■2群（画像・音・言語）－7編（音声認識と合成）

# 4章 音声応用

（執筆者：）

■概要■

【本章の構成】

■2群 － 7編 － 4章

# 4-1 音声対話／音声ドキュメント処理

（執筆者：中川聖一）［2009年7月 受領］

音声認識技術と音声合成技術を統合した代表的な応用システムは，人間と機械（ロボット）との音声による対話システムと音声翻訳システムである．後者は自然言語処理による機械翻訳が中心技術となるので，ここでは述べない．音声認識技術の代表的な応用システムは音声ドキュメントの検索システムやブラウジングシステムである．

## 4-1-1 音声対話システム

典型的な音声対話システムの構成図を**図 4･1** に示す．主な構成要素は音声認識部，音声理解部，対話管理部，問題解決部，応答文生成部，音声合成部である．

音声認識部と音声理解部は，話し言葉特有の現象である助詞落ち，間投詞，倒置，言い直し，言い淀みなどを含む入力音声を認識し，これに認識誤りが加わったものを正しく理解する部分である．対話管理部は，対話の履歴を管理し（通常はスタックを使用することが多い），現時点での対話の状態をもとに，次発話の予測や照応・省略表現の解決，不足情報の要求を行うものである．一般に，対話システムは，大規模な知識データベースを参照しながらユーザの質問に答えるための問題解決部が必要である．更に，適切な応答文を合成するための文生成部とイントネーションや感情を付与する音声合成部（概念からの音声合成）も必要である．

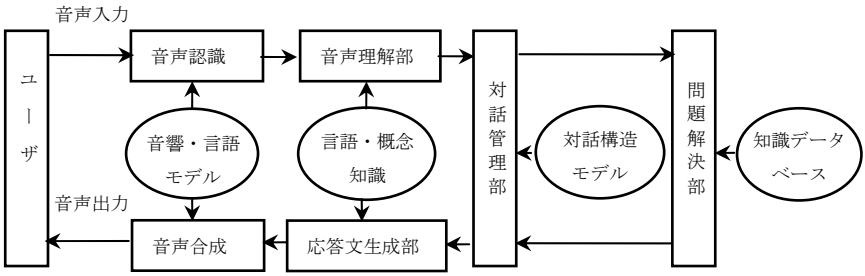


**図 4･1　音声対話システムの構成図**

対話構造の特徴としてグライスの対話協調（Grice's maxims）の原理（必要なだけの情報を伝える，必要以上の情報は含めない，偽であると信じることは言わない，など）などをもとに対話の管理を行い，文脈を参照しながら意味表現を形成していくのが一般的である．しかしながら，対話構造のモデル化とその利用は比較的難しい．最近では，大規模な対話コーパスから部分観測マルコフ決定過程を機械学習でモデル推定し，利用しているものがある．

音声対話は，即興的に行われるので，相当「さぼった」発話がなされる．このため，音声，言語，概念，知識のいずれかにおいても，誤り，あいまいさ，省略などの不確実さが多くなる．したがって，音声対話の研究では，書き言葉とはかなり異なったアプローチにより，各

レベルを密接に関連させて，一体的に行う必要性がある．

通信衛星を介した国際通話で，少しでも遅延があるとスムーズな対話が実現されないことからも明らかなように，「オンライン・リアルタイム，漸次的発話・生成，割込み・相づち」が音声対話では重要である．このほかに，使い勝手の良いシステムとしてユーザインタフェースの設計が重要である．シュナイダーマンが与える対話設計における八つの黄金律 [1] を，特に音声対話システムに勘案して以下にまとめる [2]．

1. 似たような状況では一連の手順に一貫性を持たせる，自由度を与えすぎない（ただし，代名詞や省略の使用，多様な言い回しは許可）
2. 頻繁に使うユーザには近道を用意する．簡略表現やマルチモーダル入力，代替入力を許す．例えば，エキスパートはユーザ主導型の対話システムを好み，初心者はシステム主導型の対話システムを好む
3. 有益なフィードバックを提供する相槌や確認（ユーザからの応答がない場合のプロンプト・ヘルプ機能）
4. 段階的な達成感を与える対話を実現する途中確認の表示，対話履歴の表示．漸次的な発話を許す
5. エラーの処理を簡単にさせるリジェクト機能・確認機能，簡単な修正・再入力法．
6. 逆操作を許す．間違った発声に対するキャンセル機能，直前の発話状態への移行
7. 主体的な制御を与えるユーザ主導の対話制御法やユーザとシステムの主導切り替え（混合主導）
8. 短期記憶の負担を少なくする途中結果の表示と音声合成以外の応答表示機能，漸次的文生成

音声対話システムは，対話の主導権により，システム主導，ユーザ主導，混合主導に分類される．ディスプレイを使用するかどうか，ディスプレイに擬人化エージェントを表示するかどうか，タッチ入力などのほかのモダリティと併用するかどうかによっても対話システムの構成は異なってくる．また，情報のフローによって次のように分類される [3]．

- 説明・メニュー選択型（システム→ユーザが主）
- フォーム入力型（ユーザ→システムのみ）
- 検索・相談型（ユーザ⇔システムの双方）
- 共同作業型（ユーザ⇔システムの双方）

なお，メニュー選択型やフォーム入力型などの比較的簡単な音声対話システムは，Voice XML を使用して音声認識，DTMF（電話のプッシュ音によるトーン信号），キー入力，録音，音声合成などを組み合わせて開発することができる．

### 4-1-2 音声ドキュメント処理

音声認識の応用としてまず考えられたのが音声ワープロである．1980 年前後，各大手電機メーカーは音節ごとに区切って入力する「音声ワープロ」の開発に凌ぎを削ったが，一般の人には受け入れられなかった．その後「単語単位発声ワープロ」「文節単位発声ワープロ」「連続発声ワープロ」と技術開発は進んだが，いまだに受け入れられていない．音声は人と人とのコミュニケーション手段であって，機械へのデータ入力手段としては違和感があり，キーボード（タッチキー）に負けている．これは認識精度だけの問題でなく，機械に向って話し

かける習慣がないこと，人に聞かれるのに抵抗があること，人に騒音として迷惑をかけること，などいろんな問題があり，実用化を予想するのは難しい．バンキングシステムの音声入力が手作業中の音声制御コマンドとして実用化されたが，一部の応用に限定されてきた．最近では，カーナビのインタフェースとして音声入力が標準装備されつつあるが，普及しているとは言いがたい．また，ロボットとの音声対話も有望と言われながら，ロボット自体が普及していない．

一方，インターネット上に流通しているコンテンツのうち，音声の占める割合は年々大きくなっている．例えば，動画の検索をする場合，そこに付随している音声をキーとして検索するのが有用であると考えられる．検索では，音声認識率が 70～80%程度でも十分可能であり，また，この程度なら，認識結果から内容の理解ができるため，新たな応用が開けると考えられる．図 4･2 のように音声認識が役立つ音声ドキュメントは多岐にわたっている．音声ドキュメント処理には，音声認識，検索，要約，整形，メタデータ付与などが重要な技術である [4]．特に，大規模音声ドキュメントからのキーワード発声部分の検索は，音声認識辞書（通常，2～8 万単語）にない未知語（主に固有名詞などで検索には重要）が検索キーワードの場合は難しく研究課題となっている．

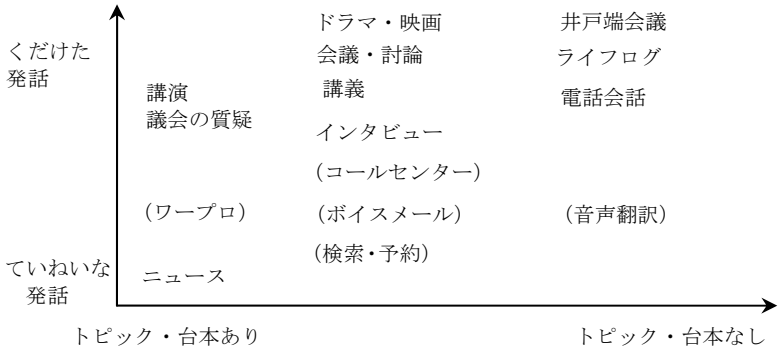


**図 4・2　音声ドキュメントの分類（括弧内は対機械入力）**

**■参考文献**

1) B.シュナイダーマン（東，井関 監訳），“ユーザインタフェースの設計，”第 2 版，日経 BP 出版，1995.
2) 中川聖一，“小特集に寄せて―音声対話システム構築の課題―，”日本音響学会誌，vol.54，no.11，pp.783-790, 1998.
3) 河原達也，荒木雅弘，“音声対話システム，”オーム社，2006.
4) 中川聖一，“音声ディクテーションから音声ドキュメント処理へ，”日本音響学会講演論文集 vol.1-3-1，Sep. 2007.

■2群 － 7編 － 4章

# 4-2 車載型音声インタフェース

（執筆者：坂井　誠・武田一哉）［2009年9月 受領］

音声は，誰もが自然かつ容易に使うことができるコミュニケーション手段である．音声を機械が認識できれば，人間と機械をつなぐ自然なインタフェースの実現が期待できる．更に，音声には手や目が離せない状況下でも使うことができるという大きな特長がある．このような利点を有しているため，音声認識を車載機器操作へ応用すれば，運転時のように手や目が離せないときでも車載機器の操作が安全に行える．そのため，車載機器，その中でも特にカーナビゲーション（以降，カーナビ）の操作に音声認識が応用され広く実用化されている．

## 4-2-1 カーナビゲーションの市場と音声認識の実用化

カーナビは1980年代に実用化が始まり，1990年代には一般に普及していった．2007年には市場規模が6300億円を超え [1]，累計出荷台数は3000万台を超えた [2]．カーナビは目的地まで自動で案内してくれるなど利便性が高いが，一方で，操作の際にはドライバーの視線が前方からそれたり，ハンドルから手が離れるなど危険な状況になることが早くから問題視されていた．そのような中，音声認識はハンズフリー，アイズフリーで安全なインタフェースとして注目が集まり，1995年にカーナビで実用化された [3]．

実用化初期は，音声認識の処理を専用ハードウェアで実現していたため高価格となり，一部の高級車のオプション機能として提供されていた．その後，専用ハードウェアからカーナビに内蔵されている CPU を共有できるソフトウェア処理に置き換わることで安価に提供できるようになり，徐々に普及が加速していった．現在ではほぼすべてのカーナビに音声認識が標準機能またはオプション機能として搭載されている．

## 4-2-2 音声認識サービスの形態

音声操作が可能な初期のカーナビでは，事前に自分の声を登録しておき，登録音声と入力音声とのマッチングを行う特定話者認識方式が採用された．しかしながら，事前に音声やキーワードを登録する手間があり，音声認識を使うまでが煩わしいという課題があった．その後，ユーザの音声登録が不要な不特定話者認識方式が実用化された．実用化当初の認識語彙は，数十から数百語の機器操作コマンドのみを認識するだけであったが，その後，施設名称のような POI（Point Of Interest）入力，地名入力，電話番号入力，郵便番号入力など，認識可能な語彙数を増やして利便性が高められてきた．現在，POI 検索では数万から数十万の，地名入力では数十万から数千万の語彙を認識できる製品が実用化されている．

また，カーナビの記憶媒体が HDD になると，大量の音楽を HDD 内に格納できるようになり，カーナビから再生できるようになった．CDDB などから楽曲情報や音声認識に必要な楽曲の読みが自動で付加されるようになったこともあり，所望の音楽を簡単に再生するために音声認識による楽曲検索が実用化されている．

## 4-2-3 応用上の課題

音声認識はカーナビへの応用から10年以上経過し，その間様々な改良が施されてきた．し

かしながら，いまだユーザが十分満足して使えるインタフェースには至っておらず，いくつかの課題が指摘されている．

カーナビへ音声認識が応用された主目的は音声によるカーナビの操作を実現することである．そのためカーナビの誤操作につながる音声の誤認識は誰のどのような発話に対しても基本的には許されず，音声認識には高い認識性能が要求される．しかしながら現在の音声認識はまだその要求を満たすレベルには至っていない．今後は不特定話者認識技術の更なる開発が必要となる．一般に，音声認識性能は認識語彙数が少ないほど高くすることができる．しかしながら，少ない語彙数では操作できることが限定されるので，魅力的なインタフェース実現のためには音声認識は大語彙化されていく．そのため，大語彙認識でも性能の劣化が少ない性能向上技術が必要となる．更に，車室内には，走行ノイズ，エンジンノイズ，風切り音，ウィンカーやハザード音など様々な雑音の混入がある．このような中でも高い認識性能を確保するためには，高精度な雑音除去手法が必要である．それ以外にも，人間が音声を発する際は，無意識に「えー」や「あのー」などのフィラーが混入されることがある．そのためフィラーに対しても頑健な音声認識が要求される．このような様々な要求に対し，カーナビに用意されている計算資源（CPU，メモリ）は限られているため，いかに少ない計算資源で高い認識性能を確保する手法を開発するかが音声認識の大きな課題となる．また海外に目を転ずると，ヨーロッパなど多様な言語が陸続きで共存している地域が存在する．このような地域ではある特定の言語だけ認識できるのではなく多言語を認識可能にすることが必要になる．

### 4-2-4 カーナビ以外の車内音声認識アプリケーション

カーナビへの応用以外にも音声認識は車載型音声インタフェースとしていくつか実用化されている．例えば，車室内でハンズフリー通話を可能にするために，携帯電話操作を音声で行う応用例が実用化されている．運転中の携帯電話の操作は安全のためにアイズフリー，ハンズフリーでの操作が求められており，国内では 2004 年の道路交通法の改正により手で携帯電話を保持したままの通話が禁止されている．そこで，ハンズフリー通話キットや Bluetooth 無線機能付き携帯電話が登場し携帯電話を手で保持することなく車室内で安全に通話ができるようになった．しかしながら，この場合でも運転中に電話をかけたくなった場合は，手操作で発信相手を検索しなければならず携帯電話の使用が制限される．これに対し，音声認識を使用すれば発信相手を発話するだけで容易に検索でき，携帯電話の通話開始から終了までを一貫してハンズフリーで操作可能となる．

そのほかの応用として，一部の車両では車両操作（例えば，エアコンやワイパーの操作）が音声で操作可能になっている．また，北米では XM ラジオなどの衛星ラジオの選局を音声認識で操作する機能が実用化されている．

**■参考文献**

1) 日本自動車部品工業会，“2007 年度出荷動向調査，”2007.
2) 国土交通省道路局，“カーナビ・VICS の出荷台数，”2009.
3) 石井和夫，小川浩明，角田弘史，加藤靖彦，表雅則，南野活樹，本田等，藤村聡，渡雅男，“カーナビゲーション用音声認識ユニット，”日本音響学会 H8 秋季講演論文集，vol.2-Q-30，1996.

■2群 － 7編 － 4章

# 4-3 音声と福祉

（執筆者：市川　熹）［2009年7月 受領］

2001年WHO総会で採択された人間の生活機能と障害の分類法「国際生活機能分類ICF」[1)]では，障害＝活動制限，社会的不利＝参加制約が，環境すなわち社会によって引き起こされるという非常に重要な観点に基づき規定された．現在社会環境の中で情報環境は非常に大きな部分を構成している．また，情報発信権や情報取得権など，情報環境への参加を可能とすることは，基本的人権であるという認識が国際的にも認められている．

社会的存在である人間にとって他者とのコミュニケーションは必要不可欠な機能である．その手段としては話し言葉や書き言葉，更にはメールなど多岐に広がっているが，最も基本的なものは，音声対話に代表される揮発性の対話言語により実現される対面の実時間対話である（揮発性対話型自然言語には他に手話などがある）．とかく言語情報に注目が集まり，以下の点は見落とされがちであるが，実態は周辺言語情報と非言語情報が存在することにより，円滑な対話が実現されている．ここで，周辺言語情報とは，発話文の構造情報と話者交替意図の表示などを実時間で知覚認知可能とする支援情報などを，非言語情報とは話者の感情や個人性などの多様な情報を意味する．

このような様々な情報の発信から受信までのいずれかの課程に障害があれば，円滑なコミュニケーションが困難になる．揮発性言語を用いるコミュニケーションの障害には様々な種類が存在することが分かる．更に，その障害は，先天的な原因や，病気や加齢による後天的な原因などが存在し，その内容は極めて多様である．また，視覚障害や加齢により視覚機能が低下している人にとっては，視覚機能の代替手段としての音声による対話は必要不可欠な手段となる．

心身障害の実態は極めて多様である．大きく身体障害，知的障害，精神障害に分類され，更に，複数の障害を持つ（重複障害）の人も多い．障害について考えるとき，上記の障害別分類のほか，その程度，障害の部位，先天性か後天性かなどを考慮する必要がある．また，その人のこれまでの生育環境，教育環境なども考慮に入れる必要がある．また，高齢者は65歳以上と定義されている．75歳未満を前期高齢者，75歳以上を後期高齢者と呼んでいる．日本は2007年1月ころ65歳以上の人口が全人口の21%を超え，世界で最初に超高齢社会*に突入した．

以下に音声による障害者支援の関係について概要を示す．音声と福祉の関係には，(1) 音声機能の障害，(2) 他の障害に関する支援技術への応用，(3) 音声の研究開発手法の応用，の三つの視点が存在する．手話なども揮発性言語として音声と共通の性質を有しており，横断的検討から有益な知見が得られている．

## 4-3-1 音声機能の障害

情報が紙で提供される時代には，視覚障害者は他の人に読んでもらう必要があった．電子

* 国連が高齢者の問題について報告した際に，欧米で社会的に高齢者の問題がクローズアップされた人口比がおおむね7%であったことを参考に，7%以上を高齢「化」が進行しているという意味で高齢化社会と命名された．その後，その倍の14%以上を高齢社会，3倍の21%以上を超高齢社会と習慣的にと呼ぶようになったとされている．

データで提供され，音声合成技術や点字ディスプレイが開発されると，それらを利用して自立して情報を獲得する道が開けた．しかし GUI の出現や URL などからの情報が増大している現在，それらには情報が 2 次元的に表示されており，また画像情報も多用されているため，再び大きな問題となっている．画面表示の固定された情報に対して，音声は 1 次元の揮発性の情報であり，単純に音声合成技術で変換するだけでは，情報取得には大きな負担がかかる．

脳性麻痺などで発話が困難な場合は VOCA と呼ばれる携帯用の機器を利用し，スイッチから文字を入力して合成音声に変換するものが色々開発されている．

声帯を癌などで切除され，発声が困難な人には人工喉頭と呼ばれる機器を喉の外壁にあて，声帯波形の代行を行い，発声支援を行う [2]．単に合成音声で出力するだけなく，生涯当事者の個人性や感情の音質も再合成する試みも行われている [3]．

### 4-3-2 他の障害に関する支援技術への応用

総務省は 2007 年までに字幕付与可能なすべてのテレビ番組に字幕を付加することを目標に設定した．これを実現すべく，放送音声を認識し，字幕に変換付与する技術の開発が進められてきた．この技術は学校における講義をビデオに収録し，字幕を付けて聴覚障害学生が利用できるシステムなどの応用が可能である．また，盲ろう者（視聴覚重複障害者）は文字も見ることができないため，音声認識結果を触覚情報である点字や指点字に変換する支援技術も開発が進められている．

発声が不安定のため，聞き取りが困難な音声を認識し，聞きやすい合成音声に変換する技術の開発が期待される [4]．しかし，個人差も大きく，当事者は緊張すると発声がその都度様々に変動するため現状技術では認識性能は十分ではない．文章レベルの認識も課題である．

自治体など公共の組織の HP はアクセシブルであることが求められる．しかし，残念ながら HP を作製する大部分の人はアクセシブルな HP の備えるべき条件を把握していない [5]．またアクセシビリティ Web に関する規定の各項目を遵守していても，我々の調査によれば必ずしもユーザビリティは良くはなっていない [6]．

漢字の読みの問題も存在する．日本語は漢字の導入により音韻の種類が少なくなり，漢語の利用もあいまって同音異義語が多くなっている．このため，視覚障害者にとっては，仮名漢字変換する場合に，適切な漢字を選択することに大きな課題を抱えている [7]．

視覚の利用が可能な先天性の聴覚障害者も実はその漢字の読み方の音声を聞いたことがないため，妥当な読み方が分からず，仮名入力ができず，仮名漢字変換が困難な例も存在する．

### 4-3-3 音声の研究開発手法の応用

障害に対し支援を行う視点として，アクセシビリティという言葉がよく用いられる．しかし，障害のため不慣れな代替手段によるコミュニケーションによる負担が存在することになる．したがって，単にアクセシビリティの視点だけでは不十分で，健常者の場合以上にユーザビリティの視点が重要になる．対話言語についていうならば，周辺言語情報や非言語情報が利用可能な自然なかたちで具備されていなければ，円滑な対話の実現は困難になり，揮発性の言語を主な代替手段とする人にとっては負担が非常に大きくなることを意味する．

一般のヒューマンインタフェースの開発では，多くの場合，使用効率のような客観的尺度や，ユーザにとってどのように見えるかといった主観尺度を用いて評価され，開発されてい

る．しかし，高齢者や障害者は，その障害の内容や程度が多様であり，各ユーザの経緯や経験などの相違からも，個人差が極めて大きい．したがって，その機器を実際の場面で使用するときに，どのような負担を負っているかも，一人ひとりにより大きく異なる．

認知心理学では「メンタルワークロード」（心的負担）という概念があり，機器使用時のユーザ個人ごとの心的負担を評価する主観的評価手法に活用可能である．支援技術の体系的開発を行うための基本的視点を与える [8]．

「心的負担」を軽減し，ユーザビリティを改善するためには，対話言語が実時間言語情報として認知的に最適化されているか，というようなことが今後重要な視点となろう．

### 4-3-4 今後期待される音声知見とその応用

音声対話では何故連続している音韻の列の中から言葉と言葉の境目が直ちに判るのか，何故脳の中に記憶されているであろう何万語，何十万語もある単語から特定の言葉が直ちに取り出せるのか，何故取り出された単語と単語の関係（文の構造）が直ちに判るのか，何故円滑な話者交替が実現されるのかなど，いずれも非常に不思議である．余裕を持って予測できるような仕掛けと，効率的な心内辞書のアクセス法が存在することが予想される．

これらの性質を明らかにし，活用できれば，障害者や高齢者だけでなく，誰に対しても負担の極めて小さいインタフェースの実現が期待される [9]．言い換えれば，発声された音声から最終的に得られる情報（伝達内容情報）と，このような情報の知覚認知過程を支援する情報（伝達支援情報）が同時に存在し，伝達される情報の円滑な獲得を可能にしていると考えられる．

また，話し手の発声と平行してその音声は聞き手により聴取され，理解が進行し，反応が現れ，またその反応に応じて発声が影響されるという，ダイナミックな制御を行う情報が存在すると考えられる．話者交替の制御などの対話制御なども含む情報である．まとめて示すと，

A．伝達内容情報　最終的に伝えられた内容

B．伝達支援情報　実時間での理解を支援する情報

- 伝達内容構造情報
- 伝達内容構造予告情報
- 話者交替予告（対話進行支援）情報

我々は，この伝達支援情報 B には様々なレベルの予告情報が含まれ，それを利用し予測することによって，聞き手の知覚や理解の負担を軽減する非常に大きな役割を果たしていると考えている．ユーザビリティを上げるためには，この情報 B の解明と活用が極めて重要となろう．また，心的辞書への高速なアクセス構造やコンテンツ構造の配慮も重要である．

文の構造や発話終了か継続かの情報（話者交替の可能な部分）がプロソディの中に先行して予告的情報が存在していることなどの事実が分かってきている [10-12]．

アクセントの役割は，主にセグメンテーション機能である．自然な言葉と同じ調子で発声された無意味な言葉を聞かせて，何処で切れると判断されるかという認知実験を行った結果は，80%以上の確率でアクセント節の切れ目と一致する位置で切れると判断された [14]．

対面対話では，単に言葉というかたちでのコミュニケーションだけでなく，感情や，誰が話しているかといった個人性などの情報も極めて重要な役割を果たしている．これらの情報

がセットになって，豊かなコミュニケーションが成り立っている．個人性（身体的・生理的特徴や言語的習慣）や心理的特長（感情などの現れ方など）などはプロソディ[15]やスペクトル特性などに現れる．

なおニュース文のアナウンスにおいても，談話構造には間や発話速度，基本周波数の構造が，ニュースのタイトル部分，概要，本文の区別の制御に利用されている[16]．

複数対話では，参加者に情報の授受の速度に差がある場合，遅い者が話題の展開に取り残されることが最大の問題である．コミュニケーション障害者がこれに相当する．意思決定の場に実時間で参加できることは人権上重要である．話題の展開に追従できる（一種の実時間性）ように，対話の流れを制御する工夫が必要になる．外出が難しく，触覚情報でのコミュニケーションを余儀なくされている盲ろう者が実時間合意形成に参加可能なインターネット利用会議システムが試作されている[17]．

音声と同じく揮発性である手話や指点字†も，実時間での対話が可能である事実から，これら手話や指点字にも音声のプロソディに相当する情報が存在するものと考えられる．

実際手話や指点字でも，音声と同様に，プロソディ機能に対応すると思われる情報を外すと，文の理解は約 80%から 50%程度に低下することが観測される[18]．これらの比較検討から，逆に音声に関する視点や理解が進む側面が存在する．

手話は音声と同様に対話型自然言語であり，実時間でのコミュニケーションが可能な性質を備えている．日本手話では，基本周波数は存在しないが，顔の表情や身体の動きなどの非手指動作（NMS）と，時間構造，動作の大きさの時間パターンなどがその情報を担っていると考えられる．これらを手話のプロソディと呼ぶことにする[19]．日本語対応手話‡も日本語音声と類似した時間構造のプロソディが存在する．

聞きやすい音声合成の実現に的確なプロソディの付与が重要なように，CG などによる手話動画合成には，手話のプロソディの的確な付与が欠かせない．

失語症には様々なタイプがあり，その解析からヒトの音声処理のメカニズムを知るうえでのヒントも多数存在する．手話が自然言語であることから当然予測されることではあるが，手話者の失語症にも聴者の場合と同じ現象が観測されている．特に手話においては，同じく揮発性の対話メディアとしての音声と共通の課題が存在しており，先行する音声の研究開発手法は参考になる項目が多い．

音声品質の評価方法は，手話の画像伝送の品質評価に応用されている．しかし，手話では音韻レベルの単位の解明がいまだ十分行われていないことや，空間に並列して情報が配置されている（両手や顔，身体などの動作が並行し組み合わされている）ために，現状では明瞭度の評価は難しく，了解度試験が中心である．

このような手法を参考に，手話画像伝送における情報圧縮技術の開発が進められている[22]．音声通話品質の課題に伝送系の遅延の影響の問題があるが，手話伝送においても，画像処理の負荷が重いこともあり，遅延の影響評価が必要になる．この課題に対する評価手法として

† 指点字は，盲ろう者のためのコミュニケーション手段の一つである．点字のできる視覚障害者から盲ろう者になった方が主に使用している[19]．障害者の 6 本の指を点字タイプライタに見たて，その指に点字を打って情報を伝えるもので，実時間性に優れた方法である[21]．熟達した盲ろう者は 1 分間に 300 から 350 字を読み取ることができるといわれている．点字のコード系を利用している．

‡ 日本の手話には，通説として，日本手話，日本語対応手話，中間型手話があるとされているが，学術的にはその実態はいまだ十分解明されていない．

音声通信における遅延評価手法が参考とされている [23].

音声の規則合成技術の開発では，音韻性の実現だけではなく，音韻の時間配置やアクセント，イントネーションなどプロソディの規則の開発が重要である．手話の規則合成（アニメや CG）においても，ろう者にとって読み取りやすい手話の合成には時間構造は重要である．音声の時間構造の認知的基準点を見いだすためのメトロノームとの同期発声分析手法を参考に，周期的振動と同期したろう者の手話を分析し，手話文の時間構造規則の手がかりを得ようという試みがなされている [24].

合成音声の品質の良さとして，どの程度余裕を持って楽に聞き取れるかという評価法として二重課題法がある．二重課題としては音声の聴覚チャンネルとバッティングしないように視覚情報を用いることが多い．しかしろう者の場合は視覚で手話を見ており，聴覚は使えない．最近では，二重課題法ではなく，NASA-TLX のような心的負担の評価法 [25]の利用の検討が進められている．

手話認識では音声認識に開発された HMM 法利用の検討が進められている．しかし手話ではいまだ音韻に相当する単位に関する知見が十分ではないため，模索の段階である．なお評価には，音声対話システムの評価法を参考に，単語認識率や理解率，完了率などの利用が検討されている．

対話の持つ特徴を活かすことは，障害者支援技術のユーザビリティを向上するうえで大きなヒントを与えるものと期待される．情報の受け手の状態が送り手の情報表現に影響を大きく及ぼすことが見られる．例えば子供や外国人に話す場合は，相手の理解の程度を勘案して自然に話し方が変わることは極普通に見られる現象である．表や数式などは 2 次元情報となっており，それらの情報獲得に対話形式の利用が考えられる [26].

肢体不自由者は文字盤操作が困難であり操作回数の少ない入力方法などが検討されている．例えば音韻の出現頻度情報やバイフォンモデルなどの応用である [27].

障害者や高齢者の一人ひとりの実態は極めて多様である．その一人ひとりがユーザブルに機器を使用できるようにするためには，従来の理工系の視点である主観性を排し統計的に行う量的評価手法は原理的に成り立たない．人間科学の方法論としての質的アプローチが不可欠である．人間を尊重する時代の現在，理工系の学会も研究開発評価の視点を質的評価法 [28]にまで広げてゆくことが不可欠である．

**■参考文献**


1) 厚生労働省，“国際生活機能分類－国際障害分類改訂版－（日本語版），”2003. http://mhlw.go.jp/houdou/2002/08/h0805-1.html （2003）
2) 伊福部達，“福祉工学の挑戦－身体機能を支援する科学とビジネス，”，中公新書 1776，中央公論新社，2004
3) A.Iida, et.al., “Communication aid for non-vocal people using corpus-based concatenative speech synthesis,” Proc. Eurospeech 2001, pp. 2401-2409, 2001.
4) 松政宏典，他，“情報家電操作における脳性麻痺構音障害者の音声認識評価，”電子情報通信学会福祉情報工学研究会資料，WIT2007-7，pp.33-38，May 2007.
5) 紺野加奈江，“失語症言語治療の基礎，”診断と治療社，Sep. 2001.
5) “「公共分野におけるアクセシビリティの確保に関する研究会」報告書の公表，” http://www.soumu.go.jp/s-news/2005/051215_1.html
6) 飯塚潤一，他，“視覚障害者のウェブサイトの検索効率と心的負担に関する考察，”電子情報通信学会

福祉情報工学研究会資料，pp.55-60，March. 2007.
7) 西田昌史，他，“意味情報を利用した視覚障害者が連想しやすい仮名漢字変換手法，”日本音響学会春季講演論文集，vol.2-8-8, pp.343-344, March. 2007.
8) 情報福祉の基礎研究会 編著，“情報福祉の基礎知識―障害者・高齢者の使いやすいインタフェース，”ジアース教育新社，April 2008
9) 市川熹，手嶋教之，“福祉と情報技術，”オーム社，Sep. 2006.
10) 大須賀智子，他，“音声対話での話者交替／継続の予測における韻律情報の有効性，”人工知能学会論文誌，vol.21，no.1，pp.1-8, Jan. 2006.
11) 木村太郎，他，“遺伝的アルゴリズムによる F0 モデルパラメータ推定法と話者交替分析への適用，”電子情報通信学会音声研究会資料，SP2006-82，pp.37-42，Dec. 2006.
12) 小松昭男，他，“韻律情報を利用した構文推定およびワードスポットによる会話音声理解方式，”電子情報通信学会論文誌，vol.J71-D, no.7, pp.1218-1228, 1988.
13) Ohsuga.et.al., “Estimating Syntactic Structure from Prosody in Japanese Speech,” IEICE Trans. Inf.&Syst., vol.E86-D, no.3, pp.558-564, March. 2003.
14) Hatano, T., “Prosody Based Speech Segmentation,” In Proc. The 5th International Conference of the Cognitive Science (ICCS2006), pp.103-104, 2006.
15) 市川熹，他，“合成音声の自然性に関する実験的考察，”日本音響学会秋季講演論文集，pp.95-96，1967.
16) 市川熹，“ニュース文のポーズとピッチ，”日本音響学会講演論文集，vol.2-8-7, March 1994.
17) 宮城愛美，他，“発言権を考慮した指点字と文字による会議システムの構築，”電子情報通信学会論文誌 D，vol.J90-D, no.3, pp.732-741, March 2007.
18) 北原義典，他，“音声言語受容における韻律効果の検討，”電子情報通信学会創立 70 周年記念総合大会 1339，1987.
19) 市川熹，“人と人をつなぐ声・手話・指点字，”岩波書店，Oct. 2001.
20) 市川優子，他，“認知科学的手法による手話読取特性の検討，”日本手話学会第 22 回大会予稿集，vol.5, no.3, pp.71-74, June 1996.
21) 宮城愛美，他，“指点字のプロソディの分析，”ヒューマンインタフェース学会論文誌，vol.1，no.3，pp.35-40, Aug. 1999.
22) 中園薫，“手話動画像通信に関する研究，”千葉大学学位論文，March 2006.
23) 寺内美奈，他，“遅延手話対話における話者交替時の信号表出に関する解析的検討，”電子情報通信学会福祉情報工学研究会資料，WIT-2007-11，May 2007.
24) 平山望武，他，“日本手話における時間構造の分析，”ヒューマンインタフェース学会論文誌，vol.3，no.3, pp.9-14, Aug. 2001.
25) 芳賀繁，“メンタルワークロードの理論と測定，”日本出版サービス，July 2001.
26) 藤原敦史，他，“表および数式の音声化の検討，”ヒューマン・インタフェース・シンポジューム論文集，pp.643-648, Oct. 1997.
27) 森大毅，“連続音声認識の手法を応用した走査型文字入力方式，”日本音響学会春季講演論文集 1-Q-31, March 2007.
28) ウヴェ・フリック，“質的研究入門＜人間の科学＞のための方法論，”春秋社，2002.

■2群 － 7編 － 4章

# 4-5 音声コーパス

（執筆者：板橋秀一，大須賀智子，山川仁子）［2009年 5月 受領］

音声認識・音声合成などの音声情報処理の研究を進めるうえで音声データが必要なことは言うまでもない．そのデータは多種多様であることが求められる．最近では統計的手法の発達により，大量のデータがシステムの学習のために必要とされるようになった．一方，音声処理システムの研究・開発を進めるためには，各種の手法を適切に比較・評価することが必要であるが，これを行う方法としては現在のところ，共通の音声データを用いて処理を行い，その結果を比較するという方法以外は知られていない．

このようなことから共通利用可能な各種・大量の音声データを作成し，その利用体制を整備することは，研究・開発過程での利用及び各種の装置やシステムの性能評価の両面から求められている．このような目的に利用される音声データを一般に音声データベースあるいは音声コーパスと呼んでいる．音声情報処理の分野では「音声データベース」というとき，データベースシステムよりも「大量の音声データの集積」そのものを指すことが多い．そのため最近では，後者を意味する「音声コーパス」を使うようになった．音声コーパスは本編でとりあげている分野はもとより言語学的研究にも広く利用できる．またテキストコーパスや音声・言語処理ツールなどを含めて「言語資源」ということもある．

音声コーパスの必要性やその意義については近年広く認められるようになってきたが，更に音声及び関連する分野の研究の発展のためには，音声データを作成・収集・蓄積・配布するための共通の枠組を用意することが必要であると考えられるようになり，各国で体制の整備が進められている．

## 4-5-1 ラベル

音声コーパスには音声（波形）信号のほかに，音声を分析して得られる基本周波数（声の高さに相当する）などの分析パラメータや発声時の声道（声帯から口唇までの音の通路）のX線写真や発声器官の筋電図などを含むものもある．更に音声コーパスはラベルつきとラベルなしに分類される．ラベルとは音声データのどの部分がどのような音声に相当するかについてのマークである．また，音声の抑揚（イントネーション）やアクセント，強弱，休止などを示す韻律ラベルもある．ラベルづけ処理の完全自動化はできず，音声波形やスペクトルを表示して人間が音声を聞きながら確認を行うため非常に手数がかかるので，ラベルつきのコーパスは少ない． 図4･3に「音声コーパス」と発声した女性の音声の分析例とラベルを示す．

## 4-5-2 海外における音声コーパスの状況

アメリカでは前述のような背景のもと1992年にLDC （言語データコンソーシアム）が設立された．これは音声・言語コーパスに関する国際的なコンソーシアムで，米国の大学・企業を中心に100機関余が会員となっている．ヨーロッパではELRA （ヨーロッパ言語資源協会）が1995年に設立され，音声・言語コーパスの構築・供給体制を確立している．LDC/ELRAからは各種の音声・言語コーパスが入手可能である．代表的なものとしてはLDCコーパスの

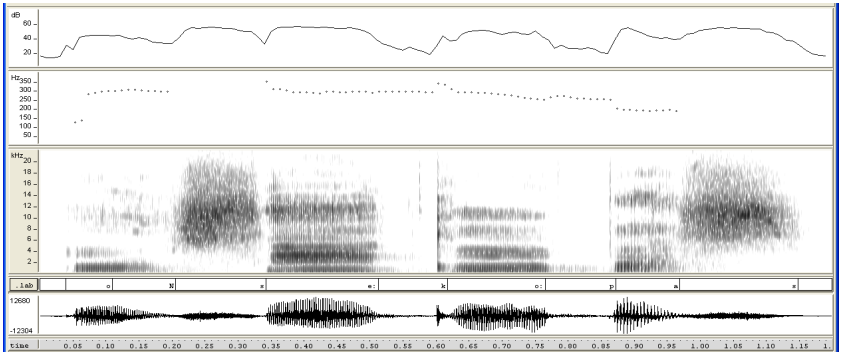

**図 4・3　女性が発声した「音声コーパス」の分析例**

**下から順に音声波形，ラベル，スペクトログラム（周波数分析した結果を濃淡模様で表したもの），基本周波数（F0），及び音声パワーを示す．**

中でこれまでに多く配付された次の 4 種が挙げられる．TIMIT（音響音声学連続音声コーパス），TIDIGITS（連続数字認識用音声コーパス），NTIMIT（TIMIT コーパスの電話帯域音声版），YOHO（話者認識用音声コーパス）．

音声データベースに関する国際協調を推進するために COCOSDA（音声データベースと評価技術標準化国際協調委員会）が 1991 年に設立された．音声関係の国際会議に合わせて毎年ワークショップを開催して，各国のデータベースの現状報告や，多言語データ収集の協調などについて話し合っている．

アジアの言語についても同様の動きがあり，上記 COCOSDA の東アジア部会として Oriental COCOSDA が活動を行っている．1998 年に日本で第 1 回の会議を開催して以来毎年会議を開催し 2008 年には 10 周年記念大会が日本で開催された．一方，韓国では SITEC（音声情報技術・産業振興センター）が，中国では Chinese LDC と CCC（中国コーパスコンソーシアム）が 21 世紀の初頭に発足して音声・言語コーパスの供給・利用を推進している．

### 4-5-3 日本国内における音声コーパスの状況

日本においても費用の負担を分担できるような組織を作ることで，大規模な音声データの開発及び普及が促進されるようにすることが必要であることから，1999 年に言語資源協会（GSK）が設立された．GSK は 2003 年に特定非営利活動法人（NPO）として東京都の認可を受け，主にテキストコーパスの配布を中心に活動を進めている．また 2006 年には国立情報学研究所（NII）に音声資源コンソーシアム（NII-SRC）が設置され，音声コーパスの取り扱いを始めた [1]．情報通信研究機構（NICT）では特にテキストコーパスの構築・配布に力を入れているが [2]，更に 2009 年には高度言語情報融合フォーラム（ALAGIN）が発足して，NICT が開発した言語資源の公開を始めた [2]．これまで国内の主なプロジェクトで作成されたコーパスを以下に紹介する．

日本音響学会では，音素バランス文，読み上げ文，模擬対話音声を含む「研究用連続音声データベース」を刊行した．ATR（国際電気通信基礎技術研究所）の音声関連研究所では単

語音声，文音声，多数話者音声，英語音声のコーパスなどを多数構築した．文部省の重点領域研究「音声言語」「日本語音声」「音声対話」では連続音声，方言音声，対話音声のコーパスを作成した．通産省のリアルワールド・コンピューティングプロジェクト（RWCP）では，海外旅行や自動車購入の話題に関する対話音声コーパスを構築した．日本音響学会と情報処理学会の音声関係者の協力により，「新聞記事読み上げ音声コーパス」が作成・公開された．開放的融合研究推進制度による「話し言葉工学プロジェクト」では延べ 700 時間，700 万形態素の「日本語話し言葉コーパス（CSJ）」が作成された．名古屋大学の統合音響情報研究拠点（CIAIR）では自動車内データを中心に各種のコーパスを作成した．文科省の特定領域研究「メディア教育利用」では日本人学生による読み上げ英語音声コーパスや留学生による読み上げ日本語音声コーパスが作成され，特定領域研究「韻律と音声処理」では韻律コーパスを作成した．情報処理学会の研究会では，実環境下における音声認識性能を評価するための雑音下音声データや評価スクリプトを CENSREC シリーズとして作成している．これらのコーパスは NII-SRC や各作成機関から入手可能である．音声コーパスの詳細については文献 1) を参照されたい．

音声・言語研究者あるいは研究機関が作成・保有している音声・言語コーパスは多数あるが，保有者自身がそのデータを配布し利用に供することは容易ではない．GSK や NII-SRC などの介在により，それが容易に実現できるようになり，データの有効利用を促進することが期待される．上述のように日本では音声・言語コーパスを扱う機関が多数並立しているが，これらの間の有機的な連携が望まれる．また利用者のコーパス選択を助ける工夫が必要である [3]．

**■参考文献**

1) 板橋秀一, 大須賀智子, “NII 音声資源コンソーシアムについて,” 信学技報, vol.SP2007-7, pp.35-40, May 2007.

2) 中村哲，隅田英一郎，鳥澤健太郎，“NICT における音声・言語研究拠点 MASTAR プロジェクトについて,” 信学技報，vol.NLC2008-74, pp.19-24, Jan. 2009.

3) 山川仁子，松井知子，板橋秀一，“複数音声コーパスの類似性の視覚化,” 信学技報，vol.TL2008-7, pp.35-40, May 2008.