SEMINAR REPORT

字幕放送拡充のための音声認識技術の開発

NHK 放送技術研究所付
（一財）NHK エンジニアリングシステム出向
佐藤 庄衛　氏

ＳＣＡＴ優秀賞ということで、「字幕放送拡充のための音声認識技術の開発」を今日は御紹介いたします。安全・安心というのに何で字幕放送なんだとか、何で音声認識なんだとか、ちょっといろいろ不思議なこともあろうかと思いますけれども、今まで起こったストーリーも含めて聞いていただければ、そういうことなんだなと御理解いただけるのではないかと思っています。

まず最初に、NHKには放送技術研究所というのがありまして、御存じない方もたくさんいらっしゃるかと思いますので、紹介させていただきます。放送法の中で、放送とか受信が進歩するような研究しましょうねということが書かれていて、最初はラジオ放送のための研究を始めて、そのすぐ後にはテレビ放送をやろうなんていうような研究をしています。

研究所の規模としては、ちょっと古いデータですけど、200人そこそこでそんなに大きな研究所ではありません。やっていることは基礎デバイスみたいなあまり人目に触れないところから応用システムまで、いろいろなことをやっています。受信料で研究しているものですから、研究した成果は技術移転や標準化を行って、いろいろな方々に御利用いただいて、もちろん放送に役立つようにということをやっています。

字幕放送とは

その中で、字幕放送ということですけれども、一応、全ての受信機で字幕放送が見られるようになっているかと思います。お元気な方々はそんなリモコンのボタン押したことないということもあるので紹介すると、耳の聞こえない方に放送でしゃべっていることを全て文字で表示する機能です。クローズドキャプションともいいますが、字幕放送というのはボタンでオン・オフができるものです。対して画面にしょっちゅう出てきている文字はオープンキャプションといいます。

図１　字幕放送とは

この仕組み自体はデジタルになったからできるようになったというわけではなくて、アナログのときからありました。下に「おらへいきだー」とかいって、「おしん」で字幕がついているところを出してますけど、NHKが一番最初に字幕放送を出したのは、この「おしん」というドラマだったと聞いています。今でもワンセグだとか、音を出せないようなところで字幕をオンにして使われている方々もいるので、障がい者だけではなく利用されているというサービスです。(図1)

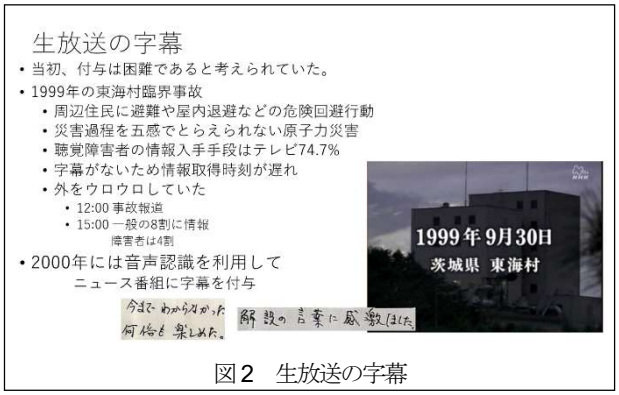


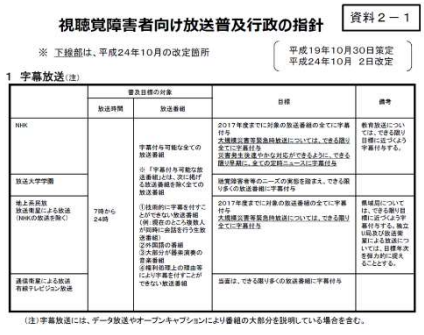
図２　生放送の字幕

そして、今日のお話で出てくる生放送の字幕というのが、安心・安全というところにつながってまいります。生放送に字幕つけるというのは、最初これは無理だよね、何しゃべるか分からないしとか思って、みんな「おしん」みたいなドラマから順番に字幕をつけていきました。

転機となったのが 1999 年の臨界事故のときでした。周辺住民の人たちに屋内退避だとか危険回避行動が勧められ、放送を通じてそれらを皆さんにお伝えしました。しかし、原子力ということで、熱いだとか目に見えたり感じられるものではないので、なかなか分かりにくいと。

当時、聴覚障がい者さんの情報入手手段というのは、74.7%はテレビだと言われています。それで回避行動や大事な情報を伝える字幕がないために情報が得られなくて、健常な人たちはちゃんと家でじっとしていたけれども、障がい者さんだけが外をうろうろしていたということがものすごく大きな問題になりました。

具体的に言うと、12 時には事故の報道があって、そのあと何回かニュースで放送していましたが、15 時の段階で一般の人たちの８割はそれが分かっていたけれど、障がい者さんは４割ぐらいしか分からなかった。これはだめだよねということで、2000 年には音声認識という技術を使って、全部ではないけれども、ニュース番組の一部では生放送にも字幕をつけましょうということを始めました。(図 2)


視聴覚障害者向け放送普及行政の指針　資料２－１

※ 下線部は、平成24年10月の改定箇所

平成19年10月30日策定
平成24年10月 2日改定

1 字幕放送（注）

普及目標の対象				
	放送時間	放送番組	目標	備考
NHK	7時から24時	字幕付与可能な全ての放送番組 ※「字幕付与可能な放送番組」とは、次に掲げる放送番組を除く全ての放送番組 ①技術的に字幕を付すことができない放送番組（例：現在のところ複数人が同時に会話を行う生放送番組） ②外国語の番組 ③大部分が募集演奏の音楽番組 ④権利処理上の理由等により字幕を付すことができない放送番組	2017年度までに対象の放送番組の全てに字幕付与 大規模災害等緊急時放送については、できる限り全てに字幕付与 災害発生後速やかな対応ができるように、できる限り早期に、全ての定時ニュースに字幕付与	教育放送については、できる限り目標に近づくよう字幕付与する。
放送大学学園			視聴障害者等のニーズの実態を踏まえ、できる限り多くの放送番組に字幕付与	
地上系民放 放送事業による放送（NHKの放送を除く）			2017年度までに対象の放送番組の全てに字幕付与 大規模災害等緊急時放送については、できる限り全てに字幕付与	県域局については、できる限り目標に近づけよう字幕付与する。独立U局及び放送事業による放送については、目標年次を各々が定めることとする。
通信事業による放送 有線テレビジョン放送			当面は、できる限り多くの放送番組に字幕付与	

（注）字幕放送には、データ放送やオープンキャプションにより番組の大部分を説明している場合を含む。

図３　視聴覚障害者向け放送普及行政の指針

こういったことを受けて、NHKだけではなくて、国のほうも動いてくれました。NHKも民放さんも含めて、何年にはこれだけ字幕をつけましょうねなんていう放送普及行政の指針を総務省さんが出してくれました。(図 3)

これもざっくり言うと、字幕付与可能な全ての番組に字幕つけなさいということで、大変なことではあるけれども、これを目指すためにやってきたNHKの取組みを紹介します。

生放送番組に字幕をつけるための課題

生放送番組に字幕つけるので生字幕と私たちは呼んでいます。

生字幕が何で難しいかというと、１つ目は事前に何をしゃべるか、文字を全然用意できない。２つ目は番組音声を遅延なく文字に変換して出さなければいけない。３つ目はあまり字幕が誤ると、それはよくないですねといったところです。当然のことではありますが、これを実際にやるのは大変です。

図４　NHK の生字幕制作手段

NHKでは音声認識だけではなくて、いろいろな方法で字幕をつけています(図 4)。左上のほうは、人手で頑張ってつける方法です。あと残りの３つが音声認識でやる方法で、リスピークだとかハイブリッド、原稿推定といった方法を次から次へと繰り出して、字幕が付与される番組を増やしてきました。

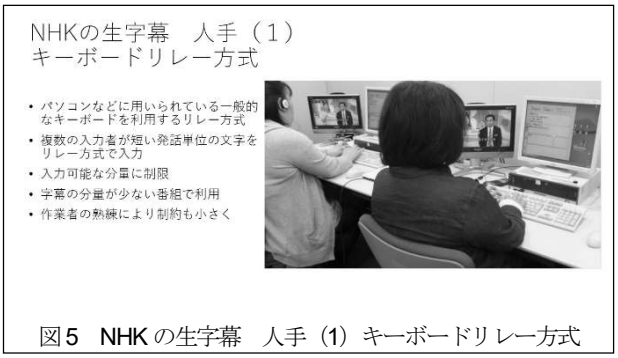


図５　NHK の生字幕　人手（1）キーボードリレー方式

まず、人手のほうですけれども、一般的にいろいろなところでもよく使われているのはパソコンとかで使われている普通のキーボードを使って、短い文章単位を複数人でどんどん打っていくという方法です。入力できる速さに限界があるので、あまりいっぱいしゃべらない番組で使われています。(図 5)

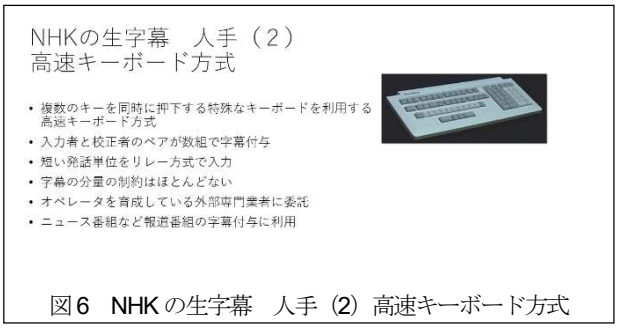


図６　NHK の生字幕　人手（2）高速キーボード方式

そして２つ目が、高速キーボードというのがあります。これは御存じの方いるでしょうか。複数のキーを同時に押すと、その組み合わせで文節「今日は何とかで」みたいなところまでべろべろっと出てくる速記キーボードというものです。これも１人では全部つくれないので、入力する人と漢字変換の間違いを直す人がペアで字幕にしていきます。これはものすごく速いので、字幕の分量の制限はほとんどありません。ただし、こんなキーボードを操れる人たちというのは専門の方々しかいないので、外部業者にお願いするしかないということです。(図 6)

音声認識技術

この後は、音声認識とはというあたりと、字幕制作のときに音声認識はどんな性能が必要なんだよというところと、字幕制作用の音声認識として私たちが開発してきたのはこんなものですよというお話をしたいと思います。

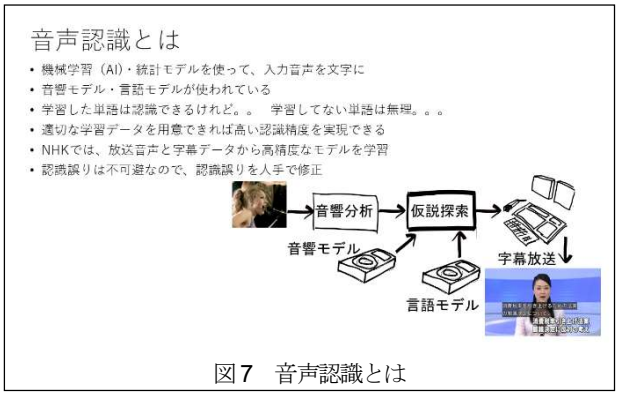


図7　音声認識とは

最近、携帯でもSiriだとかグーグルも音声認識しています。これがどんどん使えるようになったというのは、機械学習とかＡＩ、統計モデルなどを使っていろいろ変換する技術が進んできたからということになります。

中には音響的なモデル、例えば、「あ」という音はこんな波形だよみたいなことを学習したものや、言葉のモデル「今日」というのは１つの単語だよとか、そういう言葉と言葉がどんなふうにつながっていくんだよみたいなモデルが使われています。

このモデルというのは、世の中のいろいろなものから学習したものでして、学習した単語は認識できるけど、学習していない単語はやっぱり認識できませんというのが、基本的な動作になるかと思います。

だとしたら、適切な学習データをたくさん用意できれば、意外と高い認識率が得られますよねというところで、早い段階から音声認識技術を実用化してきたのがこの字幕用のモデルです。NHKでは、幸い放送音声というのは山ほどありますし、それに人手で一生懸命字幕をつけていた部分もありますので、それらから高精度なモデルが学習できるというところがある意味強みです。

とはいえ、間違わない音声認識をつくるというのは永久に無理です。ニュースに字幕をつけるのに、間違った情報をお伝えしたら大変なことになりますので、認識の間違いは人手で直して出しましょうというのが基本になります。(図7)

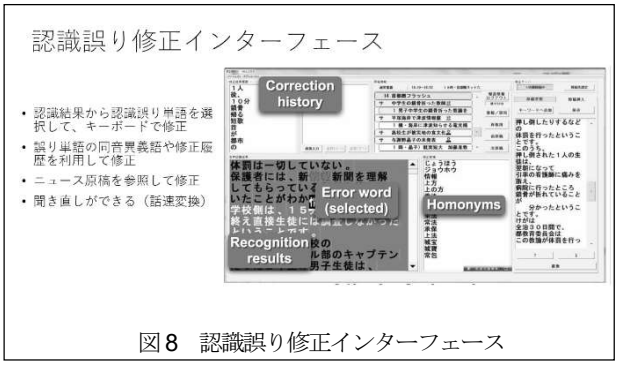


図8　認識誤り修正インターフェース

音声認識術と同時にこの音声の認識誤りを人手でどうやって修正するかみたいなインターフェイスも一緒に開発してきて、最終的に字幕がつけられるというところになっています。

字幕用のインターフェイスなので、認識結果から認識が誤っているところをタッチして選んで、それでキーボードで修正するのが基本ですけれども、全部それを繰り返していたら大変なので、前に同じ単語を直したよとか、同音異義語の履歴を同時にバラバラ出しておいて、即座に直せるようにという工夫がしてあります。

また、大本のニュース原稿みたいなものもありますので、そういったものを参照して修正するといったことができるようになっています。

あとは、当然、聞き漏らしたりすることがあるので、聞き漏らしているところを聞き直ししながら修正していくというような仕組みが必要になってきます。(図8)

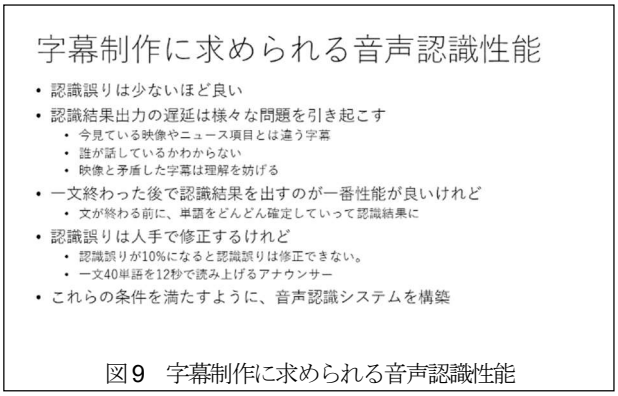


図9　字幕制作に求められる音声認識性能

この修正端末の仕組みがすごくよかったとしても、音声認識にどういう性能が求められるかというと、あまり間違わないでくださいというのが当然のことです。

あと、認識結果があまり遅れて出てくると、字幕として様々な問題が起きてきます。今見ている映像とかニュースと違う字幕が出てきても相当気持ち悪いですし、複数の人が会話しているようなところでは、誰がしゃべっているか分からなくなるのと話の筋が追えなくなってしまう。特にスポーツで起こるのが、映像で誰かがゴールしているのに、字幕は全く違うことを言っていたりすると、まるで分からなくなってしまう。要は、遅延を少なくしましょうねということが、ある意味すごく大事なんです。

音声認識は１文全部聞いてから、後から全部の文脈を振り返って、この単語だねと決めると認識率が一番よくなりますが、こういった遅れという問題があって、ちょっと認識率が悪くなっても単語をどんどん確定してお伝えしていくという仕組みを入れた音声認識というのをつくっています。

そして、認識誤りを人手で修正すると言ってもどのぐらい修正できるかという話でして、アナウンサーさんの原稿は１文大体40単語ありまして、それを12秒ぐらいで読んでいます。認識誤りが10%になると、３秒に１回誤りが出てくる計算になってしまいます。３秒に１回間違えたら直せる気が全然しません。ただ、認識誤りが５%、逆に言うと認識率が95%になったら、６秒に１回の誤りという計算になります。それだったら直せそうな気がします。(図9)

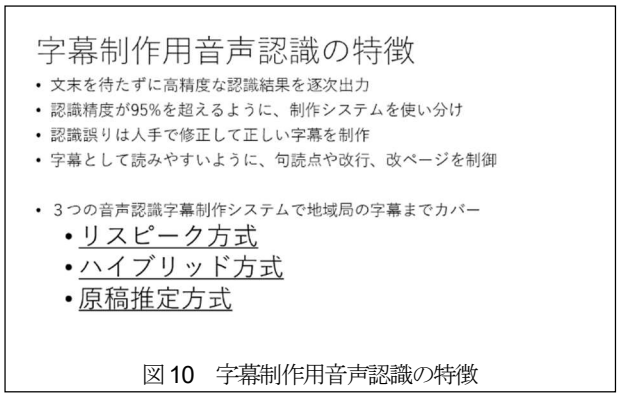


図 10　字幕制作用音声認識の特徴

こういった条件を考えていくと、私たちが構成した音声認識はこのようになります(図 10)。

文末を待たないで、高精度に認識経過をどんどん確定していきます。直すための速さというのがあるので、認識精度が 95%を超えるようにシステムを使い分けて、字幕をつけていきます。

あとは、やっぱり字幕として読みやすいように、句読点とか改行とか改ページということも考えて、音声認識の中でここに無音というか間があったよとか、そういう情報をちゃんと使って、句読点とか改行とかにもつなげていくといった工夫が入っています。

その中でつくってきたのが、冒頭のほうでも紹介したリスピーク方式とハイブリッド方式、原稿推定方式ということになります。この後、これらを詳しく説明していきます。

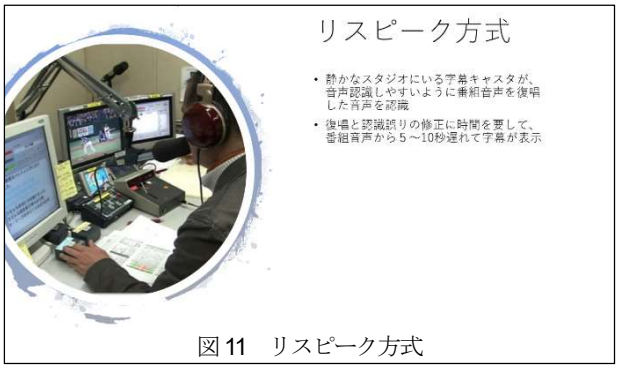


図 11　リスピーク方式

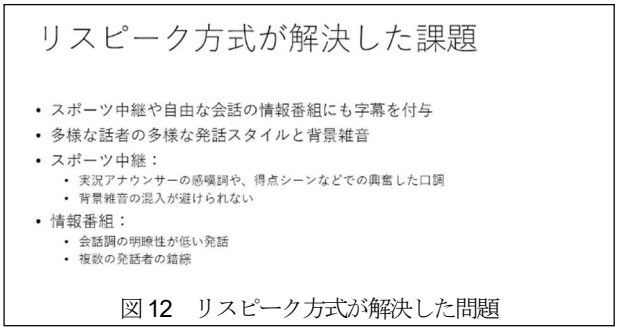


図 12　リスピーク方式が解決した問題

リスピーク方式ですが、これはずるじゃないかとかいう雰囲気もありますけど、番組の音声を認識するのではなくて、静かなスタジオにいる字幕キャスターが音声認識しやすいように番組の声を復唱して、それを認識します。そんなことをすると字幕の遅れが大きくなりますが、5秒から 10 秒の遅れで字幕が出るようにと頑張っています。(図 11)

どんな番組に使えるかというと、スポーツ中継や、自由発話が多い情報番組、いろいろな人たちが対談するようなものです。こういったものは結構いろいろな話者が出てきて、いろいろな発話スタイルがあります。あと中継ですと、特にスポーツ会場で中継実況している場合、背景雑音が避けられないケースがあります。そういったものに対応できます。

スポーツ中継ですと、実況のアナウンサーの感嘆詞とか得点シーンなどの興奮した口調だとか、そういったところでちゃんと字幕がつくようにしましょうと。

情報番組では、相手がある程度分かっているところはぐにゅぐにゅとしゃべっていたりとか、あと複数の話者がしゃべっていて話が錯綜してしまい、そのまま全部文字に換えると順番がひっちゃかめっちゃかになるというところは、ちょっと整理してからリスピークするようにしています。(図 12)

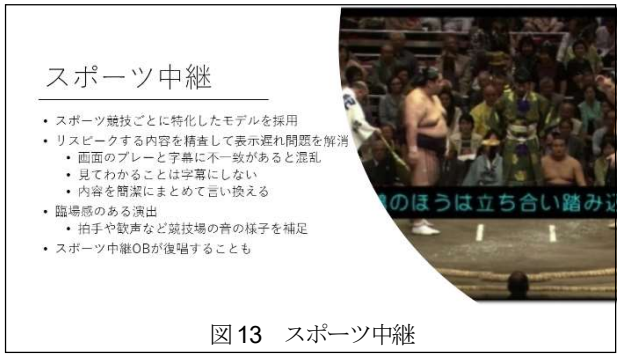


図 13　スポーツ中継

もう少し詳しく言うと、スポーツですと例えば相撲だとか野球に特化した言葉のモデルを教えてやって、そういったものを認識していくということになります。

あとはリスピークする内容を相当精査して、表示遅れがあると困るやつはしゃべらないということをやっています。例えば、アナウンサーさんが絶叫することもたくさんあって、サッカーとかの「日本ゴーーーール」とか叫んでいるやつをリスピークする人は、「日本がゴールしました」と淡々と言って字幕にします。プレーで「誰から誰へのパス」と言っているのも、字幕になる頃にはボールは全然違うところにいっちゃうでしょうから、見ていれば分かるものには何も言いません。また、トータルの試合の内容で「誰のアシストで何が起こりました」みたいな、簡潔にまとめて言い換えるようなことをしています。

あとは、聴覚障がい者さんのための字幕なので、「会場がワーッと沸きました」というのは、番組のアナウンサーさんは言っていなくても、「今、会場がワーッと沸きました。○○さんが入ってきたのではないでしょうか」みたいなことをちょっと補足として入れています。

スポーツの実況というのは子弟関係があって、こうやってやるんだよみたいなのを師匠のアナウンサーさんから習います。特に相撲なんかはそうで、リスピークするアナウンサーさんにＯＢの方を頼んだりすると、「俺が教えたんだよ、あいつは」とか言って、実況のアナウンサーさんがしゃべる前に「ここはこう言うんだ」とぺろっとしゃべって字幕にしちゃうとか、そういったことも起こります。(図 13)

図 14　情報番組

情報番組の場合は複数の人たちがしゃべっているやつを分かりやすく整理してリスピークします。あと、情報番組が難しいのは、いろいろな話題を扱います。料理だとか流行の話だとかそういったもの。そういった単語をどうやって音声認識に教えるかというのが一番の課題です。事前に入手した、原稿とは言わないのですが、こんな話題ですよみたいな構成表から出てきそうな単語や、番組のリハーサルを一生懸命文字に起こして音声認識に教えたりします。あとはリハーサルのときに実際にリスピークして、ちゃんと字幕が出るかだとか、どう言い換えたら認識しやすいかなんていうことを試しながら、音声認識がうまくいかないところを一生懸命人手でカバーして、字幕がついている状態です。(図 14)

図 15　番組音声を直接認識

そして次がハイブリッドと言いまして、番組要素を直接認識するという方法です。もちろんリスピークがある意味最強なんですけれども、言い直す人を用意するだとか、かなり準備が必要なので、その準備を減らしたいと。減らすことができれば、渋谷の東京だけではなくて地域放送局でもやれるよね、緊急報道なんかでも字幕がつけられるよねということになりました。(図 15)

ハイブリッド方式は、定時の短いニュース「２時のニュースです」みたいなやつに使われています。運用コストはリスピークする人がいないので低いです。どうしても認識できないところは修正する人がそのところだけリスピークして、復唱した音声を認識して、番組全体をカバーしています。(図 16)

こちらは記者さんが全国で集めてきたニュース原稿がデータベースになっています。そういうニュースのデータベースを事前に、また放送中も人手で教えることなく音声認識が学習して

図 16　番組音声の認識例

字幕をつけていけるといったことで、事前準備の手間も少なくなっています。

これだけコストが下がると、東京だけでなく、大阪とか名古屋とか福岡とか仙台発の地域色の番組にも字幕がついてくるということになっています。

図 17　地域局、災害時の字幕に

あとは､先ほどのリスピークだとか速記用キーボードだとか､そういったものに比べると修正するオペレーターさんを局内で確保できるので、地域局でもそうですけれども、突発的なニュースや災害時の字幕付与に使われています。(図 17)

音声認識の結果と修正の人が直す様子を御覧いただきます。平成 22 年なので相当古いですけれども、実際に使えるかどうかという実験をしていたときのビデオです。

（ビデオを流しながら説明）

音声認識の結果は、御覧のとおり音声が入ってから１秒か２秒後ぐらいにどんどん作成していきます。それに対して、間違いは人が見つけて直してから字幕として出していく。

この直す作業が３秒に１回になると誤り率が 10%ぐらいになるので、これぐらいならば直せるよねとなりました。

（ビデオ終了）

災害報道への字幕付与

そして次、先ほど災害報道の話が出ましたが、災害報道への字幕付与ということで、もう１つ工夫があります。字幕付与の目標が修正されましたというのと、災害報道が何で難しいのかという話をちょっとさせていただいて、その難しいことを何とかやっつけようとした取組みを御紹介します。

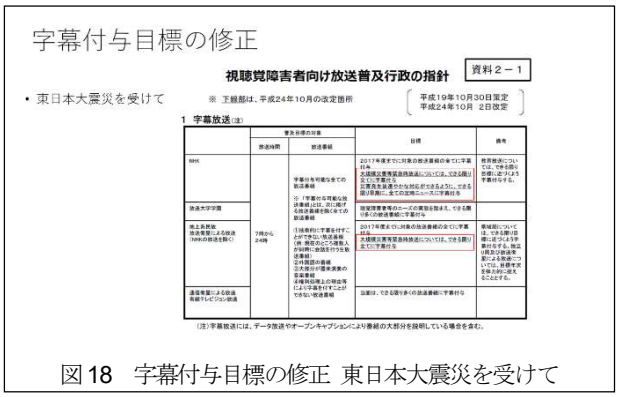


図18　字幕付与目標の修正 東日本大震災を受けて

やはり、東日本大震災がショックだったと思います。一番先のほうで御覧いただいた目標(図3)に、大規模災害時緊急放送についてはできる限り字幕付与みたいなのが追加されました。当然、そういうことに対応していけるようにしていかなければいけないのですが、なぜ難しいのか。(図18)

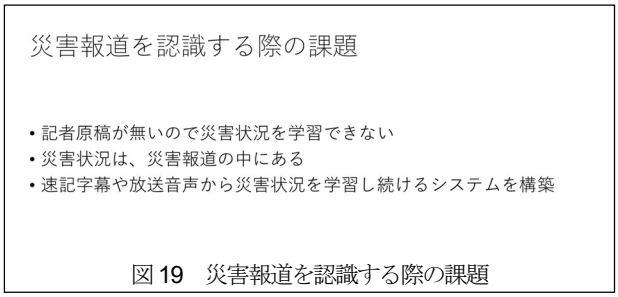


図19　災害報道を認識する際の課題

先ほど話しましたように、世の中のニュースは原稿データベースから学習できるので、「今日の自動車を盗んだ犯人のタロウさんの漢字はこうだよ」なんていうことは知っているんです。だけど、災害報道の場合は、もはや記者が原稿を書く前に状況を伝えなければいけない。なので、何も情報がない。よく繰り返される、起こっているようなことならば認識できるんですけれども、突発的にわけの分からないことが起こるときに、そういったものをどうやって教えていくか。

結局、災害状況というのは、そのときに頑張って災害報道すると、１つの報道を繰り返し繰り返しお伝えするので、自分たちで頑張ってつけた字幕だとか、そういったものから学習しましょうと。速記字幕とか放送音声から災害状況を学習し続けるシステムというのをつくりました。(図19)

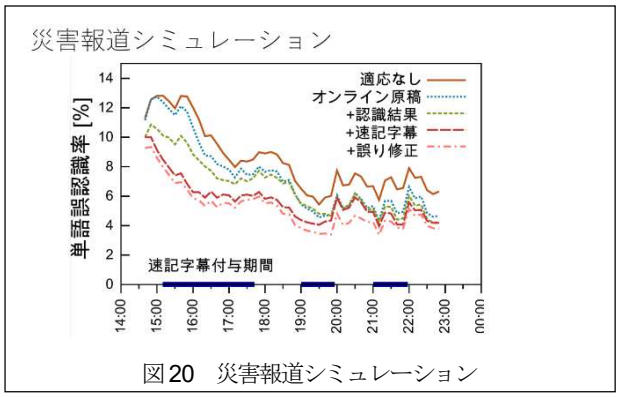


図20　災害報道シミュレーション

こちらが東日本大震災のときのシミュレーション(図20)で、大体この14時何分に地震が起きました。NHKは速記の人たちにお願いして頑張って字幕をつけていきますが、５分ぐらいのニュースかなと思ったら一晩中ニュースが続きました。さすがに最初お願いした速記の人たちも疲れてきて、ちょっと休ませてくださいという時間帯がこの辺（18～19時、20～21時）にありました。この時間帯は残念なことに字幕がついていません。なので、音声認識はこういう時間帯に字幕をつけていきましょうということをします。

縦軸が音声認識の誤りですけれども、何もしない方法（オレンジ色）だと、いつまでたっても目標値、誤りが５％まで下がりません。いち早く５％とか、しいて６％ぐらいまで下がればいい。なので、５分おきとか10分おきでいろいろなリソースをひたすら学習し続ける一番ベストケースがこのピンクの点線です。学習を始めて１時間後ぐらいには、誤り率５％は超えているけど、放送として成立するところまでいくよねということで、自ら放送した内容をどんどん学習して、災害の報道に対応するという仕組みを入れています。

図21　認識誤り修正の課題（原稿推定）

今のところまでで、仙台局ぐらいまで字幕がついたかと思います。まだ地方局がたくさんあります。今度は認識誤りを修正する人を用意するというのが難しくなってきます。意外と世の中の感じだとかちゃんと知っていてもらわないといけないし、結構急いでパパパッとやる作業なので、手の速い人でないとだめですということで。とはいえ、仙台局止まりではなくてもっと先のところまで字幕つけていきたいので、こんなことをしました。(図21)

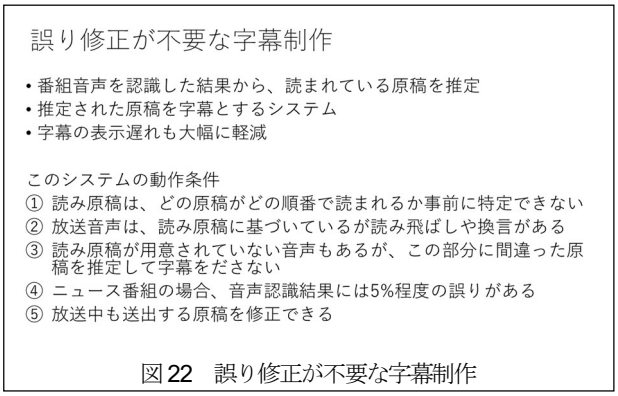


図22　誤り修正が不要な字幕制作

もう誤り修正するのをやめましょうと。結構捨て身な対応ですけれども、番組音声を認識した結果から、どの原稿が読まれているかというのを推定して、推定された原稿を字幕として出します。そうすると誤りがない。

何でできるかというと、地方局のほうにいくと、結構ニュースの差し替え、原稿の差し替えだとかアドリブ的なのがどんどん減っていくので。あとは字幕の表示の遅れというのも随分解消できます。

一瞬聞くとすごい簡単そうですけど、やってみると意外と難しくて、要は、原稿を読む順番がちゃんとはっきりしていて、頭から順番に出してくださいというタスクだったらとても楽なんですけど、どの順番で読まれるかよく分かりませんと言われました。

それから、放送原稿自体は読み原稿に基づいているけど、読み飛ばしたりとか言い換えたりとかするので、音声が原稿そのままではないけど一番近い原稿を出してくださいという話です。

また、読み原稿が全く用意されていない音声もあるので、そこで間違って別の原稿を出さないでくださいと。

あとは、重傷者の数とかどんどん変わるので、放送中でも送出する原稿を修正できるようにしてくださいと。意外と大変な感じです。(図 22)

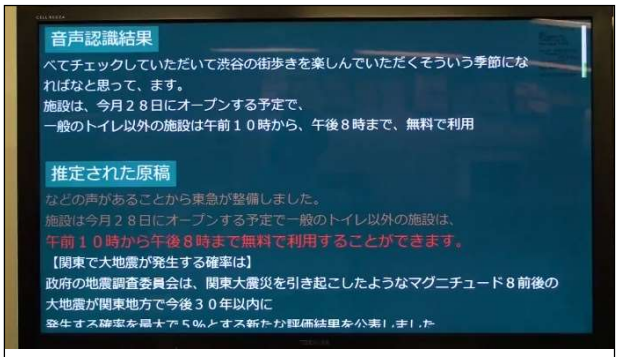


図 23　音声認識結果から原稿を推定する様子

私たちがつくってきたのはこんな仕組みです(図 23)。この後、ビデオをご覧いただきます。

上のほうに音声認識結果、下のほうに推定した原稿が出てきて、これですよと確定できたところでパッと赤字になります。今までは認識が全部終わってから修正して字幕を出していたのが、半分ぐらい読めば、もうこれだよねと確定できる。結構字幕が出るのが早くなります。

あと、最初のほうでパラパラッと動くのですが、原稿の読み飛ばしとか順番を自動推定している様子がお分かりいただけると思います。

（ビデオを流しながら説明）

パラパラッと項目を見つけました。半分ぐらい読めば原稿を確定できます。

この辺は認識しますが、インタビューなので原稿がありません。ただ、インタビューの部分はＮＨＫの場合オープンキャプションがあるので、字幕はつけなくていいでしょうとなります。こういったところで間違って字幕を出さなくて、その後、また原稿に戻って字幕を出せる仕組みというのを作りました。

（ビデオ終了）

地方局展開

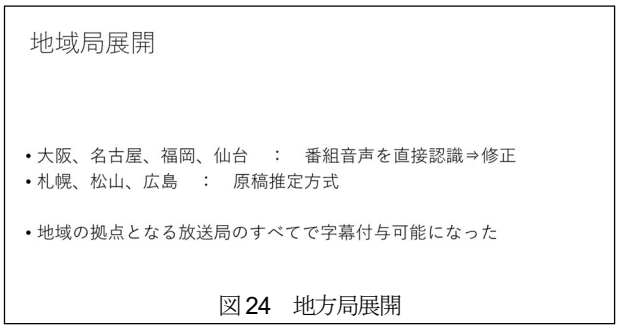


図 24　地方局展開

札幌、松山、広島にこういった方法が入って、一応、各地方の拠点となる放送局の字幕は出せるようになりました。そうすると、災害時だとか緊急時に地方の状況を伝えられるようになります。(図 24)

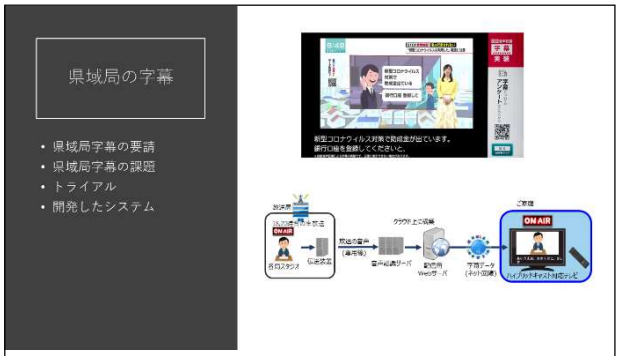


図 25　県域局の字幕

さて、今取り組んでいるのが県域局の字幕というところになります。県域局にも字幕つけてくださいねという要請がきまして、みんなでつけましょうということになりました。県域局に字幕つけるにはこんな難しいですよ、今こんなことに挑戦していますよというのを御覧いただきます。(図 25)

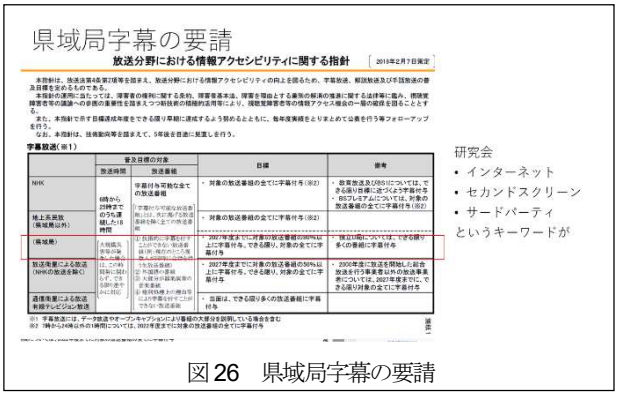


図 26　県域局字幕の要請

先ほどの総務省の目標(図 26)。2018 年に新しい目標がつくられました。そこで入ってきたのは、県域局にも字幕をつけてくださいねというものです。これを決める研究会では、インターネットとかセカンドスクリーンとかサードパーティとか、放送の枠だけにとらわれずに、広く考えて県域局のものに字幕をつけませんかという話でした。

県域局字幕の課題

- 県域放送局の字幕提供の機会を増やしたいけれど。。。
  - オペレータ、設備、設備運用は現実的ではない
  - 県域間格差が生じない全国展開

解決案
- 県域局番組をクラウドで音声認識
- 認識結果をそのままセカンドスクリーンや
  ハイブリッドキャストに配信
  - 正確性が保証される字幕放送とは異なるメディア
- これを、字幕放送の代用としてもらえないか

図 27　県域局字幕の課題

例えば、私、秋田出身ですけど、秋田局の夜８時の番組とか、３人ぐらいしか局にいないんですよ。そこに修正の人を用意してくださいとか、その人が泊まりで対応しなければいけないのはちょっと現実的ではない。あとは局が突然増えるので、オペレーターとか設備をバーンと増やすための投資が必要だ。それに順次やっていくと県域格差だとかなってしまうので、ここら辺が課題になりました。

解決案としては、番組をクラウドにあげてクラウドで認識してしまえと。認識結果をそのままセカンドスクリーンとかハイブリッドキャスト、ハイブリッドキャストというのはネットにつながるテレビの規格ですけど、そこで配信しようと。字幕放送というのはちゃんと正確ですよと保証されているものだけど、ちょっと違うメディアで「音声認識結果です」といって出したら誤っているところは理解してもらえるだろうし、これを字幕放送の代わりとしてもらえないかなということを考えています。(図 27)

トライアル　自動音声認識実験

- 番組音声を低遅延でクラウドに送信
- 認識結果をインターネットで配信
- ハイブリッドキャスト対応テレビで表示

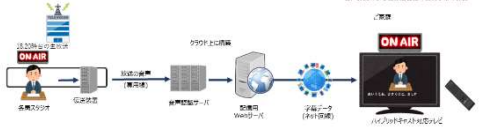


図 28　トライアル　自動音声認識実験

行った実験は、番組音声を低遅延でクラウドに送って、認識結果をインターネットで配信して、ハイブリッドキャスト対応の受信機で見れば、放送で来たものとネットで来たものが同時に画面に表示されて、何となく字幕のようにも見えるといったものです。(図 28)

今まで３回実験しています(図 29)。最初、福島・静岡・熊本で行って、このときは音声認識結果をセカンドスクリーンという別のタブレットとかに出したんですけど(図 30)、あちこち見るの大変だし、同じスクリーンで見たいよということで、2019 年 10 月からはハイブリッドキャストでやっています。あとは局数を増やして、2020 年５月からはハイブリッドキャストでトライアルをしました。(図 31)

トライアル　自動音声認識実験

- 実験 2019/2/4～8/30
  - 610ローカル　福島、静岡、熊本
  - セカンドスクリーンへ配信
  - 要望：字幕を同じ画面で見たい
- 実験 2019/10/28～11/8
  - 610ローカル　福島、静岡、熊本
  - ハイブリッドキャストへ配信　番組と同じ画面に字幕表示
- 実験 2020/5/11～7/10
  - 610ローカル、845ローカル　秋田、金沢、神戸、鳥取、高松、熊本、沖縄
  - ハイブリッドキャストで配信
- 視聴覚障害者、一般視聴者の反響を調査
  - 音声認識で県域局発の番組を認識してどの程度正確な文字がだせるか
  - 誤認識がある字幕を視聴者がどの程度許容するか
  - セカンドスクリーンに送られる字幕を視聴者がどの程度許容するか

図 29　トライアル　自動音声認識実験スケジュール

図 30　自動音声認識実験の様子 1

図 31　自動音声認識実験の様子 2

これで何を確かめたかったかというと、音声認識で県域局発の番組を認識して、どのぐらい正確な文字が出るか。あと、誤認識がある字幕を見て皆さんどう思うのか。そして、セカンドスクリーンとか見やすさというのを、どのぐらい受け入れてくれるのかというあたりを調べました。

字幕用音声認識の今後

- 人手をかけられない放送への字幕付与
- 認識精度がある程度見込めるところから始めたい
- 完璧になるまではあと何年もかかり、それは待てない
- できるところから始めるための技術開発も必要

図 32　字幕用音声認識の今後

こういった実験を通して今後やっていかなければいけないのは、もう人手を全然かけられませんという放送にどうやって字幕をつけていくかということと、音声認識を使う時点で、認識精度がある程度見込めるところからやっていきましょうというのと、完璧にはもうずっとならないので、できるところから皆様に御理解をいただいて進められればなと思っています。(図32)

取材映像の認識

ここまでは音声認識で字幕という話でしたが、安全・安心という点でもう１個大事なポイントがあります。取材映像の認識について、お話しさせていただきたいと思います。なぜ取材映像の字起こしが必要なのかということと、それのシステムの導入事例をちょっと紹介させてください。

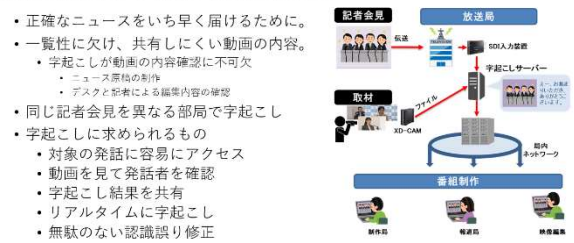


図33　取材映像の字起こしの必要性

取材映像の字起こしの必要性というのは、何か事件とか記者会見とかあったとき、それをいち早く放送に出すためにどうするんだというところです。ただ、映像というメディアは、残念なことに一覧性にものすごく欠けているし、複数人でここだよねと共有するのがものすごく難しいです。

なので、NHKだとか、多分どこの報道機関もそうだと思いますけど、取材映像を全部字起こしした上で、例えばデスクと記者の人が意見を出し合って、このコメントとこのコメントをつないで次の放送に出しましょうだとかするそうです。その文字起こしを必ずしないと、要は客観性がないので。

例えば官房長官の会見とかそういったものは、報道局の中でも政治部の人も社会部の人も何とか部の人もみんな会見内容を字起こししているともったいないし、みんなで協力してさっさと終わらせてニュースつくったらいいのにと思いました。

ここで字起こしというのは、記者会見の特性も含めて、対象の発話に容易にアクセスするというのと、今まで字起こしというのはテキストだけでしたが、やっぱり動画と一緒に見て、誰がしゃべったんだよというところも一緒にやりたいと。それに、せっかく字起こししたからその結果は共有したいよねと。あとは、ファイルで入ってきたのを後でやるのではなくて、会見があったらリアルタイムに字幕みたいに字起こしして、会見が終わったらすぐニュースに出せるようにしたいよねというあたりが求められてきました。(図33)

図34　取材画像認識システム

こんな仕組みです(図34、左上)。会見に行った記者の人たちは、大体どこら辺で何をしゃべっているか分かっていると。全部の一字一句の字起こしが欲しいのではないので、ざくっと会見を切り分けて、そこにキーワードを付与して、目的の場所にさっさと飛べるようにしましょうと。

あとはこちらです(図34、右上)。普通こういった字起こしは、動画の再生ボタンを押して、認識結果を直して、ストップボタン押して巻き戻してって大変なんですけども、単語をクリックするだけでその位置から動画が再生されます。何か直そうという動作を起こすと、そこで動画の再生が止まります。

これはウェブのインターフェイスなので、複数の別々のところの人たちが同じ素材にアクセスできるようになっていて、誰が何を操作しているかが丸見えになります。そうすると、何人かで協働してこの字起こしが作業できるといったことになります。(図34、右下)

今のところで、かな漢字変換の前の段階から見えているというところが結構ポイントになります。今回、複数の人たちが同時に同じ素材にアクセスできるようにして、いろいろな制限を全くかけていません。誰かが何かをしているというのを見えるようにするだけで、同じ目的を持って仕事している人たちの中で使うものなので、誰かが触っているときにこれを別の人が触れないとかそういうことをしなくても、できるだけ生な状態で画面を共有させるということで、うまくいくという仕組みにしてあります。

こういった仕組みがNHKの中では先ほど言った地域の拠点となる局に置かれて、記者会見だとかそういったものを字起こししています。

先ほど、共有できるという話をしましたが、それがまたおもしろいことを生んできています。１つの例ですと、働き方改革に大きく関わってきています。これはＶＰＮ経由でリアルタイムに自宅から記者会見の様子が見られます。今までは記者会見の会場にNHKから何人も行っていましたが、これならば撮る人が１人行って、記者や記事を書く人は自宅からこれを見てパパッと記事を書いて出せるようになりました。

あとは地域局にまで展開することで、例えば、熊本の気象協会の記者会見を福岡の人たちも一緒にシェアして見られるようになって、今まで福岡の人がそんなの気にしたことなかったけど、ちゃんとケアできるようになるだとか、熊本だけで手に負えないというときは福岡の人が手伝うだとか、情報のシェアというのが大きな力を生んでいくといったことになります。

そして一番最後、これまでは音声認識だけでしたが、これに顔認識とか話者認識を加えたらどうなるかというところです。

（ビデオを流しながら説明）

これはまだ実験だけです。音声認識の結果と一緒に、話者をクラスタリングした結果を出しています。同時に顔認識の結果も表示しています。話者認識は、あまり人の名前あてるところまでうまくいきません。顔認識は、人の名前をあてるところまで結構いっているので、顔認識の結果と話者認識の結果を連携させます。これで１つも名前を打たずにどんどん当てはめていって、誰が何をしゃべったかというところまで出来上がります。

（ビデオ終了）

最終的には誰が何をしゃべったかを時間方向と場所方向で見て、世の中をちゃんと把握してニュースにできるような仕組みというのを考えていきたいと思っています。

まとめ

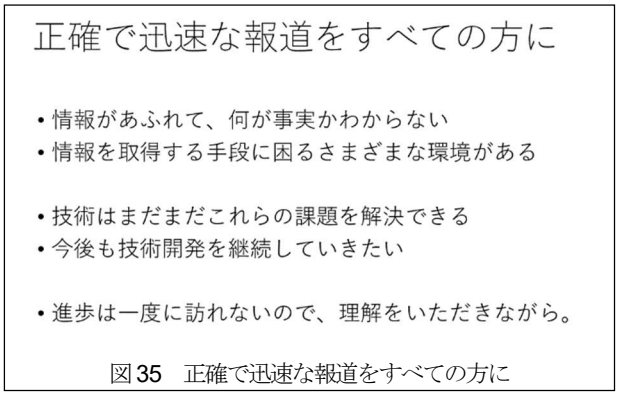


図35　正確で迅速な報道をすべての方に

まとめると、今は世の中に情報がたくさんあり過ぎて、記者の人たちも困っていると。一方で、障がい者さんとか、障がい者さんに限らず情報を取得する手段に困る様々な環境も一緒にありますよね。技術は両方を解決することができるでしょうから、私たちはこれに取り組んでいきたいと考えています。

一方で、突然解決するということは今までの私の経験でもありませんし、すぐには無理なんだろうなと御理解いただきながら、ちょっとずつ改善していきたいと思っています。(図35)

図36　NHKエンジニアリングシステムの紹介

最後です。冒頭にあったように、私は放送技術研究所からエンジニアリングシステム（ＮＥＳ）というところに出向しています。ここでは、技術を御利用いただく方用にカスタマイズや調整をして皆様に御提供できるようにしていこうと、そういった活動をしています。(図36)

ＮＨＫ技研の技術は広く、無料にはなりませんが安く御利用いただける環境があります。それらを使うためのお手伝いをＮＥＳはできますよというところを最後に紹介して、お話を終わりたいと思います。

以上です。ありがとうございます。

本講演録は、令和2年10月27日に開催されたSCAT主催「第108回テレコム技術情報セミナー」のテーマ、「Society5.0時代における国民の安心安全を支える研究業績」の講演内容です。
＊掲載の記事・写真・イラストなど、すべてのコンテンツの無断複写・転載・公衆送信等を禁じます。