人間と機械が協調する現場指向型の音声同時字幕システムの構築

Design of a Field-oriented Real-time Captioning System through Human-Machine Collaboration

○ 井野秀一（産総研）　三好茂樹（筑技大）　白澤麻弓（筑技大）　河野純大（筑技大）

Shuichi INO, National Institute of Advanced Industrial Science and Technology (AIST)
Shigeki MIYOSHI, Mayumi SHIRASAWA and Sumihiro KAWANO, Tsukuba University of Technology

Key Words: Information Support Service, Hearing Impairment, Caption, Speech Recognition, Repetition, Remote Network

1. はじめに

現在，国内において聴覚に障害をもつ人は約 35 万人いる．また，大学などの高等教育機関には，1000 名を越える聴覚障害学生が在籍している．彼らが用いるコミュニケーションの手段には，手話・読話・筆談があり，公共の場（例えば，講演会や授業など）においては，手話等の他に，ノートテイク・要約筆記・パソコン要約筆記がある．しかし，手話や読話を使うには，それらをマスターしていることが前提となる．特に，中途失聴者が新たにそれらを習得するには，外国語の学習と同様な大変さを伴う．その一方で，情報保障者のサポートによる，筆談・ノートテイク・要約筆記という会話の内容を文字に変換する方法もある．しかし，このようなディクテーションによる手法においても，「時間がかかる」「内容の要約が必要である」という性質を含んでおり，円滑なコミュニケーションにとっての悩ましさは尽きない．

一方で，近年のコンピュータを利用した情報処理技術の発達により，顧客サービス（コールセンター）を始めとする様々な分野で音声認識技術が取り入れられつつある．音声認識技術の利点としては，「（キーボード入力に頼らずに）音声を入力手段として利用できること」「文字化の処理速度が速いこと」などが挙げられる．ただし，現時点の技術レベルでは，話し手に関係なく，大規模語彙かつ不特定話者・複数話者を対象として，ヒトのように音声認識することは難しい．さらに，ヒトの場合は，多少の誤りを含んでいる曖昧な文でも，文脈の前後関係などから意味を推測し，不明確な部分をある程度正しく理解することができる．これには，ノンバーバル情報（話し手の表情・口の動き・ジェスチャーなど）の理解が大きく関わっているが，機械には，そのような柔軟な意味解釈は難しい．

そこで，私たちは，人間と機械（音声認識装置）の得意なところを互いに利用するという発想のもとで，音声認識技術とヒトの復唱能力をハイブリッドで活用した，聴覚に障害のある人たちの国際会議や授業における情報バリアフリー[(1)]のための字幕システムの開発に取り組むことにした．

具体的には，公共施設等での利用を考慮したリアルタイム音声字幕化システムを構築するために行ったヒトの同時復唱能力に関する基礎実験，聴覚遅延フィードバックを利用した復唱トレーニング法，顔情報の併用による理解度向上の効果，および，それらの結果から得られた設計仕様に基づくシステム開発について述べる．また，より一層の現場志向のシステムとしては，携帯情報端末（スマートフォン）の代表格である iPhone と Bluetooth 対応のマイクロホンを用いたモバイル型遠隔情報保障システムについても報告する．

2. 設計概念

音声認識の形態には，特定話者方式と不特定話者方式がある．現在の技術レベルから考えると，会議や講演会などのように大語彙環境を対象とする状況では，特定話者方式の利用が現実的である．しかし，会議における話者は，講演者から聴講者（質問者）までを含み，時には複数の話者が同時に混在するケースもあり，特定話者環境を想定するシステムでは，その利用範囲は著しく限定されてしまう．

そこで，このような現場でありがちな実用上の問題を解決するために，Fig. 1 に示すように，話者と音声認識装置の中間に同時復唱者を配置し，ヒトの介在により不特定話者環境を特定話者環境に変換する構成の音声同時字幕システムの研究開発に着手することにした．

Fig. 1　An overview of the proposed speech recognition system collaborating with a re-speaker.

3. ヒトの同時復唱能力

3-1 復唱精度の個人差

同時復唱者を音声同時字幕システムに配置する場合，予め，ヒトの同時復唱能力を知っておく必要がある．そこでどの程度の精度でヒトは同時復唱することができるかを定量的に調べた．

実験協力者は，発話に関するトレーニングを受けたことのない大学生 5 名（22～33 歳，男性：3 名，女性：2 名）のグループと発声・発話トレーニングを受けている TV アナウンサー3 名（29～52 歳，男性：2 名，女性：1 名）のグループである．

また，ヒトの復唱能力（精度）を算出するために，復唱課題の文章と実験協力者の応答による復唱文に対して，それぞれ形態素解析を施した．復唱精度は，復唱課題の原文と被験者の復唱文の相違を比較し，形態素の一致する割合を下記の式（1）より数値化した．

$$\text{復唱精度} = \frac{N - D - S - I}{N} \times 100\,[\%] \quad \cdots (1)$$

ここで，N は復唱課題文（正解文）の総形態素数である．D，S，I は，正解文に対する誤りの種類で，脱落誤り，変換誤り，挿入誤りを意味する．復唱課題には，同時通訳に

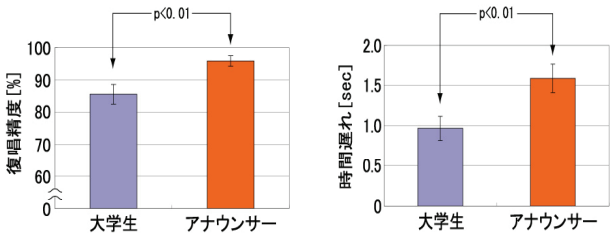

Fig. 2　Rate of correct response (left) and time delay (right) on the re-speaking task.　(left bar: students, right bar: TV announcers)

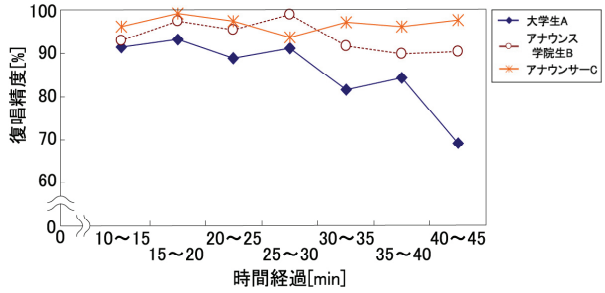

Fig. 3　Time course of correct response rate on the re-speaking task. (A: student, B: announcer candidate, C: TV announcer)

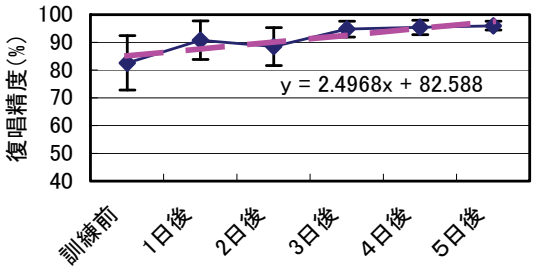

Fig. 4　Improvement of re-speaking abilities using delayed auditory feedback (DAF) effect.

よる国際会議の講演禄を編集したものを用いた．

実験結果を Fig.2 に示す．復唱精度は，大学生は 86%であるのに対して，アナウンサーは 96%であった．この 2 つのグループ間の違いを形態素レベルで比較してみると，復唱精度では脱落誤り D に大きな差があった．具体的には，大学生の脱落誤りの傾向として，同時通訳者の話速が増加した部分で多く見られた．また，その脱落は，連続して広範囲に及んでいた．これに対し，アナウンサーでは，「は」「を」「に」のような，いわゆる助詞の部分に若干の脱落が見られる程度であった．さらに，各被験者が復唱課題の音声を聞いてから復唱するまでの遅延時間を測定したところ，大学生が 0.7～1.2 秒であるのに対し，アナウンサーは 1.3～1.7 秒であり，アナウンサーの方が復唱するまでの遅れ時間が長く，復唱に溜があった．

つまり，アナウンサーは，復唱する内容を記憶・理解しながら処理しているために，大きな文章の脱落は少なく，一方の大学生は，オウム返しのように聞くと同時に処理しているために復唱に余裕がなく，少しの聞き逃しでも意味理解が混乱し，大量の脱落誤りにつながるのではないかと推察される．

3-2 復唱精度の時間推移

上記の実験過程のなかで，被験者より「復唱作業は疲れる」という内観報告が得られた．復唱による音声認識装置への入力作業は，「聞く」と「話す」という双方のタスクに常に集中しなければならないため，適度なリフレッシュ（休憩）を必要とする．このことから，復唱者は，複数名とし，交代制で行うことが望ましいと考えられる．そこで，次の実験では，復唱精度および音声認識率の時間推移を調べ，復唱者の交代時間の目安を探った．

本実験における復唱課題としては，約 45 分間の国際会議での講演（同時通訳・日本語）を用いた．実験協力者は前述の実験に参加した，大学生（33 歳，男性），アナウンス学院生（27 歳，女性），TV アナウンサー（36 歳，男性）である．

実験結果を Fig.3 に示す．アナウンサーとアナウンス学院生では，ほぼ一定の復唱精度を示した．一方，大学生に関しては，25 分あたりを経過した頃から徐々に復唱精度が低下する傾向が見られた．実験後の内観報告では，アナウンサーとアナウンス学院生からは「長時間になると辛い」という回答が，大学生からは「20 分が限界」という回答が得られた．

以上より，復唱者の交代時間の目安としては，20 分以内に設定することが望ましいということがわかった．

3-3 復唱トレーニング法の考案

一見，同時復唱は他人の話す内容を「オウム返し」する作業であるため，誰もが簡単にできると考えてしまう．しかし，先のアナウンサーと大学生の違いからもわかるように，その能力には個人差がある．また，私たちの予備実験から，同時復唱の苦手な人は聴覚遅延フィードバック（DAF）に対する感受性の高いことが示されている[2]．そこで，音声同時字幕システムの性能をヒト（復唱者）の側から高めていくために，DAF に着目した同時復唱トレーニング法を考案し，その効能を実験的に調べることにした．なお，DAF とは，自分の音声を遅延させて聞く状態のことであり，そのような状況では流暢な発話が阻害されやすい．

実験協力者は，聴覚言語的な既往症のない大学生 3 名である．トレーニング課題は，自分の音声を 0.2 秒遅延させた状態で，平易な文章（例えば，昔話の「ももたろう」「かぐやひめ」など）の音読を毎日 20 分間ほど実施するといった単純な課題である．

実験結果を Fig. 4 に示す．復唱精度の経時的変化から，初回の訓練効果が大きく，被験者のばらつきは訓練毎に小さくなることがわかった．さらに，5 日間のトレーニングで復唱精度が約 15%向上することを確認した．また，この訓練効果は，その 1 ヶ月後にも維持されいることを確認している．

以上より，同時復唱者の安定した人材の確保と育成のためには，まず，簡単な DAF 感受性テストを行い，同時復唱の得手・不得手を簡便にスクリーニングし，さらに不得手な人に対しては毎日 20 分程度の DAF 環境下で発話トレーニングを課すことで，ある基準以上の同時復唱能力を備えた人材の安定確保につながると考えている．

4. 顔情報の併用効果

どんなに優秀な復唱者を介在させても，特定話者認識の整った音声入力の環境を整備しても，100%の認識精度を実

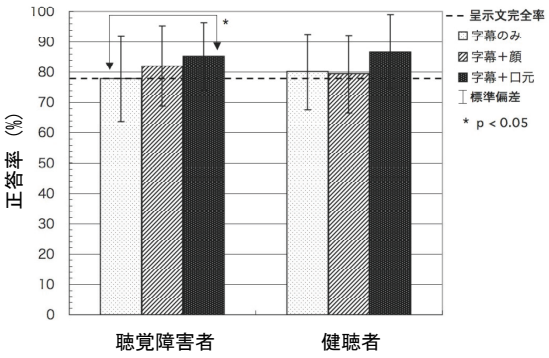

Fig. 5　Prediction ability on incomplete sentences prewith facial information.

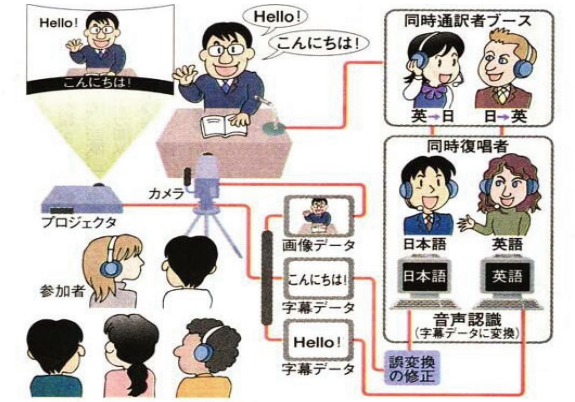

Fig. 6　Illustration of the real-time captioning system using speech recognition technology with re-speakers and simultaneous interpreters.

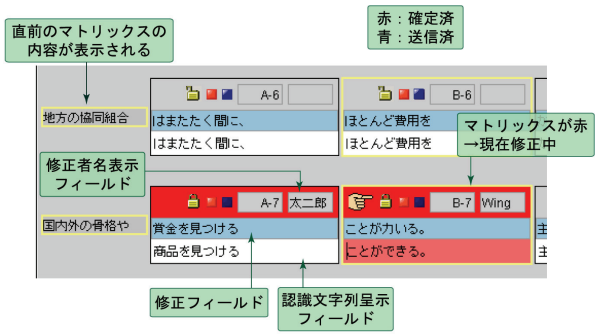

Fig. 7　Character error correcting interface of the real-time captioning system.

現することは並大抵でない．一方，ヒトのコミュニケーションについて考えてみると，聞き手は，話し手の音声のみから会話の内容を認識している訳でなく，顔の表情・口の動き・ジェスチャーなどの様々なノンバーバル情報を巧みに利用しながら，複合的に内容を理解している．

そこで，字幕には誤りを含んで表示される場合があることを前提としながら，不完全な字幕でも本来の意味を推測しやすい情報表示方式のヒントを探る実験を行った．具体的には，顔や口の映像の付与が，不完全な字幕の理解向上にどのくらい寄与するのかについて調べた．また，そのときの口の動きと字幕表示の時間的ずれの影響についても調べた．実験協力者は聴覚障害者 5 名，健聴者 2 名である．

実験結果を Fig. 5 に示す．これらより，聴覚障害者の場合，字幕に口元の映像情報を追加することで，字幕のみと比べて，類推による正答率が約 5%上昇することがわかった．また，字幕と顔情報の呈示タイミングは，「時差なし」「字幕先行」が「顔先行」よりもわかりやすく感じることがわかった[(3)]．

以上より，音声同時字幕システムに顔や口の動きの情報を映像で重ねて呈示する機能は，聴覚障害ユーザに対する内容理解の促進に役立ち，それと同時に字幕に対する顔映像の重ね合わせのタイミングへの配慮の必要性も明らかになった．

5. 国際会議向け音声同時字幕システムの開発

5-1 システム構築

上記の実験結果をふまえて，講演者と音声認識装置との間に同時復唱者を配置し，その復唱入力により字幕を作成するシステムを試作した．

試作システムの構成図を Fig. 6 に示す．本システムでは，講演者と音声認識装置の間に同時復唱者を配置し，講演者の話した内容は全て復唱により音声認識装置へ入力され，その認識結果が字幕データとなる．ここでは，国際会議にも対応可能であるように，復唱者は日本語と英語の両方に配置した．復唱者は講演者の話す言語に応じて，講演者，または同時通訳者の音声を復唱することになる．なお，これらの音声同時字幕システムは，技術の汎用性と拡張性を考えて，音声認識エンジンには比較的ポピュラーな IBM 製の ViaVoice（日本語版・英語版）を利用した．

また，字幕データをスクリーンに表示する場合，講演者の上半身の映像といっしょに呈示する．これは，先の実験結果で確認したように，聴覚障害者にとって，口の動き・顔の表情・ジェスチャーといったノンバーバル情報は意味理解に重要であると同時に，誤認識された部分の正しい意味を推測するために有効な手助けとなるからである．

さらに，会議という公共の場での情報提供という性格上，できるだけ正確な字幕の呈示が望まれる．そこで，同時復唱者の音声認識の誤変換箇所を外部から強制的に直すことのできる修正インタフェースを作成した．その PC 画面の一部を Fig. 7 に示す．

この修正インタフェースでは音声認識による字幕データが修正者サーバ PC へ自動的に送られ，修正者はサーバに接続している修正用 PC（修正者クライアント PC）上で字幕データを共有しながら修正する．4（＋1）人で 1 チームを構成し，修正者が各々の縦の列を専属的に担当し，誤変換箇所があればキーボード入力で即座に訂正する．なお，音声認識の字幕データは，修正者が瞬時に担当箇所を把握できるように，文節単位（7 文字程度）で文字が修正セルに自動的に入力される機能を付与している．なお，修正者には講演者の音声を 3～4 秒遅らせて聞かせ，字幕データと音声の照合を確実にする工夫を取り入れている．

5-2 国際会議での運用評価

試作した字幕システムの運用評価試験を実際の国際会議（DPI 世界会議）の場で実施した．上記のシステム構成による実運用評価結果は以下の通りである．同時復唱による復唱精度と音声認識率は 97%と 95%であり，共に高い値を示した．また，修正前後におけるスクリーン上の字幕表示精度（発言内容と字幕データの形態素の一致率）を比較し

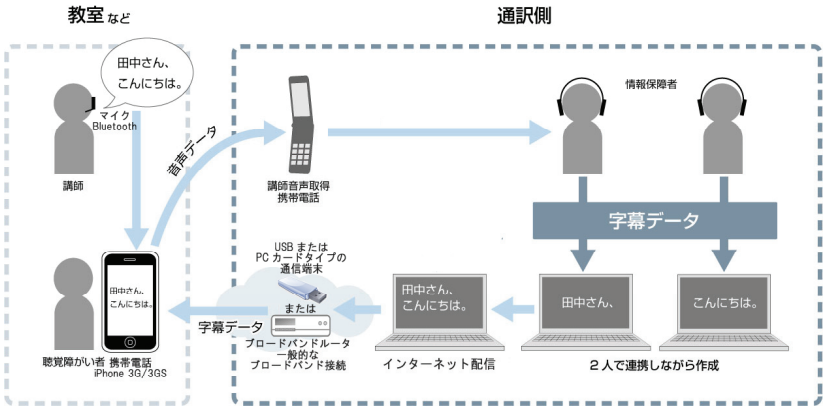

Fig. 8　Illustration of the mobile captioning system with a smartphone and a Bluetooth microphone.

てみると，修正前の 91%に比べ，修正後は 96%であり，約 5%上昇した．リアルタイム性については，修正作業が加わることもあり，講演者の発言からスクリーンへの字幕の表示までに 5～10 秒程度を要した．

本会議の参加者に音声同時字幕システムの感想を聞いたところ，「バリアフリー（情報保障）の観点からとても有用である」という意見が全体から得られ，一方では「誤字や表示の遅れが気になる」という意見もあった．しかし，「多少の誤認識がある方が楽しく字幕をみられる」（中途失聴者）というユニークな感想や，さらに「メモ取りの際に役立つ」（健聴者），「顔と文字が大きく見えるのがよい」（高齢者）などのユニバーサルデザイン的な評価も得られた．

6. モバイル型遠隔情報保障システムの開発

6-1 システム構築

最近，音声通話に加えて，本格的なネットワーク機能や様々なアプリケーションソフトを利用できるスマートフォンが携帯電話市場で広がっている．これらには，音声通話とウェブなどのインターネット機能を同時に利用できる機種がいくつかあるが，その中でも普及率が急上昇している iPhone（Apple 社製）を対象機種に選び，手軽に利用できる音声字幕システムとして「モバイル型遠隔情報保障システム」を開発した[(4)]．

本システムの主たる特徴は，前述の国際会議向け字幕システムと違って，字幕表示と音声通話を 1 台の携帯電話で担当させることによって，字幕サービスの利用者側に必要な機器構成（iPhone と Bluetooth 対応のマイクロホンのみ）を限りなくシンプルにした点にある．また，運用コストに関しては，サービス提供に要する全てのデータ通信（文字と音声）を定額制サービスに含めることを条件にして（G3 回線と無線 LAN の併用）設計し，ユーザの経済的な負担ができるだけ少なくなる字幕システムとした．開発したシステムの概略図を Fig. 8 に示す．

6-2 教育現場での運用評価

本システムを聴覚に障害のあるユーザを対象として，小・中学校の授業や遠足，大学生の講義や学外の見学会，そして，社会人に対する屋外でのセミナー活動などの様々な現場で利用し，聴覚障がい学生（ユーザ）および字幕サービスを提供する情報保障者の立場から見た運用面での評価をアンケート方式で実施した．その結果，ユーザ側である学生からは，「講師等が発話した内容の理解に役立った」「字幕による情報保障を受けづらい状況（体育や見学など）で役立った」との使用感の評価が得られ，サービスを提供する側の情報保障者からは，「（屋内外を問わず）移動を伴う状況下や初等中等教育の場での心理的配慮を要する場面（情報保障者が教室などに入ることで生徒らに何らかの心的負荷を与えてしまう状況）で効果的である」との評価であった．以上から，モバイル型情報保障システムに対する現場のニーズとその重要性を確認できた．

7. まとめ

聴覚に障害のある人たちの情報保障サービスを目的とした音声同時字幕システムの研究開発について，その基礎から応用について述べた．ここでは，人間と機械の協調作業を基軸に据えて，実用性を重視する現場対応のシステム構築をそのデザインポリシーに挙げ，同時復唱や顔情報の活用などの応用人間工学的な視点でのシステム開発とその運用評価を行った．さらに，字幕サービスに対する教育現場（学校における授業や課外活動など）のニーズからは，普及の伸びが著しいスマートフォンを利用したシンプルな構成のシステム開発について紹介した．

今後は，ユーザの利用条件に応じた音声同時字幕システムのバリエーションの拡張と情報保障者の人材育成などを含めて，常に現場のニーズを大切にした研究開発を推進していくと共に，映画館[(5)]などの生活に潤いを与える芸術・文化の場における利用も視野に入れた多様なユビキタス情報保障システムに発展させていきたいと考えている．

参考文献

(1) 井野，情報バリアフリーとVR－聴覚障害者のコミュニケーション支援技術－，日本VR学会誌, vol. 8, no. 2, pp. 70-75, 2003.
(2) 高橋，加藤，井野 他，同時復唱を利用した不特定話者の音声認識－復唱訓練方式について－，日本音響学会聴覚研究会資料, vol. 33, no. 1, H-2003-02, pp. 1-8, 2003.
(3) 黒木，井野，中野 他，音声同時字幕システムにおける内容理解の向上を目的とした話者の顔情報の呈示方法，HI学会論文誌, 2007.
(4) 三好，河野，白澤 他，聴覚障がい者のためのモバイル型遠隔情報保障システムに関する評価,第61回HI学会研究会, vol. 12, no. 3, pp. 1-6, 2010.
(5) 井野，映画のバリアフリーと複合現実感ハードウェアの関係，平成20年度厚労省障害者自立支援調査研究プロジェクト「バリアフリー映画製作事業報告書」, p. 40, 2009.