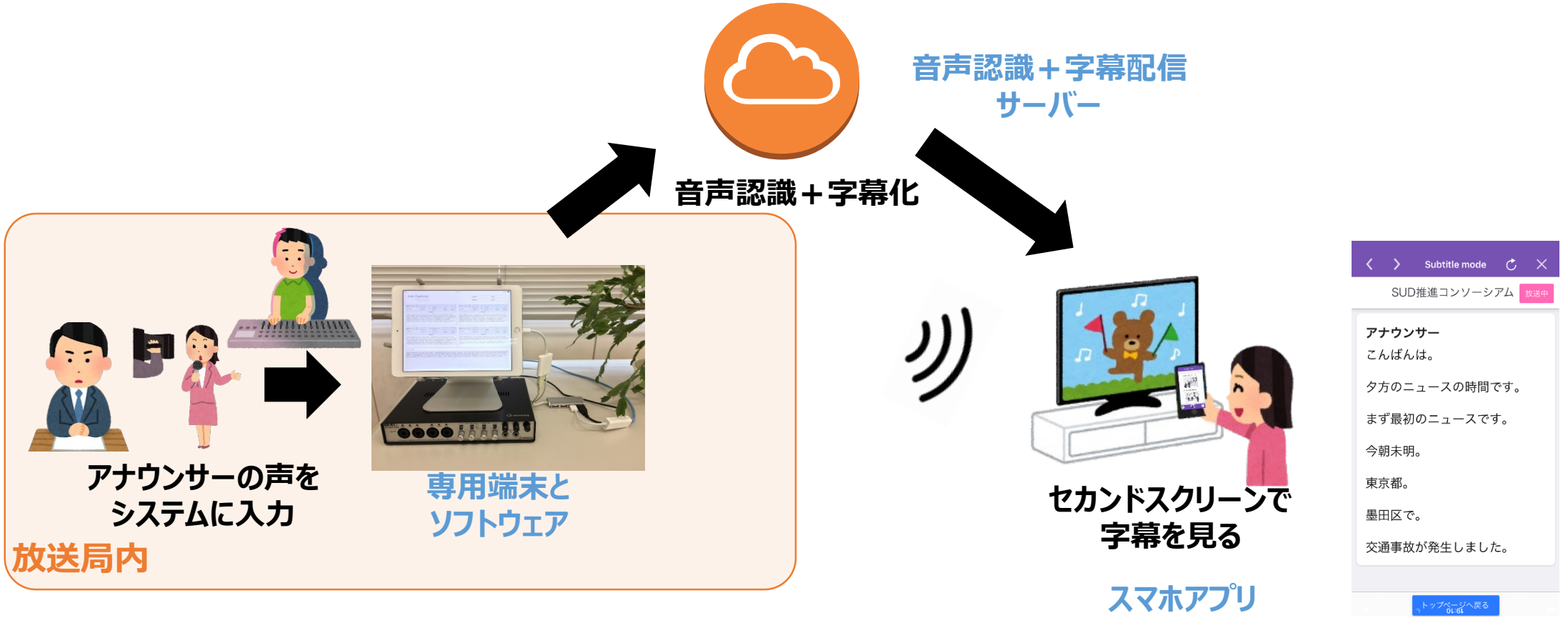
資料10

自動字幕に関する技術動向

令和４年１１月１日
ヤマハ株式会社

自動字幕システムについて

２０１７年に開催された「視聴覚障害者等向け放送に関する研究会」を発端として、
２０１８～２０２０年と３年間に渡り、NICT・放送局・聴覚障害者団体と共同して
テレビ音声の字幕を自動で生成するシステムを開発・実証しました。



音声認識の精度や速度について

【音声認識精度】※ヤマハで実証したニュース番組のデータ

２０１８年
約８５％

２０２１年
約９２％

放送番組データ約２０５０時間分を
NICTの音声認識エンジンに学習させたことで
認識率が向上

情報番組：８６％　教養番組：８１％　娯楽番組：６６％
ニュース番組以外はまだ誤認識が発生しやすく難易度が高い

【音声認識速度】 話し終わりから表示されるまでの平均遅延時間

生放送での従来字幕
約５.６秒

自動字幕
約１.８秒

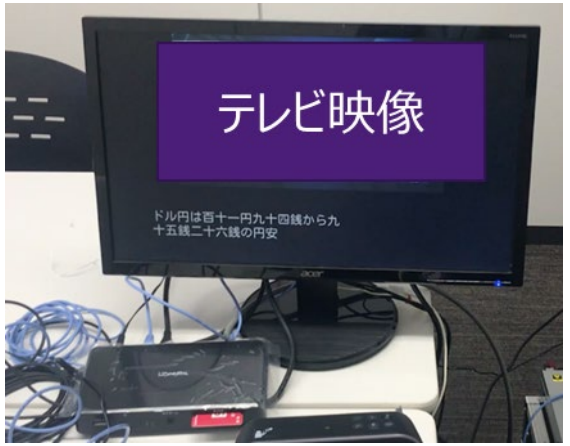
字幕表示機器について

主にセカンドスクリーン（スマートフォン）とアウトスクリーン（セットトップボックスやハイブリッドキャスト）に字幕を表示

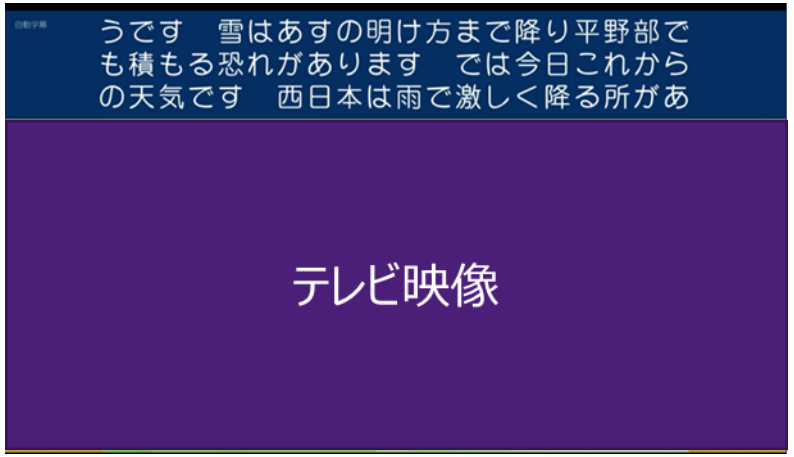
セカンドスクリーン字幕
（マルチスクリーン型放送研究会様との共同実験）

良かった点	読み返すことができる
課題	テレビ画面と手元を見比べるのは見づらい

アウトスクリーン字幕



セットトップボックス



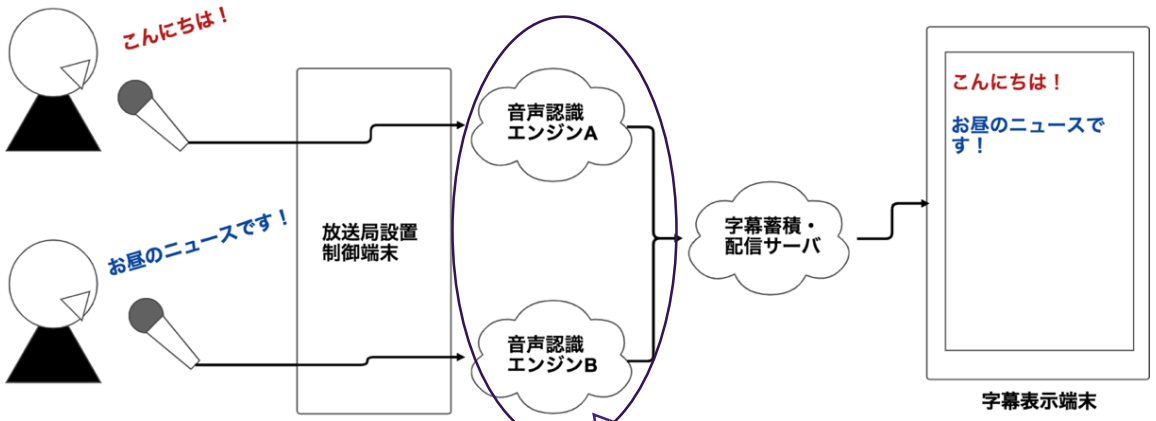
ハイブリッドキャスト

良かった点	普段見ている字幕に近い
課題	専用のテレビや機械が必要

自動字幕の技術的な課題とそのアプローチ

課題：複数人が話をするとき、誰が何を話しているかが字幕ではわかりづらい

話し手ごとに、別々の音声認識エンジンを動かして、個別に文字化して、字幕を色分けなどする必要がある



【新たな課題】
自分のマイク以外にも
自分の声が混ざり込んでしまい、
音声認識結果が乱れてしまう

「音源分離処理」技術の研究開始
マイクの音混ざり問題の解決へアプローチ

２０２０年の総務省調査委託事業での検証において
娯楽番組で２人の掛け合いに「音源分離技術」を適用
約4%の認識精度改善効果

	総字数	誤認識字数	認識精度
オリジナル音声	547	65	88.1(%)
分離処理後音声	377	30	92.0(%)

その後２０２２年現在、改善効果約10%まで改良
さらに多くの話者の分離ができるよう研究を進めている

固有名詞の音声認識精度の改善

事前に辞書データを用意して登録できる仕組みを開発
辞書登録は「副作用※」もあるため、認識サーバーへの効き具合を調整する必要がある
※副作用：辞書登録したことで、もともと認識できていた単語が認識しづらくなる現象

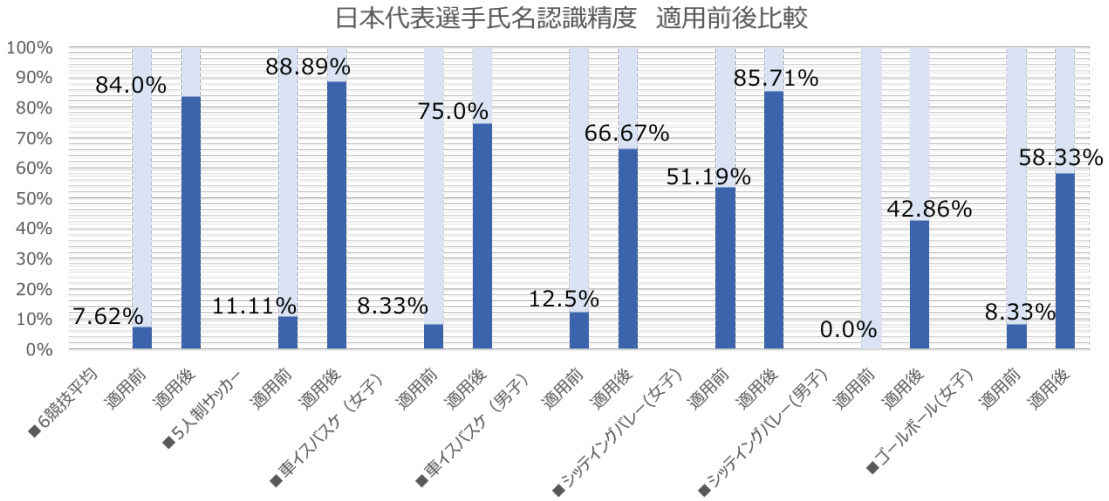
原稿登録効果　（登録を反映した結果、正しい字幕が出力された例　）
(1)　誤）猫　→　正) 理子
(2)　誤）鯖江落ち　→　正) サバイブ

カウンタ	認識結果（原稿登録なし）	文字数	要修正
0:00:13	抜いて	23	
0:00:17	きたオードリーさんからアドバイスがほしい	20	
0:00:19	です。猫さん１９歳愛知	11	1
0:00:21	県のかたです。あれば教えてあげたいね先輩	20	
0:00:25	として出すからね。ああ終わら始めたときにまだ	22	10
0:00:33	生まれてなかったら、ああ全然先輩を確かにね	21	
0:00:37	この激しい芸能かよかすれ	12	5
0:00:39	は鯖江落ちてきました。	11	4
0:00:43	なんとかで	5	
0:00:44	ギリギリ紳士で着替えてきまし	14	5

カウンタ	認識結果(原稿登録あり)	文字数	要修正
0:00:11	生き抜いてき	6	
0:00:12	たオードリー	6	
0:00:13	さんからアドバイスが	10	
0:00:14	ほしいです。	6	
0:00:15	理子さん	4	
0:00:15	１９歳愛知	5	
0:00:17	県のかた	4	
0:00:17	です。教えてあげ	8	
0:00:19	たいな	3	
0:00:20	先輩として出すからね。１９だぜ。	16	
0:00:27	終わら始め	5	3
0:00:28	たときに	4	
0:00:28	まだ	2	
0:00:28	生まれてなかったら、	10	
0:00:30	全然	2	
0:00:31	先輩を	3	
0:00:31	確かに	3	
0:00:32	この	2	
0:00:32	激しい芸能かよかすれ	10	5
0:00:34	ば	1	1
0:00:34	サヴァイヴしてきまし	10	
0:00:35	た。なんとか	6	
0:00:39	で	1	

娯楽番組へ適用
「副作用がおきにくい設定」で調整
約２.３%の認識精度改善

■ **認識精度改善の結果**
- パラリンピック日本代表選手団リスト選手 254名　競技パートナー 23名を、NICT経由でサーバのパラメータ調整の上、辞書登録をおこない適用。
- 団体競技(6競技)の選手入場時に参加選手名の読み上げ場面を利用し、適用前と適用後の認識正答率を調査。
- その結果対象の選手名について、42～77%の改善が見られた。



スポーツ実況へ適用
「辞書データが結果に出やすい設定」で調整
４２～７７%の認識精度改善

今後の課題

特に地方系列局や独立局の番組に対する字幕付与を伸ばしていく必要がある
→人的コスト削減や、自動字幕に関する取り決めを行っていくことが重要

字幕付与が求められている番組の割合：**約65%**

番組への字幕付与の達成率			自主制作
NHK	1局	100%	95%
大手民放(東阪名)	13局	100%	95%
地方系列	**101局**	**86.6%**	**10%(2.4h)**
独立局	**13局**	**35.5%**	**25%(3.75h)**

放送局の字幕付与達成率（総務省公表データを元に算出）

２０２２年にヤマハが放送局約２０局へヒアリング

現状、字幕は間違ってしまった場合、
「訂正放送の対象」という認識
↓
自動字幕が間違う可能性がある以上、
専任スタッフを配置する必要がありコストが大きい

本研究会を通じて議論していただきたいこと

データ放送やクローズドキャプションに自動字幕を使用する場合
「訂正放送の対象となるかどうか、どんな字幕形式に適用するかどうか」
「訂正放送対象外とするために必要とな認識精度を何%とするか」など
取り決めを検討していく必要があると考える