペレータは 3 秒に 1 単語の認識誤りを修正することになり，リアルタイムの字幕制作システムの構築は難しくなる．

この前提に基づいて，音声認識を用いた各種字幕制作方式は，番組の特徴に応じて使い分けられている．以降，音声認識を用いた各字幕制作システムの特長を述べる．

3. リスピーク方式

情報番組やスポーツ中継の字幕制作に用いられる方式である．これらの番組では，多様な話者が自由に発話するため，明瞭性が低い部分が多く番組音声を直接認識しても十分な認識精度が得られない．スポーツ中継を例にすると，実況アナウンサの感嘆詞や得点シーンでの興奮した口調部分では認識精度が著しく低下する．さらに，会場の騒音の混入が避けられない場合も多く，背景雑音にも起因して認識精度が低下する．

リスピーク方式では，静かなスタジオにいる字幕キャスターが番組音声を聞いて，それを音声認識しやす

図 5 認識誤り修正インターフェース

いように復唱した音声を認識する．復唱により上記の課題を回避して，必要とされる認識精度を確保するシステムである[4].

復唱音声の認識結果はオペレータにより確認され，認識誤りは即座に修正されて正確な字幕が制作される．図 5 は認識誤りの修正に用いられるインターフェースである．インターフェースの左下に逐次表示される認識結果から認識誤り単語をタッチパネルで選択して，キーボードを用いて簡単に修正することができる．また，誤り単語の同音異義語や修正履歴が適応的に提示されており，これらの単語を利用して効率良く修正することもできる．さらに，ニュース番組など，原稿を取得できる番組では原稿を参照して原稿内の単語も利用できるようになっている．このシステムでは復唱と認識誤りの修正に時間を要するため，番組音声から 5~10 秒程度遅れて字幕が表示される．

復唱の方法は,番組音声をそのまま文字にすることを基本としながらも，この時間遅れを考慮して復唱する言葉を選んでいる．スポーツ中継の場合には，画面に映し出される選手のプレーと字幕の内容に不一致があると競技の進行を理解することが難しくなるため，字幕キャスターは映像を見てわかることを省き，プレーを簡潔にまとめて言い換えることにより，字幕の表示遅れがあっても字幕番組を楽しめるように工夫している．さらに，拍手や歓声などの競技場の音の様子を言葉で補足するなど字幕ならではの演出もしている．一方，情報番組の場合には，複数の出演者の発話が交錯することもあり，このような場合には，字幕キャスターが交錯した発話を整理して視聴者にわかりやすい字幕を提供している．

音声認識では,認識対象となる単語やそれらの単語と単語のつながりやすさを与える言語モデルが用いられている．リスピーク方式の字幕制作システムでは，スポーツの競技ごとに適応化した言語モデルを用意して認識精度を高めている．一方，情報番組は放送回やコーナーごとに異なる幅広い話題を提供しているため，認識に必要な単語や言い回しが毎回異なってくる．字幕制作担当者は，放送前に入手可能な情報をもとに必要な単語や言い回しをカバーしたテキストを用意して，毎回このテキストから言語モデルを学習して認識している.放送前に入手可能な情報の代表的なものとして，番組構成表がある．番組構成表は，番組の大まかな流れが記載された表であり，字幕制作者はこの記載と予定されている番組出演者から，出演者がとり上げそうな話題に関するテキストを取集する．さらに，番組のリハーサル時の出演者の発言を書き起こして学習テキストを用意する．同様に字幕キャスターもリハーサル時にリスピークを試して，認識が難しい単語と言い回し，認識しやすい言い換え方法を確認して，生放送時の字幕制作に臨んでいる.

高速入力キーボード方式に次いで,適用可能番組が広いリスピーク方式であるが，字幕制作者の事前準備やスキルのある字幕キャスターの確保が課題となるため，コストをかけられる番組だけで利用されている.

4. ハイブリッド方式

定時に放送されている短いニュース番組や,地域局発のニュース番組では，運用コストの低い字幕制作システムが必要とされている．現在運用されている音声認識は，アナウンサの原稿読み上げ部分だけでなく，記者による現場からのリポートなどニュース番組の大部分で番組音声を直接認識して実用的な認識精度を確保できる．しかし，番組の一部には一般話者へのインタビューや会見など認識が難しい部分がある.そこで，番組音声を直接認識することを基本とし，認識が困難になる部分では，修正オペレータ自身によるリスピークを併用するのがハイブリッド方式である[5].

ハイブリッド方式は主にニュース番組で利用されており,記者が入稿する記事のデータベースと連携し，最新のニュースを認識するために必要な固有名詞や人名を自動的に学習して，必要とされる認識精度を確保している．刻一刻と変化する社会の状況に対応するため，放送中に入稿される原稿も学習して利用できるように，字幕制作システムを稼働させながら，新しい言語モデルに切り替える仕組みも備えている.

地域局発のニュース番組の一部のコーナーでは,地域ごとに特色のある演出がある．このような項目はニュースとは異なる話題を扱うため原稿データベースに登録がない場合がある．このような部分では認識精度が一時的に低下するため，図 5 に示した認識誤り修正インターフェースでは，音声認識の難易度に応じて修正者を増減できるように構築してある．端末同士で他の修正者の作業を確認できるため，複数人での共同作業が可能になり，このような認識精度の低下に対応することができる.

地域局発の情報番組では,事前収録映像の再生が多

くなる．このような事前収録映像の場合には，放送前に対応する原稿をオペレータが用意できる場合がある．この修正インターフェースは，事前に用意した原稿を番組の進行に合わせて手動で送出することもできる．

ハイブリッド方式は，明瞭性が低くて認識が難しくなる部分でリスピークすることを前提としているが，明瞭性が低い発話の多くは，放送時にオープンキャプションが付与されることが多く，実際にリスピークを要する部分は多くはない．

この字幕制作システムは，東京のほか，大阪，名古屋，福岡，仙台の地域の拠点となる放送局にも導入され，地域局発のニュースにも字幕を付与できるようになった．

5. 原稿推定方式

地域局発のニュース番組の字幕を拡充するに際しての課題の一つが，認識誤りを発見してそれを即座に修正する技能を有するオペレータの確保であった．この即座に素早く行わなければいけない作業を軽減し，だれもが余裕をもって行える作業に置き換えたシステムが原稿推定方式である[6]．

ニュース番組の多くは事前に読み原稿が用意されているものが多いものの，東京発のニュースでは放送直前や放送中に原稿が修正されるため，音声認識結果を修正して字幕を制作していた．一方，地域局発のニュースでは，直前や放送中の原稿の修正は多くはない．そこで，このシステムでは，番組音声を認識した結果からどの原稿のどの部分が読まれているかを推定して，対応する原稿を字幕として送出する．これにより，迅速さが求められる認識誤りの修正作業は，時間的な余裕がある中で原稿を修正する作業に置き換えられた．

本システムの概要を図 6 に示す．地域局発のニュース番組に字幕を付与する場合，このシステムは次の条件のもとで番組音声の認識結果から読まれている原稿を推定できれば良い．

(1) どの読み原稿がどの順番で読まれるかは事前に特定できない．
(2) 放送音声は，読み原稿に基づいているものの読み飛ばしや言い換えがある．
(3) 読み原稿が用意されていない音声もあるが，この部分に間違った字幕を付与しない．
(4) ニュース番組の場合，音声認識結果には 5%程度の単語に認識誤りがある．
(5) 放送中に字幕制作アルゴリズムを停止することなく送出する原稿を修正できる．

ニュースで取り上げられる予定の項目のリストは事前に得られ，それらの原稿は各地の記者が入稿してくるさまざまな時事に関する原稿のデータベースから参照できる．しかし，これらのニュース項目は時事の

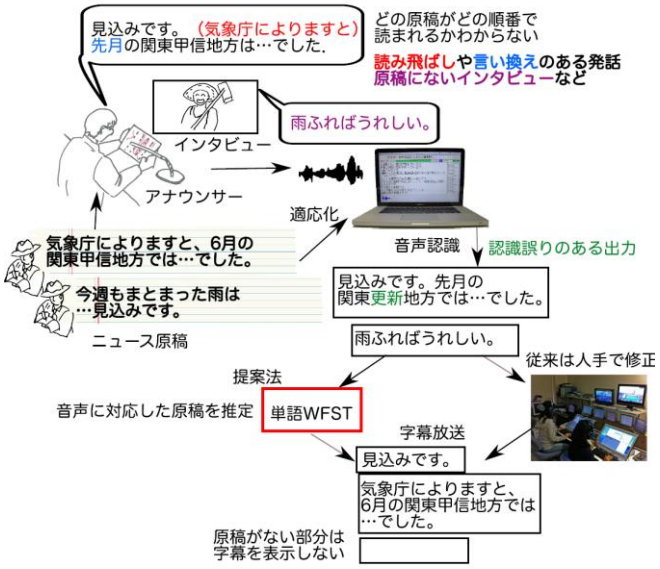

図 6 原稿推定方式の概要

推移に応じて放送中でも変更される場合がある(1)．図 6 の例では，原稿の順番が「気象庁に…」の次に「今週も…」であるが，実際には「今週も…」の次に「気象庁に…」が読まれている．さらに，原稿は読みやすいように修正されたり，番組時間内に読み終えられるように読み替えられたり，読み飛ばされたりする(2)．図の例では，「気象庁によりますと」「先月」がそれぞれ読み飛ばされた部分と言い換えられた部分である．また，事前に収録されたインタビュー映像を再生する部分などでは，対応する原稿がない場合がある(3)．図の例では「雨が降ればうれしい」が原稿の無い部分である．そして，図の「更新地方」が音声認識の誤りの例である(4)．最後に，送出する原稿を最新の情報に保つために送出用の原稿を修正できるように構成しなければならない(5)．これにより，災害時の被災者数など，放送中にも新たな情報が入ってくる項目は，読み上げる時点の情報に更新して字幕を送出できるようになる．

認識誤りを人手で修正する前述の方式では，認識誤りの修正にかかる時間分だけ遅れて，視聴者の画面に字幕が表示されており，字幕を必要とする方々からはこの遅れ時間の解消が強く求められていた．この原稿推定方式のために，原稿文をすべて読み終える前に精度よく出力字幕文を確定できるアルゴリズムを開発した．このアルゴリズムで作られる字幕は，ニュース原稿の読み始めは対応する字幕の表示が遅れるが，文の読み終わりの頃には字幕表示が読み上げ音声に追いつくため，字幕表示遅れに対する不満は半減する．

このシステムは，札幌，松山，広島の地域の拠点となる放送局に導入されている．前述のハイブリッド方式と合わせて，日本国内の拠点局発のニュース番組に

は字幕を付与できる体制が整っている.

6. 災害報道への字幕付与

冒頭に述べた通り，東日本大震災を受けて平成 24 年の「視聴覚障害者向け放送普及及び行政の指針」の見直しでは，「大規模災害時緊急放送時については，できる限りすべてに字幕を付与する.」「災害発生後速やかな対応ができるように字幕を付与する.」の 2 項目が新たな目標として加わった.

東日本大震災時 NHK は高速キーボードを用いた方式で可能な限り字幕を付与したが，災害報道が長期にわたったため，特殊技能を有するオペレータの確保が難しくなり，字幕を付与できない期間が生じた．このような部分に音声認識を用いたハイブリッド方式で字幕を付与できるようにするため，刻一刻と変わる災害状況の報道を認識するための言語モデルを開発した.

災害時には原稿のデータベースは十分に機能せず，最新の災害の情報は放送されている音声に中にあり，それが繰り返し伝えられるのが災害報道の特徴である．そこで，災害報道を認識した結果や高速キーボードで付与された字幕を定期的に学習することを繰り返す言語モデルを開発した[7][8]．図 7 は，東日本大震災の発災から 23:00 までの災害報道のアナウンサによる発話の誤認識率のシミュレーションである．提案法による情報収集方法「+速記字幕」「＋誤り修正」により，発災後 1 時間程度で認識精度を 94.3%に改善できる見込みを得た．この精度はシステムが目標とする認識精度 95%には届かなかったが，非常時であるという事と，同じ内容が繰り返し伝えられるという災害報道の特徴を考えると運用可能な認識精度であると考えている．また，5 時間を過ぎるとデータベースの原稿が役立つようになり，通常の運用に戻れると考えられる.

7. 地域放送字幕の拡充に向けて

平成 30 年に定められた「放送分野における情報アクセシビリティに関する指針」では，地域放送に字幕を付与してほしいという社会的要請を受けて，県域局に対しても目標が与えられた．県域局の字幕付与の難しさは，まさに生放送番組への字幕付与の難しさである．生放送に字幕を付与するための音声認識技術は，近年の深層学習の導入により認識精度が向上してきている．また，「視聴覚障害者向け放送に関する研究会」では，インターネットやセカンドスクリーンの活用も言及されている.

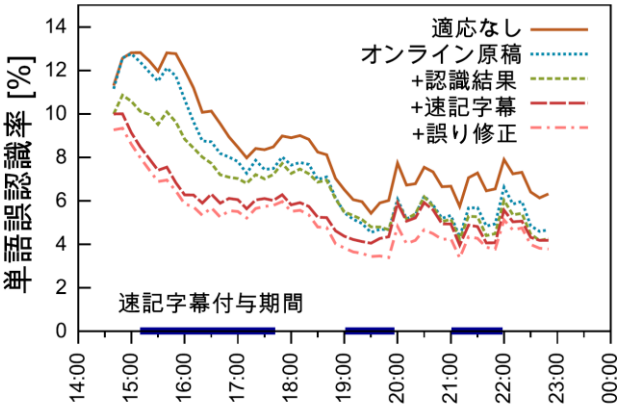

図 7 東日本大震災時の音声認識シミュレーション

NHK ではこれまで開発された音声認識字幕制作システムにより，字幕付与に要する要員を減らしてきたが，県域局を対象として全国規模での要員の確保や音声認識を含めた字幕制作設備の確保は極めて困難である.音声認識精度を向上させ,設備を集約したうえで,人手を介さずに字幕を提供できる仕組みが必要とされている.

NHK は，2019 年 2 月より，福島，静岡，熊本の 3 県で，自動音声認識字幕を配信するトライアルを開始した[9]．図 8 にこのトライアルの概要を示す．このトライアルでは，音声認識はクラウドに集約してあり，各地域局には放送コンテンツを伝送する簡易な機器があればよい．地域局発の番組音声を認識して得られた文字列はそのまま人手を介することなく，配信用 Web サーバーを介して，視聴者の PC やタブレット，スマホなどのセカンドスクリーンにリアルタイムで配信される.

このトライアルは,地域放送局の字幕提供の機会を増やすための取り組みであり．人手をかけられない状況下で，一定の誤りが含まれる音声認識結果を字幕として許容する社会環境の醸成を見定め，許容されるとの判断に至れば実サービスへの移行を検討するための実験である.

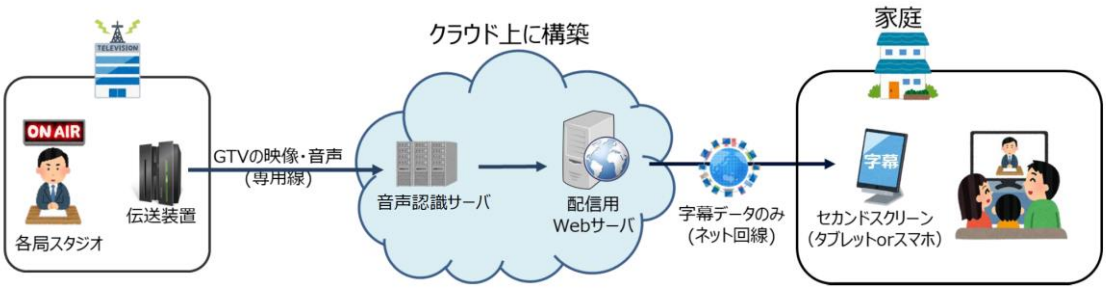

図 8 自動音声認識字幕配信トライアルの概要

トライアル対象番組は，平日夕方 6 時 10 分から 7 時に放送されている「はまなかあいづ TODAY」（福島），「たっぷり静岡」（静岡），「クマロク!」（熊本）である．実施期間は 2 月 4 日から 8 月 30 日の期間（点検・補修等の休止あり）である．この期間に，聴覚障害者および一般視聴者の反響を調査する．この調査により，

(1) 音声認識で地域局発の番組を認識してどの程度正確な文字が出せるか．
(2) 誤認識のある字幕を視聴者がどの程度許容するか．

が調べられる．

このトライアルには，認識結果をそのまま字幕とするために必要な機能が追加された音声認識が用いられている．その代表的な機能を 2 つ述べる．

一つ目は，人名の表記に関する工夫である．音声認識で用いる言語モデルは，東京発の番組に付与された字幕やニュースから学習されている．地域局で話題になる人名は学習されていないため，人名の表記誤りが多くなることが見込まれる．また，人名の表記誤りにより生じるニュースの誤った理解も避けたい．ここで用いられた音声認識では，認識結果の単語が人名であると判断された場合には，カタカナで人名を表記してこれらの課題を回避した．

二つ目は，認識困難部分での字幕の表記である．この音声認識は一般的な定時ニュースであれば 97%以上の認識精度が得られるが，地域局発の上述の番組では，地域の人々をとり上げて地域の人々の発話をそのまま放送することが多い．このような発話には不明瞭であったり，外国語が混在していたり，方言があったり認識が困難になるものがある．こうした理由で認識精度が低下した部分では，視聴者が認識結果の誤りを推定して元の発話を予測することは困難になるため，認識誤りによって番組内容を誤解する原因になる．また，このような発話では放送番組の映像にオープンキャプションが付与されており，字幕の情報が不要であることも多い．そこで，音声認識が候補単語の探索にどれぐらいの演算を要したかを自動的に調べて，認識が難しかった部分では字幕表示を停止して，画面には「。。。」を表示する．この機能により，本トライアルで提供される字幕を認識精度が高い部分だけに限定することができるため，誤認識も許容されやすくなると期待される．

8. おわりに

様々な環境下の視聴者すべてに情報を届けることが公共放送の使命であり，字幕放送はその手段の一つである．この字幕放送を拡充するため，NHK が実現してきた音声認識技術の現状と挑戦を，障害者向け放送の指針とともに紹介した．多くの方々から今なお字幕放送の拡充が求められており，その要望に応えるためのトライアルも今後継続していく．今後，トライアルで得られる調査結果をもとに，新たに与えられた字幕拡充の目標を早期に達成すべく今後も音声認識技術の研究開発を進めていく．

文　献

[1] 総務省，視聴覚障害者向け放送普及行政の指針見直しの概要，http://www.soumu.go.jp/main_content/000254131.pdf，(2012.10)（参照 2019-05-16）
[2] 総務省，放送分野における情報アクセシビリティに関する指針，http://www.soumu.go.jp/main_content/000531258.pdf，(2018.2)（参照 2019-05-16）
[3] 総務省，視聴覚障害者等向け放送に関する研究会報告書，http://www.soumu.go.jp/main_content/000531258.pdf，(2017.12)（参照 2019-05-16）
[4] 服部多栄子，椎名勉，堂免大規，“字幕放送サービス－システムとサービス概要－，” 映像情報メディア学会技報，BCT2004-24，28，5，pp. 17‐20 (2004).
[5] 今井亨，奥貴裕，小林彰夫，“音声認識によるリアルタイム字幕放送の進展，” 情処研報，SLP-88，4 (2011).
[6] 佐藤庄衛，尾上和穂，小林彰夫，奥貴裕，一木麻乃，荒井孝，“ローカル番組の字幕付与システムの開発，” 情処研報，SLP-103，1 (2014).
[7] 奥貴裕，藤田悠哉，小林彰夫，佐藤庄衛，“災害報道字幕制作のための言語モデル更新，” 信学技報，SP2013-48，pp. 101–106 (2013).
[8] 佐藤庄衛，小林彰夫，奥貴裕，藤田悠哉，“大規模災害等緊急時放送の字幕付与に向けた音声認識の改善，” 音講論集，1-9-15，pp. 129‐130 (2013.3).
[9] 高木康博，“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”，音講論集，1-3-12，(2019.3).