Vol.2016-AAC-2 No.4
2016/12/2

アクセシブルな講演を実現する自動字幕提示手法の検討

布目 光生[1,a)]　渡辺 奈夕子[1]　芦川 平[1]　藤村 浩司[1]

概要：講演や講義などの情報保障手段として，これまでの人手による要約筆記から，音声認識技術を活用したリアルタイム字幕への期待が高まっている．しかしながら，多様な話し手や内容を問わず，安定的に実用レベルの精度を持つ自動字幕を提供する事は，依然ハードルが高い．そこで我々は，社内の聴覚障碍者への試用とフィードバックを通じて，連絡型会議や講演・報告会の場などで利用しやすい Web ベースのリアルタイム字幕システムを開発した．本報告では，このシステム概要と社内での簡易評価の結果について述べる．

キーワード：音声認識，自動字幕，情報保障，アクセシビリティ，会議支援

1. はじめに

近年の音声認識技術の進展により，人と人との会話のような，話し言葉に近いスタイルの音声認識精度が向上し [1]，音声認識技術を活用した応用や用途が広がりつつある [2–5].

そうした応用の一つに，情報保障ツールとしての自動字幕への期待がある．学校での授業や講義，社内での会議や業務連絡などの場で，話し手の音声をリアルタイムに認識し，手持ちのタブレットや室内備え付けの機器で字幕を表示する事で，音声だけでなく視覚的に「見て分かる」情報伝達手段を提供する技術が提案されている [6–9].

これまでも，ノートテイクや要約筆記などが，聴覚障碍者に対する一般的な情報保障手段として提供されてきた．しかしながら，従来の情報保障はボランティアや NPO 団体，あるいは専門のスキルをもった業者によるサービスとして提供されてきたため，時間や場所の制約があることはもちろん，サービスを受ける側にとっても，手間やコストを無視する事はできず，必要な場面で必要な人に十分なサービスを提供する事は難しかった．

これに対し，音声認識を活用したシステムであれば，場所や時間を問わずいつでも音声を文字化でき，利用する側にとっても，また情報保障を提供する側にとっても種々の障壁を下げることができる．

特に，小・中規模なミーティングや報告の場など，これまで人的なサポートが及ばなかった利用シーンであっても，システムを導入する事でコストや制約を大幅に低減して，広く自動字幕化サービスを展開することが期待できる．

2. 課題・目的

実際の利用シーンでは，授業あるいは企業内の情報伝達の場面一つをとっても，発話の内容や話者・発話スタイルには様々なバリエーションがあり，さらに利用場所の環境やノイズ，設備や機材も多岐に渡っている．こうした様々な状況で，安定的に実用レベルの認識精度を達成する事は非常に難しい．

そうした技術的な困難さがある一方で，社会的な要請として既に 2016 年 4 月からいわゆる「障害者差別解消法」が施行されており，その骨子である合理的配慮が求められている．聴覚障碍を持つ当事者にとっては，日々の授業や講義，あるいは，日常での業務会議や連絡会などで必要な情報を入手する必要が常に生じている．

こうした状況を鑑み，日常業務で欠かす事ができない情報伝達や業務会議の場での困りごとやニーズ把握を目的として，当社従業員である聴覚障碍者のうち，52 名を対象にアンケート調査を実施した (2015/06/30-7/10).

その結果，音声認識技術に対する期待として，障碍等級によらず，是非使ってみたい (5 段階中の 5)，できれば使ってみたい (5 段階中の 4)，少しは使ってみたい (5 段階中 3) が 8 割を越えていることがわかった．しかしながら，そのうち障碍等級 6 級では，積極的に使ってみたい (5 及び 4) は，4 割程度に留まり，等級の違いによって期待度が異なることが分かった．また，参加している会議に関し，現状

[1] 株式会社東芝
1, Komukai-Toshiba-Cho, Saiwai-ku, Kawasaki, Kanagawa 212–8582, Japan
[a)] kosei.fume@toshiba.co.jp

の理解度と満足度についても調査を行った．その結果，事前知識として内容がある程度想定できる「定例会議」では，問題なく理解している (5 段階中の 5)，概ね理解している (5 段階中の 4) が 6 割以上を占める一方で，朝礼や昼礼など，資料提示や周囲のサポートが難しいシーンでは，あまりわからない (5 段階中の 2) が 35%程度，半分は理解している (5 段階中の 3) と併せて 50%を占め，約半数の人が半分以下の理解度である現状であった．また，会議の満足度としても，朝礼・昼礼および，ミーティングは最も満足度が低く，4 割近くの人が，仕方なく参加している (5 段階中の 3)，できれば参加したくない (5 段階中の 2)，参加する事が苦痛である (5 段階中の 1)，と感じていることが分かった．

また，普段の会議で困っていることとしては，細部や用語がわからない (77%)，ちょっとした発言やコメントがわからない (83%) といった項目が多く，これらは障碍の程度に関わらず挙げられた．ただし，細部に限らず，話題自体がわからない，という回答も 1 級から 3 級でそれぞれ約 20%の割合で存在する状況が明らかになった．

3. 基本コンセプト

アンケートの結果，音声認識に対する期待は高く，リアルタイム字幕が必要とされていることがわかる．特に，健聴者が主体で時間が限られる情報共有の場では，情報保障が忘れられがちであるため，手軽に導入でき簡単に使える手段が必要である．

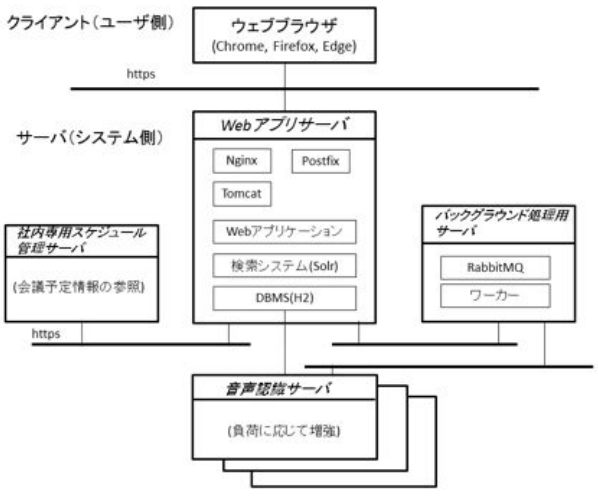

図 1　システムの基本構成

これに対応するコンセプトを具体化する方法として，リーンスタートアップ的な開発手法により，当事者や関係者での通常業務での利用を経て，具体的な機能仕様や妥当性を検証していくこととした．

3.1 ターゲットとする利用シーン

想定とする利用シーンは，企業や学校などの組織内での情報伝達の場，すなわち講演や講義といった情報提供の機会や，報告型の会議などを想定している．前段のアンケート結果からも分かるように，こうした利用シーンは，日頃の活動を回していく上で欠かせない状況でありながら，情報保障が取りこぼされてしまう場合が多い．例えば，主催者の配慮がある大きなイベントなどでは，主催者責任で聴覚障碍者に限らない障碍者向けサポートがあり，逆に，ごく少数での場や対面での場合には，周囲の人が聴覚障碍者に配慮しつつ，コミュニケーションをとることが普通である．そのため，その中間の規模の情報伝達では，情報保障が忘れられがちである．本システムは，そうした場面でも障碍者自身やあるいは話し手自身が簡単にセットアップができ，さらに，話し手はその場のマイク利用の負担のみで，字幕サービスの提供を実現する．

3.2 利用形態

情報保障ツールとしてのリアルタイム字幕機能は，日常業務の中で無理なく自然に使える事が大前提である．字幕を閲覧する聴覚障碍者はもちろんのこと，業務上，当事者に情報や指示を伝える上司や同僚にとって，業務の妨げとなるような手間や準備，運用，または大掛かりな機材または高価な装置が必要となると，現実的な手段ではない．

そこで我々は，業務利用の PC の延長で簡単に普段使いができるように，Web アプリケーションによる提供を基本構成とした (図 1)．

エンドユーザは，Web ブラウザを用意すれば字幕を閲覧する事ができ，また，PC に（適切なオーディオインターフェースを介し）マイクを接続することで，話し手の音声の取り込みが可能となる．さらに，一般的なサーバ／クライアント構成とすることで，音声認識に必要な音響モデルや言語モデルの更新，また社内特有の専門用語や略称，固有名詞などをサーバ側で一括管理することもできる．そのため，エンドユーザとしても，機器や場所を問わず，社内のアクセス可能な PC であれば同一環境での字幕機能が利用でき，さらに一度作成されたモデルやユーザ辞書などのリソースを社内で共有・再利用することも可能である．

4. 自動字幕システム

まず，基本的なシステム構成として，Web アプリケーションの構成をとった．エンドユーザにとって，専用アプリのインストールなどの負担を最小限に減らすとともに，リーンスタートアップ的な手法では欠かせないサーバ側の改良を，エンドユーザ側で意識せずに享受することができる．また，開発者側にとっても，アプリケーション単独の改良だけでなく，基盤技術である音声認識エンジン／辞書／モデルに関わる更新も可能であり，その際にエンドユー

ザの負担を強いることや，運用中の字幕サービスを極力妨げることなく更新が行える．

4.1 構成

全体構成は図 1 に示した通りサーバ／クライアント構成とし，今後の拡張性やアプリケーションからのデータ流用性や再利用性を踏まえ各部のデータのやり取りは，一般的な HTTP や websocket を用いた．

音声認識エンジンはサーバ化されており，Web アプリケーションと疎結合で構成されている．これによって，各モジュールの可搬性や独立性が維持され，それぞれの改良を行った場合に，迅速にそれらの更新をシステムに反映でき，ユーザが改良結果を享受できる．処理の流れは以下の通りである．

(1) Browser で音声を取り込んでアプリサーバへ websocket で送信する
(2) アプリサーバは会議情報の管理を行いつつ
(3) 音声認識サーバに音声を送信し，音声認識結果を受け取る
(4) 音声認識結果はアプリサーバから Browser へ websocket で送信され
(5) Browser はその内容を表示する

4.2 音声認識エンジン

音声認識エンジンでは，モデル学習または認識アルゴリズムの抜本的な改良に加え，実用を踏まえた場合，そのターゲット分野の音声データや書き起こしデータ，関連するテキストコーパス等の学習データを，いかに大量に高品質なものを収集できるか，あるいはその疑似学習データを用意できるか，が課題の一つとなっている．今回の講演や情報伝達型の会議に関しては，既存の利用可能なデータは限られており，かつ企業内の会議音声であれば，基本的には外部に出る事は無く，自社内での活用に留まるのが普通である．

前者のアルゴリズムについて，我々は大語彙への対応や話し言葉対応についても，頑健性の向上を図っており，音韻とフィラー・言いよどみを同時に識別する LSTM-CTC 音響モデル [10] を用いることで，精度向上を図っている．

また，後者ではこの Web アプリケーション自体が土台となり，社内業務に直結した実際の音声データを収集するという役割を担う他，プライベートなクラウドソーシング [11] との連携で，人手のかかる正解データ作成作業を分散し迅速に行う枠組みも用意している．

4.3 ユーザインタフェースの検討

字幕提示については，日常的に利用するシステムである

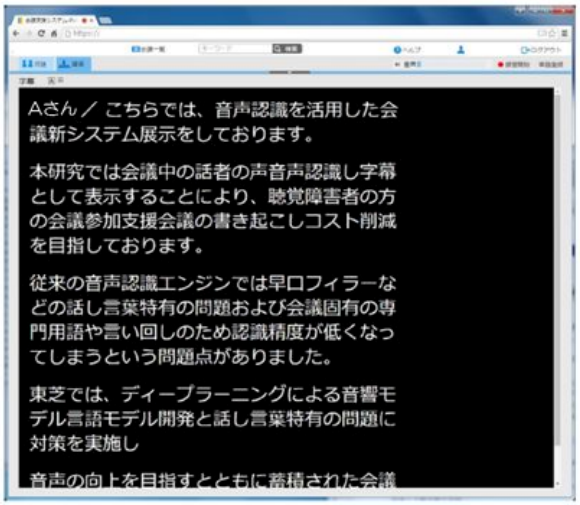

図 2 講演型の字幕表示

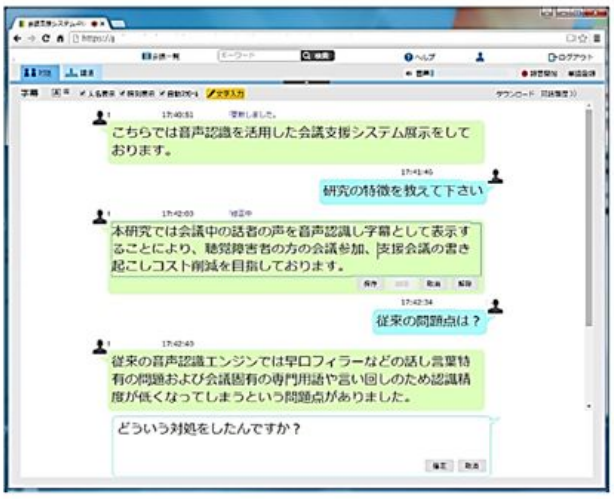

図 3 対話型の字幕表示

ため，エンドユーザにとって直感的で混乱が無く，見やすい字幕提示にすることが重要である．筑波技術大学などでのヒアリングを踏まえ，字幕提示では一般的なハイコントラストで提供されている黒背景白文字のベーシックな画面 (図 2) を踏襲した他，聴覚障碍者側からの意思表示，または発言者以外からの注釈表示の要求を考慮し，対話型の画面 (図 3) を用意した．これらの表示スタイルは，字幕を閲覧するエンドユーザが，必要に応じてタブを切り替えることでいつでも変更することができる．

5. 社内試行に基づくリーンスタートアップ的開発

以上のようなシステムを段階的に構築し，これまで社内の聴覚障碍者の方に業務内での情報伝達を主目的とする会議などで試用してもらった．

利用に際しては，説明時に実演しながらマイクの使い方やアプリケーション機能を示すと共に，実際の認識精度の感触や利用イメージを掴んでもらった．

5.1 利用者フィードバックに基づく改良

最初に，話し手の利便性と音声を入力する事の心理的な障壁を考慮し，負担の少ないと思われるスタンドマイクを提供した．

まず機器の課題として，毎回会議室にノート PC とマイクを持参し，セットアップを行うという利用者が多い中で，外付けマイクの接続前にアプリケーションを起動し，外付けマイクではなく，意図せずに PC 内蔵マイクで音声入力をしてまうケースなどがある事がわかった．さらに，実際の利用の場での行動観察として，話し手は，手元の PC とプロジェクタ等に提示したプレゼン資料との間で頻繁に視線移動を行う事が分かった．それにより，マイクと話し手の位置が変動し，認識精度の低下が見られた．また，情報伝達型の場においても，参加者からの質疑応答やコメントなどの発声があり，それらに回答する時には，話し手の意識がマイクから遠ざかってしまい，音声が正しく拾えないことが多々あった．またターゲットとしていない距離の離れた参加者からの音声が中途半端にマイクに入ってしまい，誤認識されたフレーズや単語などの単位が湧き出して表示されることで，悪い印象を与えている事も明らかとなった．

さらに，アプリケーションを新しく試行する場合はいわゆるコールドスタートに相当するため，部署特有の専門用語や略称，組織名，人名，イベント名などの固有名詞などの認識精度が悪く，悪い印象を与えていることも明らかとなった．

こうした状況を鑑み，効果が高く，比較的実装が容易と判断された以下の機能などを，公開期間中に随時機能拡張や改良を実施した．

- 録音ステータスの表示
- 音声レベルインジケータの表示
- 入力に使用されている音声デバイス名の表示
- マイク利用のガイドライン
- ユーザ単語登録機能

部署	累計音声時間 (hour)	備考
A(スタッフ系)	16.96	週一回の連絡会
B(技術系)	25.17	週一回の連絡会
C(スタッフ系)	29.13	週一回の連絡会
D(スタッフ系)	45.35	教育や講演など
E(技術系)	52.33	勉強会等
F(技術系)	18.75	技術定例，勉強会等
G(スタッフ系)	7.45	連絡会など
H(技術系)	22.56	報告会など
I(スタッフ系)	36.83	報告会など
J(技術系)	2.53	報告会での利用

表 1　利用部門と利用状況例

5.2 蓄積された統計情報の分析／考察

本来のコンセプトに対する仮説検証とそこから導き出される根本的な課題にフォーカスするために，一定期間，実際に本システムを利用している社内の利用実態と利用シーンを概観した（表 1）．集計期間は，2015 年 8 月 18 日から 2016 年の 9 月 28 日の範囲だが，システム自体は現在も継続利用中である．なお，システムの試用開始時期は，一斉ではなく，部署ごとに数週間から数ヶ月程度のずれがある．そのため，累積の利用期間にも大きなばらつきあり，上記に示した累計の音声収録時間が，必ずしも利用頻度を反映している訳ではなく，あくまで試用開始から現時点までのスナップショットとしての位置づけである．

また，今回の対象ユーザについては，表 1 に示した通り，便宜的に社内の利用部署を大きく二つに分類した．技術系とあるものは，設計・開発・製造などに携わる部署であり，スタッフ系とは企画や管理，総務といった部署で業務に携わるユーザを示したものである．

利用状況と傾向であるが，まずどちらにも共通する利用シーンとしては，連絡会とされているものがある．これは，週一回程度，数名から十数名の場で，リーダーに相当する人物がマイクを利用し，社内の連絡事項などを伝達する場で，情報保障を受ける人は，持ち込んだノート PC や，会議室に備え付けのディスプレイ等でリアルタイム字幕を閲覧する，というものである．部門による違いとして，技術系では，専門用語や略称が頻出するような技術ミーティングがあり，一方，スタッフ系では，非定型な打ち合わせや報告会，社内教育などで情報保障ツールとしての試行利用が見られた．

また技術系利用では，システム利用を希望している聴覚障碍者自身が，主担当として業務に参加することが多くなるため，周囲の理解もあり定期的に利用されることが多くなる一方で，これまでの方式や習慣などで，重要な期日や

単語などの伝達については筆談との併用が見られた．聴覚障碍のある利用者自身からは，主体的に参加する会議での利用よりは，一参加者として参加の必要があるが情報保障が無かった，部署横断での比較的規模の大きな報告会などの場で利用できたことが，大変役に立った，との声もあった．規模の大きな報告会では，報告者がスライド資料等を用いるため，通常よりも理解が期待できそうであるが，従来は行間として発表者が何を発言されているかを，全く把握できなかったため，こうしたシステムの価値があるとのコメントもあった．

スタッフ系利用では，情報保障ツールとしての利用場面や内容が多岐にわたるため，会場の違いや発表者の話し方に対応した機材の選択や利用が適切でない点が多いことが目立った．例えば，演台でのスタンドマイクを利用した結果，話し手の口元のマイクの距離が不安定になってしまったり，距離を置いて話した結果，適切に音声が取り込めていないことが多かった．ハンドマイクであっても，発表者個人の持ち方のクセなどで，音声が適切に入らず，結果，適切でない認識結果が頻発し，字幕結果を閲覧した場合に，認識精度として良くない印象を与えるなどの状況が散見された．

5.3 講演向け自動字幕システムの試作

我々はさらに，ここまでに述べた試行結果を受け，エンドユーザが実際に利用するシーンでの根本的な課題を，以下の通りにフォーカスした．

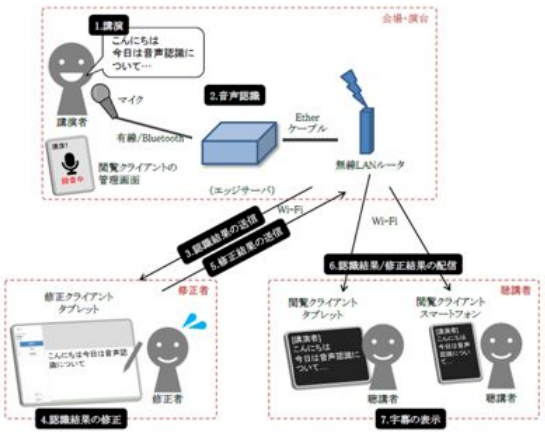

図 4 講演向けスタンドアロン構成

- 多忙な業務において，機器準備の手間や起動にかかる手間と利用のハードルの高さ
- 機器設定の不備による，意図していない状態での音声収録による音声認識精度低下
- 情報保障ツールとして，音声認識誤りをカバーし正確な情報を伝える実用的な修正手段

これらの課題に対応するため，これまでの社内実証中のWebシステムの派生型として図4の構成によるシステムを構築した．

コンセプトとしては，エンドユーザが簡単にシステムの設置，立ち上げ，利用，終了までを行える情報保障ツールであり，複雑な操作を極力省き，設定の揺らぎを回避し，できるだけシンプルで使い勝手の良いものとした．特に学会やイベントで行われる講演は，環境が多様であり，外部ネットワークへの接続も不安定になる場合もある．こうした外部ネットワークから遮断された場でも，簡単利用できる手離れの良いシステムとした．

5.4 今後の課題

この試作システムについては，会議室などの備え付け利用の他，講演などの会場に持参し設定する事を想定している．今後，社内／社外 ([12]) で情報保障が必要な場面での試行を重ね，上記の課題に対する我々のコンセプトの実証，また実利用からのフィードバックを踏まえた，GUIやUXを含めた機能改善等を早期に回し，実用的なシステムへと精錬化していく予定である．

6. 結論

音声認識を活用したリアルタイム字幕システムは，情報保障が必要な部署で定期的な利用があり，情報保障を実現する一つの手段として，提案コンセプトが受け入れられたと考えている．しかしながら，初期設定含めた十分な配慮／利用者周辺の協力がまだまだ必要であり，運用保守も含め，ユーザ単語辞書等の登録などの作業に理解や実働を割ける部署でなければ難しいことも明らかになった．こうした課題に対応するために，Webベースシステムの改良・派生版として，エンドユーザや周囲のサポートする人の手間を軽減するためのシステム構成・試作を実施した．

今後は，システム側のユーザビリティ改善と共に，全自動の音声認識では避けられない認識誤りをカバーし，簡便な修正手段を実現するための手段の検証／評価などにも取り組んでいく．

参考文献

[1] 中川聖一：音声認識研究の動向，電子情報通信学会論文誌 D-II， Vol. J83-D-II, No. 2, pp. 433–457 (2010).
[2] 秋田祐哉，三村正人，河原達也：会議録作成支援のための国会審議の音声認識システム，電子情報通信学会論文誌 D， Vol. J93-D, No. 9, pp. 1736–1744 (2010).
[3] 今井 亨：リアルタイム字幕放送のための音声認識，NHK技研 R&D， Vol. 2012/1, No. 131, pp. 4–13 (2012).
[4] Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P. and Duerstock, B. S.: Using Speech Recognition for Real-TIme Captioning and Lecture Transcription in the Classroom, *IEEE Transac-*

tions on learning technologies, Vol. 6, No. 4, pp. 299–311 (2013).
[5] 高橋麻理子，近藤修明：効率的なビジネス活動を支援する会議音声活用システム，東芝レビュー， Vol. 70, No. 1, pp. 52–55 (2015).
[6] 富士通ソーシアルサイエンスラボラトリ：LiveTalk. http://www.fujitsu.com/jp/group/ssl/products/software/applications/ud/livetalk/.
[7] シャムロック・レコード：UD トーク コミュニケーション支援・会話の見える化アプリ. http://udtalk.jp/.
[8] 栗 田 茂 明：IPtalk. http://www.geocities.jp/shigeaki_kurita/.
[9] NICT: 聴覚障がい者とのコミュニケーション支援アプリ SpeechCanvas. http://speechcanvas.nict.go.jp/.
[10] 那須 悠，藤村浩司：LSTM-CTC を用いた音響イベント検出・除去音声認識システムの検討，信学技報， Vol. 116, No. 208, pp. 121–126 (2016).
[11] 芦川将之，川村隆浩，大須賀昭彦：プライベートクラウドソーシングにおける精度向上手法，人工知能学会全国大会論文集， Vol. 28, pp. 1–4（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/40020082940/⟩ (2014).
[12] 秋田祐哉，塩野目剛亮，白石優旗：音声自動認識による字幕情報保障トライアル (2)，情報処理学会アクセシビリティ研究会 (予定)， Vol. 2016-AAC-002.