4.2. コミュニケーションのための代行技術

4.2.2.視覚による聴覚代行

中途失聴者にとっては、手話や読話など主に先天的な重度難聴者が使っている言語手段を覚えるのは大変な努力が要る。したがって音声を文字にして見せる筆談やそれを代行するような音声認識装置が役に立つ。音声が文字として提示されれば、それから記憶している音声言語を惹起させることができるからである。最近の人工知能（AI）とデータベースの著しい発展と拡大により、音声認識精度は急速に上がってきている。とくに「生成 GPT（Generatable Pre-learning Translation)」の登場により、質問に対して AI が的確な答えを正しい文法で出力するようになってきた。音声認識技術の進歩は、文字言語の概念や文法を獲得してから失聴した後天的な聴覚障害者にとってはこの上ない福音になろう。このことについては、「音声字幕システム」の項で述べたい。

ただし、音声言語の概念を獲得する以前に失聴した先天性の場合には、脳の中に音声情報の記憶やそれによって構築される言語体系も有効に活用することはできない。そのため、言葉の組み合わせによる抽象的な思考をすることが難しくなり、「９歳の壁」といわれるように、９歳程度の思考能力で止まってしまう恐れがあった。現在は、それを超えるための教育方法が採られているが、全ての聴覚障害者に対して、音声をそのまま文字にして与える方法が良いかというと、そう簡単なことでもない。本項の前半では、音声を視覚的なパターンに変換する方法について述べ、後半では聴覚障害者用の音声字幕システムについて述べる。

（１）音声情報の視覚提示

①　歴史

音声認識技術の研究とは別に、古くから音声を目で見えるパターンに変換して、聴覚障害者の発声訓練や音声認識の補助として利用しようという研究が行われていた。音声を視覚パターンにして提示するアプローチでは、いうまでもなく視覚の情報受容能力が大きいことに期待している。確かに視覚は指先などの限られた触覚に比べると、時間分解能は劣るものの、２次元あるいは３次元画像として情報を受容する能力は桁違いに大きい。視覚を有効活用するためには、時間的に変化する情報を電光掲示板のように、２次元的な情報に変換して直観的に見せる方法がとられる。

1947 年にポッター（Potter, R.）らによって考案されたビジブルスピーチ、すなわち音声の時間スペクトログラム（声紋）を利用することが提案されていた（文献 4.2.16)。ところが、幾人かの研究者によって、視覚は聴覚に存在する音声解読機構のようなものが欠如していることから、音声スペクログラムは、いかに画像的加工を施そうと、また、いかに熱心に訓練を行おうと本質的に読めないだろうと指摘されている（文献 4.2.17)。このような論争が続いている中で、ビジブルスピーチの研究は、音声の特徴となる成分、例えば/n/、/m/

のような鼻音成分や/s/、/sh/のような摩擦音成分などだけを抽出し識別しやすいパターンに変換して見せるディスプレイなど、支援方式は形態を変えて進められた。

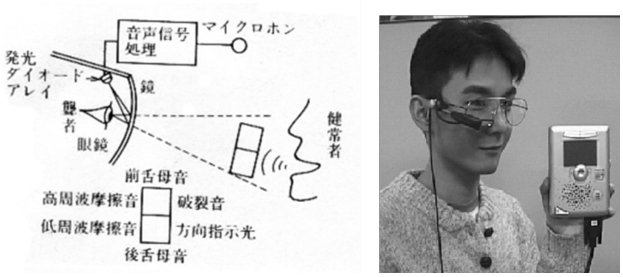


図 4.2.15　左：音声の特徴情報をメガネを介して話し手の口元に提示する「アプトン眼鏡」、右：単音節音声タイプライタとウェアラブル眼鏡ディスプレイを接続して話者の口元に文字を表示する機器

また、視覚を利用する方法は、舌、顎、唇などの正しい動きを教えるための発声・発話訓練に発展していった。米国の MIT では、ヘレン・ケラーの音声取得法に着目し、話し手の口元に取り付けたセンサで口唇、鼻、喉からの情報を検出し、それを視覚で分かるようなパターンに変換して見せる「タドマ法」を開発している。眼鏡の縁に８の字型に LED ディスプレイを付けて、検出した音声の特徴要素で LED を点滅させる「アプトン眼鏡」なども開発され（図 4.2.15 **の左**）、「読話」と併用して使用された（文献 4.2.18）。日本では、似鳥・伊福部らが、１音１音区切って話した単音節音声を自動認識によりかな文字にし、さらに日本語ワープロで漢字混じり文に変換する「単音節音声タイプライタ」を開発し、しばらく印刷会社などで使われた（文献 4.2.19）。さらに、それとウェアラブル眼鏡ディスプレイを接続して、話者の口元に認識された文字を移す機器を開発したことがある（図 4.2.14 の右）。1980 年半ばの頃は、眼鏡ディスプレイは高価であったので実用には至らなかったが、最近は、VR（バーチャルリアリティ）技術の進歩により、高解像度のものが安価で手に入るようになったことから、これらの表示方式も見直されてきている。

一方、日本では古くから、渡邊、上田らによって、目の錯視を利用して、視覚でもセグメンテーションがしやすくなるように工夫した一種のビジブルスピーチを提案していた（文献 4.2.20）。方式としては、音声中の第１から第３ホルマント周波数を「３原色」に変換し、ホルマント情報を画面の下から上へ流れるように提示する。すると、対比効果により色が変化する中間のところが強調され、例えば/ai/では/a/と/i/が分離して見えるという効果が得られる(図 4.2.16)。評価実験から、３母音連鎖を読み取らせたところ初

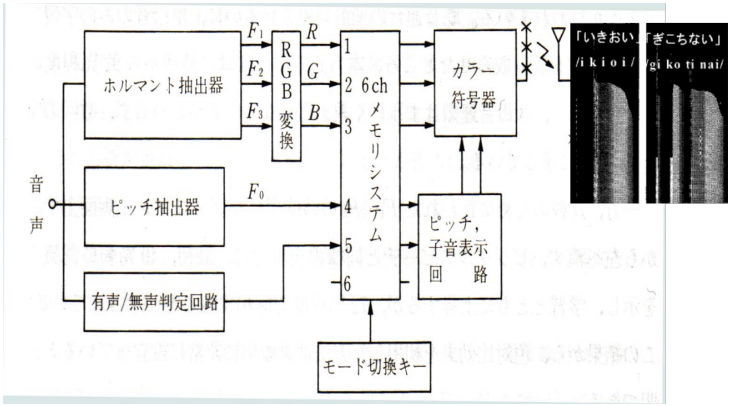


図 4.2.16　目の対比効果を利用した音声視覚表示

期の段階でも数回の学習で 98%の認識率が得られたと報告されている。この方式はその後も改良されており、発音訓練のフィードバックを得たり、訓練結果の評価に利用されたりしている。

手話（sign language）や読話(lip reading)は一種の動画であることを考えると、ビッグデータと深層学習などの人工知能（AI）を活かした情報変換により視覚による聴覚代行の道は十分にあるといえる。さらに、このような技術の延長として、VR 分野で開発された眼鏡型のウェアラブルディスプレイを利用して、音色、音楽、音源方向を視覚に提示できれば、音声ばかりでなく　報音など環境の音を知らせることができるし、音楽を目で楽しむという道も開かれるかも知れない。

なお、本書では詳しくは述べなかったが、手話の認識・合成の研究も盛んに行われ、「手話工学」という分野も生まれている。ただし、手話の認識といっても、手の動きだけでなく顔の表情や口唇の動きが重要な情報になっていること、しかもセグメンテーションをどうするかという難題もあり、現時点では実用化が難しい。NHK の放送技術研究所が取り組んでいるように、手話合成については、人間の類推機能を期待できるので、認識に比べると実現性が高く、放送などで利用する価値がある。今後は、手話の研究は「ジェスチャー認識・合成」という形で、これもビッグデータと AI を活用したヒューマンインタフェースの１つとして発展していくものと思われる。

② 音声認識技術の利用

聴覚障害者が利用する「読話」は視覚による聴覚の代償機能といえるが、極めて少ない情報からでも意味を類推する能力が潜在的にあることを裏付けている。したがって、誤変換の多い不完全な音声認識技術でも、この類推機能が働くように認識結果を提示することによって、言葉の理解に十分に役に立つ。このような観点で音声認識技術を聴覚障害支援に活かす研究もが芽生えていた（文献 4.2.21)。

音声認識は極めて多様な手法が提案されていたが、1960 年代の後半には、発声時間長の伸縮を正規化する DP（Dynamic Programming）マッチング法が考案され、単語単位ではあるが認識精度は大きく向上した。ところが、1980 年代になって、大量の音声データが手軽に扱えるようになり、音声を統計的手法で認識する方法が主流になってきた。その枠組みに登場したのが隠れマルコフモデル(HMM：hidden Markov model)であった。最近は自然言語処理に関する人工知能とも結びついて、障害者・高齢者のコミュニケーション支援に相応しい機能を備えてきている。ただし、HMM を有効に使うためには、いかに正確な「音素モデル」や「言語モデル」を構築するかが鍵を握る。

HMM では、音素間の状態遷移確率をあらかじめ多くのデータから求めておき「音素モデル」として格納しておく。また、音素列である単語や助詞などの「形態素単位」にまとめて、それらの間の状態遷移確率（例えば「私」から「が」に移る確率）を利用して計算を簡略化す

る。形態素間の状態遷移確率もあらかじめ多くのデータから求めておき「言語モデル」として格納しておく。したがって、文章データの数が多いほど音素モデルや言語モデルはより精巧になるので、文章データの数とともに認識率も向上することになる。さらには音素モデルや言語モデルと同様に「文構造」「意味」「文脈」「場面」「感情」に関するモデルを作ることにより、認識精度も上がる。このように、例文となる文章データの数が多いほど音素モデルや言語モデルはより精巧になるので、文章データの数とともに認識率も向上することになる。

近年はインターネットなどで集めた文章データが膨大になり、これを使うことによって各種モデルなどが正確になり、認識結果の精度を大幅に高めることができるようになってきている。それを利用した方法は既にスマートホンのアプリケーションソフト（アプリ）として提供されている。また、認識結果の意味が分からなかったとしても、対話形式で聞き直すことで正しい文章やそこに含まれている意味や感情などを引き出すことができる。さらに、最近の人工知能や「自然言語処理」の技術を使うことで認識精度はあがっており、曖昧な会話音声による対話などにも対応できるようになっている。このような音声認識技術の進歩により、聴覚障害者支援の方法も大きく変わろうとしている。

ただし、日本語文字は脳の中で音声に変えてから理解する「表音文字」と音声に置き換えなくてもいきなり意味が伝わってくる「表意文字」すなわち漢字を併用しており、同じ発音でも違った意味を持つ「同音異義語」が多い。これらの日本語特有の問題は認識精度を落としているが、この問題も人工知能の進歩によりで解決されることが期待されている。

③　音声字幕システム

(i)復唱による音声自動字幕システム

服部・伊福部らは、2002 年に札幌で開かれた障害者インターナショナル世界会議（DPI：Disabled Peoples’International）からの依頼で聴覚障害者のために講演者の声を「復唱」と音声自動認識の組み合わせによって文字にして見せる「音声自動字幕システム(automatic caption system)」の開発を試みた。従来、字幕システムでは、話者の音声をキイボードで入力して文字にする「PC 要約筆記」などを利用しているが、音声自動字幕システムで得られた知見は、聴覚障害者や同時通訳者のためのコミュニケーション支援にも生かされるので、少し詳しく述べたい。

DPI は 4 年に 1 度、世界の主要な都市で開かれる国際会議で、109 の国や地域から約 3000 人が集まる大イベントでもあり、日本では初めての開催であった。当時、不特定話者の音声認識の技術はあまり高くなかったこともあり、話者の声を特定の「復唱者」が復唱してコンピュータに入力することで、結果的に「特定話者音声認識」にするという方法をとった。コンピュータにはあらかじめ復唱者の単語辞書を登録しておき、誤りの修正のたびに単語辞書のみならず音素モデルや言語モデルも自動的に更新されるようにした。この復唱者を介在させる方式は、NHK がスポーツや歌番組などで行っている「リスピーク方式(re-speak

method)」と基本的には同じ考え方である(文献 4.2.22)。図 4.2.17 にネットワークを使った音声字幕システムの概略を示した。

なお、字幕化は日本語と英語のみとし、日英以外の言語はいったん英語に通訳した。予備

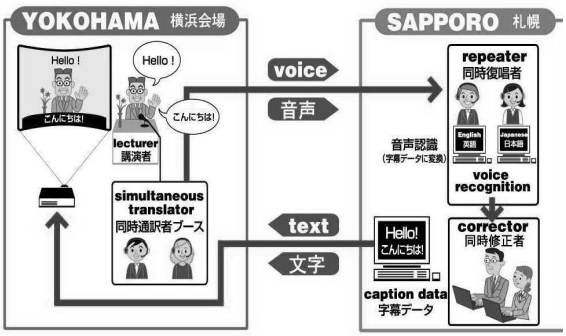


図 4.2.17　ネットワークを利用した音声字幕システム

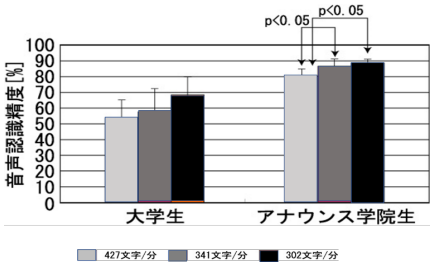


図 4.2.18　話速度および復唱者の違いによる音声認識精度の変化

試験の復唱者として数名の大学院生（5 名）と放送のアナウンサー学院の生徒(3 名)が加わった。両者で、音声認識の精度を比較したところ、アナウンサー学院生が明らかに高かった。とくに図 4.2.18 に示したように、話速が速くなるにつれてその傾向が大きくなった。その理由について調べた結果、脳内の音声処理とくに DAF（Delayed Auditory Feedback）現象と深く関わっていることが推論された。

TV 放送の国際中継では自分の声が長い回線を伝わってくるうちに遅れて自分の耳に入り、そのため吃音になる場合がある。これは DAF による吃音といえるが、これから逃れるために、アナウンサー学院生は復唱した自分の声はできるだけ聞かないようにして、話者の声に集中する訓練を受けている。この訓練効果が復唱にも大きく関与していると想像された。

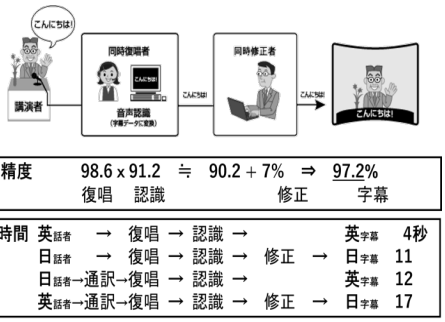
図 4.2.18　左:DPI プレ大会、右:夕張国際映画フェスティバル

図 4.2.18 に、2001 年に札幌で開かれた DPI プレ大会で日本語と英語を提示したり（左)、夕張で開かれた国際映画フェスティバルでさらに韓国語と仏語も提示したりした様子を示した。その後、20 回を超える運用を通して、97%の正答率を得るのに要した時間を調べた結果、図 4.2.19 に例を示すように、英語音声から英語文字が一番速く、英語音声から日本語文字が遅いことが分かった。この傾向は生成 GPI とビッグデータによる最新の音声自動認識においても同様になるといえる。

以上のような経験を通じて、聴覚障害者に音声字幕方式を適用した場合の特有の課題について浮き彫りにされた。

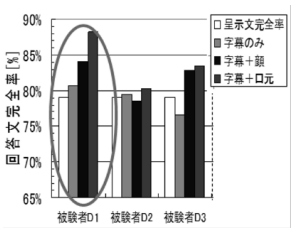
(ii)聴覚障害者の字幕認識の特徴

黒木らは、さらに誤変換のある字幕の認識に読話がどこまで寄与するかを調べている。そのため、①字幕のみ提示、②字幕と一緒に「顔画像」を提示、③字幕と一緒に「口の動き」を提示、の３種類の提示方式を作成し、普段、口話法や手話を使っている聴覚障害者３名(大学生)と、健聴者（大学院修了者、26 歳）2 名に協力してもらい、文章の認識率を求めた（文献 4.2.23)。その結果、「字幕+口元」の場合に全被験者で文章の認識率が最も高かった。その中で、口話法を使っている聴覚障害者 D1 の場合は「字幕+口元」がとくに文理解に貢献しており、１秒ほどの「字幕先行」が好まれることが分かった（図 4.2.20 の上図)。



精度	98.6 x 91.2	≒	90.2 + 7%	⇒	97.2%			
	復唱 認識				修正		字幕	
時間	英話者	→	復唱 → 認識 →				英字幕	4秒
	日話者	→	復唱 → 認識 →	修正	→		日字幕	11
	日話者→通訳→		復唱 → 認識 →				英字幕	12
	英話者→通訳→		復唱 → 認識 →	修正	→		日字幕	17

図 4.2.19 97%の認識率を得るのに要した文字―音声変換時間（国際ユニバーサルデザイン会議の結果か



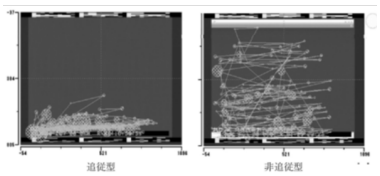


図 4.2.20　上：顔情報の提示による言語理解率の変化、下：聴覚障害者における画面上の字幕文字の注視（停留点軌跡）の例

一方、自らが聴覚障害者である中野らは、字幕が下から上へスクロールするようにして提示したときに、眼球運動がどのように変化するかを計測している(文献 4.2.24)。そこから注視点を予測した結果、図 4.2.20 の下図に例を示したように、聴覚障害者（25 から 37 歳の先天性難聴８名）は上方に大きく広がる「非追従型」が多いことが分かった。

さらに、詳細に調べると、健聴者の場合は「、」の場所で停留することが多いが、聴覚障害者ではその傾向が見られなかった。「、」は健聴者にとっては内言語では「間」を置く場所とみなすが、とくに先天性の聴覚障害者にとっては内言語を獲得することが難しいので、「間」を置くという意味が分りづらかったのであろう。「、」のある場所を改行に変えると聴覚障害者は遥かに読み取りやすいと答えた。最近、テレビなどで字幕を付けるのは当たり前のように身近になってきているが、健聴者と聴覚障害者とでは字幕の読み取り方に違いがあるので、このことを考慮した提示方法をとるべきであろう。

(iii)実用化・ビジネス化に向けて

音声字幕システムは、インターネットがあれば講演者、通訳者、復唱者やPC要約筆記者などの支援者および聴衆がどこにいて字幕化できることを意味している。したがって支援者が家にいてもよいことから、支援そのものが在宅ビジネスになり、外に出ることの少ない障害者の新しい雇用につながる可能性も出てくる。また、リアルタイム性と活用シーン（理解補助、情報保障、記録/データ）の観点から考察すると、音声字幕は主に「理解補助」と「リアルタイム性」を重視する会議や授業などで効果があることが想像される。

実際、三好らは、音声認識技術の代わりにPC要約筆記を活用して聴覚障害者が受ける授業の補助として活用し、その有用性を実証するとともに（文献4.2.25)、「PEPネット」と呼ばれる音声字幕システムを実用化している。また、河原らは、国会などで速記の代わりに字幕システムを活用し、高い評価を得ている。前述のように、インターネットから得られる文章のビッグデータと人工知能による推論機能の進歩により音声認識技術は新しい局面を迎えている。さらに話者の音声から「感情」を抽出する方法も研究されており、その研究は認知症者など脳機能障害者の生活を支援するロボットとのコミュニケーションに有効でないかと期待されている。（文責　伊福部　）

文献（4.2　続き）

(4.2.16) Potter, R.K., Kopp, G.A., Green, H.C., “Visible Speech” (Van Nostrand, NewYork,1947)

(4.2.17) Liberman, A.M., Cooper, F.S., Shankweiler, D.P., “Why are speech spectrograms hard to read?” Am. Ann. Deaf 113,127–133(1968)

(4.2.18) Upton, H.W., “Visual speech reader design considerations”, in Preprints of the Research Conference on Speech Processing Aids for the Deaf, Gallaudet College(1977) 5.

(4.2.19) 伊福部　「音声タイプライターの設計」,ＣＱ出版 (1983)

(4.2.20) 渡遺 亮，上田裕市:“連続音声の色彩表示システムにおける母音連鎖の視覚的イメージ"，信学論，J-64 A，pp.574-581 (1981)

(4.2.21) 篠原正美:“視覚障害者用文字・音声変換システム"，音響会誌 43(5)pp.336-343 (1987)

(4.2.22) 中村章，清山信正，池沢龍，都木徹，宮坂栄一:“リアルタイム話速変換型受聴システム"，音響会誌，　50(7) pp.509 520 (1994)

(4.2.23) 黒木速人，井野秀一，中野聡子，堀耕太郎，伊福部　:”聴覚障害者のための音声同時字幕システムの遠隔地運用の結果とその評価”、ヒューマンインタフェース学会論文誌, 8(2) pp.225-262（2006）

(4.2.24) 中野聡子，牧原功，金澤貴之，中野泰志，新井哲也，黒木速人，井野秀一，伊福部　:"音声認識技術を用いた聴覚障害者向け字幕呈示システムの課題－話し言葉の性質が字幕の読みに与える影響－"，電子情報通信学会論文誌 D, Vol.J90-D(3) pp.808-814 (2007)

(4.2.25) 三好茂樹,河野純大,白澤麻弓,磯田恭子,蓮池通子,小林正幸,小笠原恵美子,梅原みどり,金澤貴之,中野聡子,伊福部　:“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の提案と情報保障者による評価”,ライフサポート, 22(4) pp.11-16 (2010)