

音声情報の能動的再構築処理に関する研究：騒音環境における音声認識を支援するシステム

勝瀬, 郁代

<https://doi.org/10.15017/1398257>

出版情報：九州芸術工科大学, 2001, 博士（工学）, 課程博士
バージョン：
権利関係：

Chapter1

序論

本章では、本研究の目的、背景を示し、その必要性、採用したアプローチの意義を明らかにする。

1.1 研究の目的

機械による音声言語理解の過程で、環境騒音などの要因により情報の損傷が存在した場合、その欠落した情報の修復を行う。その上で、物理的な復元ではなく、人間の音声言語理解過程の特徴である能動的な再構築過程により「修復」を行うための処理を工学的に実現することを目的とする。

1.2 本研究の背景と必要性

音声言語はそもそも人間同士のコミュニケーションの道具として発展してきたものであるが、近年の情報技術の発展により、音声言語は人間同士のみならず、人間と機械間のコミュニケーション手段のひとつとして注目されている。特に、音声を機器操作のための入力手段として用いることができる音声インターフェースの実現は、情報端末を扱う上での人間の負担軽減を期待できるものとして望まれている[1]。そのような要求を背景として、人間が話した音声信号を言語テキストへ“復号化”することを目的とする自動音声認識技術が開発された。近年、自動音声認識システムは、発話の様式や話題が限定されるにしても、大語彙連続発話音声の認識率が95 %を超えるに至っており[2]、パーソナルコンピュータで実時間動作が可能となり[3]、望まれているほとんどの目的への応用には適用可能であるように見える。しかし現実には自動音声認識技術を応用システムに適用しても、その性能を十分に発揮できないことが多い。その原因は、次の3つの課題が残されたままであるためと考えられる。

(1) 現行の自動音声認識システムは、その性能を発揮できる発話様式が読み上げ様式¹に限定

¹書き言葉の様相を持つ音声言語。

されている。人間の発話負担が少ない会話音声²では、十分な認識性能を發揮できない。

(2) 現行の自動音声認識システムは、発話された言葉を忠実にテキスト化することを目指しているが、音声タイプライタとしての使用目的以外では、正しく聞き取れた言葉の数の割合よりも、話し手の意図を正しく理解できたかどうかの方が重要であることが多い。そのため、アプリケーションの立場から音声認識システムの性能を新たな尺度（confidence measure）で評価しようという動きがある（例えば[4]）。つまり、実際の応用では言葉の聞き取りだけでなくその理解までもが必要とされる。

(3) 自動音声認識技術の応用が望まれる環境は、駅、空港、路上、車内、レストラン、工場など、制御不能な環境騒音が存在する場所であることが多い。そのような環境騒音の存在する状況では、自動音声認識システムは十分に性能を發揮できないことが多い。

本研究では、上記3つの要因のうち、主に（3）に関する課題を克服するためのものである。ただし第5章では課題（2）の克服も含まれる。課題（1）や課題（2）は、自動音声認識技術だけでなく、自然言語処理や推論、メディア理解、センサーフュージョンといった、他の高度な情報処理技術との融合により成り立つシステムが抱える問題であるのに対し、課題（3）は、音声コマンド方式という単純な場合から、自然な会話を実現する音声対話システムに至るまで、音声入力機能を有するすべてのインターフェースが直面する課題である。それゆえ、課題（3）の克服は音声入力機能を有するすべての応用システムの実用化を強力にサポートするものであり、本研究の意義は大きい。以下の節では、工学的立場と聴覚科学の立場からの騒音下の音声認識の捉え方について述べ、本研究の目的を達成するためのアプローチについて言及する。

1.2.1 騒音下自動音声認識を実現するための工学的研究

工学的観点からは「音声認識」の定義は厳密に規定される。自動音声認識とは、「入力音声の表層的内容（文字列情報）を、計算機を用いて同定する処理」と定義される³。それゆえ、騒音下自動音声認識とは、「目的音声に騒音が混合した混合信号から、目的音声の表層的内容（文字列情報）を、計算機を用いて同定する処理」と定義できる。騒音環境での頑健な音声認識の実現を目指してこれまで数多くの手法が提案してきた。これまで提案してきた手法は次の4つのカテゴリーに分類できる。以下では、各カテゴリーの代表的な研究事例を挙げ、その特徴について述べる。

- 複数マイクロホンによる音源分離 (Microphone Array)

自動音声認識システムの前処理として位置し、混合信号から音声信号のみを抽出しようとす

²省略、繰り返し、言い直し、言い換え、挿入、倒置はもちろんのこと、構文的誤り、言い淀み、繰り返し、発声のなまけが含まれる。

³これに対して、入力音声の意味内容を同定する処理は音声理解と呼ばれて区別されている[5]。

るもの。原音声をできるだけ物理的に忠実に抽出することを目的とし、原則的に自動音声認識システムとは独立して構築される。適応形雑音抑圧マイクロホンアレイ[6]では、小規模、小素子数のマイクロホンアレイを用いて、使用する音場における雑音到来方向に対して感度の低い指向性パターンを自動的に形成する。独立成分分析[7][8]のアプローチは、互いに独立な信号が混合した観測信号から、信号の独立性以外の先見的知識なしに、もとの信号を抽出する手法である。騒音源の数や伝播空間特性が既知または推定可能でなくてはならぬいため、適用可能な騒音環境の範囲が限られる。

- 耐騒音性音響特徴抽出 (Noise Resistant Features Extraction)

このアプローチは、音声信号におけるノイズの効果に焦点をあて、耐雑音性の音声特徴やパラメータの抽出を図るものである。代表的な手法は、CMN (Cepstrum Mean Normalization) である[9]。雑音の特性に関する仮定があまり一般的でないため、適用可能な騒音環境の範囲が限られる。

- 音声強調 (Speech Enhancement)

このアプローチは、騒音と音声の混合信号から音声信号のみの抽出を図るものである。代表的な手法として、スペクトルサブトラクション[10]、櫛形フィルタ[11]などがある。一般に抽出された信号にはスペクトル歪みが見られ、認識システムの認識性能の劣化につながる。

- 音響モデルのノイズ補償 (Speech Model Compensation)

ノイズモデルを利用して、音声認識システムで使用される音響モデルのノイズ補償を行うものである。HMM 分解[12]・合成[13]手法、その拡張形であるPMC[14]により実現される。一般的にノイズモデルは騒音環境ごとに用意する必要があり、騒音の変動にもある程度の定常性が必要となる。

1.2.2 騒音下音声知覚における聴覚の能動的な働き

工学における音声認識の研究では、これまで音声の表層的内容を同定することに重点がおかれてきた。一方で、人間にとって「音声を聞くこと」は、音声波形などといった表層的な信号の受容だけを意味するものではない。むしろ話し手が「音声」という媒体を通して聞き手へ伝えたいのは、「意図」であり、同様に聞き手が知りたいのも話し手の「意図」である。その意味では、たとえ音声信号が耳に到達した時点でなんらかの歪みを受けていたり雑音が重畠していたりしても、それから話し手の発声した原音声を物理的に復元する必要はないということになる。さらに、聞き手の脳内で音声情報から言語情報へ変換される場合も、話し手の発した一語一句を完全に復号する必要はなく、その後の意味解釈のために十分な情報が得られればよい。

音声知覚理解における人間の振る舞いを観察すると、音声波によって運ばれてきた情報は、聞

き手側が持つさまざまな処理機構が働くことにより逐次処理されるというよりもむしろ能動的に再構築されるという方が適当であることがわかる。

本研究では、聞き手が音声という媒体を通じて話し手から情報を伝えられたとき、様々なノイズによって部分的に欠けた情報を能動的に再構築しながら話し手の意図を理解する過程に焦点を当てる。そして聞き手によるそのような能動的な働きを次の3段階に分け、以下の各節にて概説する。

第一段階 : 原初的聴覚情景分析の段階～生態学的制約による聴覚情景分析

第二段階 : スキーマに基づく聴覚情景分析の段階～音声言語の知識に基づくトップダウン処理の導入

第三段階 : 音声言語の、コミュニケーションの道具としての役割を利用した処理が行われる段階～音声対話による情報の再取得

第一段階：原初的聴覚情景分析

音声言語という高次機能に注目していると、聴覚の本来の役割について見失いかねない。聴覚は、そもそも生物が環境を把握するための一環したシステムの一部として発達したものである[15][16][17]。聴覚の主な役割は、耳に到達する音を分析し、環境のどこでどのような音響事象が起きているのかを推定することである[18][19][20]。そのためにはまず、聴覚は時間一周波数上で分解された音の基本的要素を音源ごとに統合しなければならない。このような統合は群化(grouping)と分凝(segregation)の問題として捉えることができる。Bregmanは、ゲシュタルト心理学の群化の法則にならって、聴覚的な群化と分凝を支配する現象的法則を整理した[15][21]。彼は、聴覚体制化(auditory organization)のための手がかりを、音源一般に関する物理的(または生態学的)制約のみを反映するものと、特定の音源に特化した、知識に基づく制約の導入が含まれるものとに分類した。前者を原初的聴覚情景分析(primitive auditory scene analysis)、後者をスキーマに基づく聴覚情景分析(schema-based auditory scene analysis)と位置付けている。スキーマに基づく聴覚情景分析については次節で述べる。

原初的聴覚情景分析は、一般的な音響的規則性が音源分離の手がかりとして使われ、スキーマに基づく聴覚情景分析に先だって行われる、より一般的な分析である。

原初的聴覚情景分析における音脈形成(stream formation)の手がかりは、さらに系列的統合(sequential integration)と周波数統合(simultaneous grouping)に分類される。系列的統合は、局所的情報の時間的な関係に関する制約(音の高さ、周波数成分の分布、音の強さ、空間的位置、時間的变化パターンなど)によって、それらを音脈へ割り当てる。周波数統合は、局所的情報の周波数的な関係に関する制約(主に調波的関係)によって、それらを音脈へ割り当てる。

本論文第2章では、原初的聴覚情景分析の系列的統合における制約として、動的な追跡過程に

よる連続性の手がかりを新たに提案する。さらに第3章では、第2章で提案された手がかりを工学的に実現するための計算機モデルを提案し、これまで明らかとなっている原初的聴覚情景分析の手がかりと組み合わせて、能動的聴覚体制化を行う原初的聴覚情景分析の計算機モデルを提案し、音声信号に対して適用した例を示す。

第二段階：音声言語の知識に基づくトップダウン処理の役割

人間の音声言語理解の過程を計算機による情報処理のアナロジーで捉えると、音声信号の感覚的変換物をその信号に意味を与える概念過程に連結するための一連の認知過程であるといえる。そしてこのような一連の認知過程はボトムアップ処理（データ駆動型処理）とトップダウン処理（概念駆動型処理）に大別される。ボトムアップ処理は、外界の情報を直接知覚対象に変換する処理であり[22]、トップダウン処理は、可能な解釈についての知識、すなわち、何かについての概念化がその事物の知覚を助ける時、そこで起こっている処理である[23]。音声言語理解の過程では、概念化において音韻論的、統語論的、及び意味論的知識が積極的に利用されていると考えられている[24]。

人間が持つ音声言語に関する様々な知識は、音声言語理解の段階だけでなくさまざまなレベルで人間の音声知覚に影響を与えている。音声言語に関するスキーマに基づく聴覚情景分析は、音脈の形成に対して強固に働く。時には、原初的聴覚情景分析の結果と矛盾するような答えを選択することも厭わない。そしてこのようなトップダウンの働きは、騒音下での音声知覚に対して重要な役割を果たしている。

人間の、騒音下における音声知覚の頑健さを示す典型的な証拠として、音韻修復現象(phonemic restoration)[25][26]がある。音声の一部が雑音に置換されたものを聴いた時、聞き手はその置換されたはずの音声を完全な形で聴くことができ、置換雑音は付加雑音として知覚されるというものである。音韻修復現象には、断続的に雑音で置換された信号に連続感をもたらす聴覚の原初的働き[27]や置換雑音の前後の調音結合情報[28]、文脈からの予測可能性[29]などが関係しているようである。音韻修復は聴覚処理とは関係のないスキーマによる錯覚現象である[30]という見解もあるが、一般的には、トップダウン情報に基づく聞き手の「期待」と音響信号の特徴の交互効果によるものである[31]と考えられている。音韻修復現象のように音声が完全に雑音に置換される場合に限らず、騒音下の音声知覚において、知識に基づく「予測可能性」が音声知覚に与える効果についても報告されている[32][33][34]。

このように、人間の音声知覚では、トップダウン情報とボトムアップ情報は共に用いられるものであるが、トップダウン情報はボトムアップ情報の不完全な部分を補う働きをするというよりもむしろ、ある音声が存在しているという「期待」を作り出し、この期待に合うように処理にバイアスをかけるというきわめて能動的な働きをするものである[35]。このような能動的働きを工

学的に実現するためには、人間が音声に関する「どんな」知識を「どのように」用いているのかについて考慮する必要がある。

本論文第4章では、人間の音声知覚モデルを考慮した上で、トップダウンの知識とボトムアップ処理の融合を図り、かつ工学的に意義のあるシステムとして騒音下での音声認識を実現する手法について議論する。

第三段階：コミュニケーションの道具としての音声言語

元来、音声は人間同士のコミュニケーションのための道具の一つとして発展してきた。コミュニケーションの多くは、対面または電話などの通信媒体を通じての音声対話の形態で実現される。音声対話の様相[36]は次の3つに分けられる。

- (1) 音声・言語レベル：言葉のやりとり自体に意味があり、対話にははっきりとしたゴールや話題はない。
- (2) 情報レベル：相互に説明、理解、質疑応答などにより、知識や意図の交換を行うもの。
- (3) 強調レベル：共同作業、共同行動などのために話し言葉が利用されるもの。情報レベルに加えて、作業結果が逐次対話にフィードバックされる。

間身体的な場の構築[37]を目的とした(1)の様相を除けば、通常、対話者は相互に説明・理解・質疑応答を繰り返す。このような情報の授受形態は、聞き手が取得できなかった情報を、音声対話を通じて話し手から再取得することを許容する。音声信号は時間情報であり、一瞬のうちに消え去るものであるため、音声信号という表層的な情報は、一度失われれば二度と取得できない可能性がある。しかし、音声対話を通じて我々はその音声波形が運んでいた情報（話し手の意図）を再取得することが可能なのである。

第5章では、第1段階、第2段階の再構築過程では修復が困難であった情報を、音声対話を通じて再取得することにより情報の再構築を行う、音声対話システムを構築する。

3つの段階の関係

聞き手による能動的働きは、各段階においてそれぞれ異なるレベルの再構築を行うことができる。そして現実問題として、音声言語という複雑な情報を扱うにはこのような様々なレベルでの情報の再構築が必要である。

第1段階では音響的規則性に基づく再構築を行うものであった。いわば第1段階はデータ駆動型処理（ボトムアップ処理）に重点を置いた段階であるといえる。第2段階では、第1段階のデータ駆動型処理に加えて概念駆動型処理（トップダウン処理）を導入することにより、より頑健な再構築が可能となる。しかしながら入力情報なしでは音声言語は知覚すらされないことからわかるように、第1段階は第2段階の実現のための基本的構成要素であり、その存在意義は大きい。

ところで、第1段階、第2段階における「修復」とは、背景騒音の影響を受けた音声情報を受信した後、目的とする情報(第1段階では音声信号であり、第2段階では言語的シンボルであった)を得るためのものであった。音声波は時間情報であり瞬時に消え去るものである。それゆえ聴覚は、ある時刻で耳に届いた音響信号から十分な情報が得られなかつたからといって、空間情報を対象とする視覚がもう一度視線を向けることによって情報を確認するように、音響情報を再取得することはできないのである。そのため、第1段階、第2段階における修復だけでは、信号が受信時点で修復不可能な状態であった場合にはその情報は永遠に失われることになる。ところが、対象を音声言語情報に限定すると再取得が可能となる場合がある。すなわち、音声言語は人間のコミュニケーション手段として存在するがゆえ、音声対話を通じて情報の再取得の道が残されているのである。これが第3段階である。

では第3段階の修復があれば、第1段階、第2段階は必要ないのであろうか。答えは否である。通常我々が音声言語を用いてコミュニケーションを行う環境では、背景騒音がまったく存在しない状況の方が稀である。「聞き返し」による情報の再取得のみに頼るのは、際限なく「聞き返し」が発生することにつながり実用的ではない。あくまで、「聞き返し」は第1段階、第2段階において再構築ができなかつた場合の非常手段である。

以上のように、レベルの異なる再構築過程を設けることによって、より頑健で柔軟な音声言語情報の再構築が可能となるのである。

1.2.3 本研究のアプローチ

音声言語は人間が話し、聞くために発明され、発展してきたものである。そのため、高品質で頑健な音声認識技術を実現するには、人間の音声生成や音声知覚に関する知見を踏まえて、機能モデルを構成することが王道であると言える。しかし、現時点では音声生成や音声知覚に関する人間の持つ機能のほとんどはまだ解明されていないため、知見に基づく計算モデルのみでは工学的に十分に機能する技術とすることは難しい。それゆえ、現時点では、人間の機能の工学的実現を目指すというよりも、むしろ従来技術では困難な技術課題に対して工学的な観点から取り組む過程で、聴覚科学の知見を導入するというアプローチを取るべきである。本研究はこのような枠組みで行われている。本研究では、主として環境騒音を中心とした ill-formed な音声入力機構を有するシステムに対して人間の能動的な再構築機能の導入を図ることにより、実用的なシステムの実現を目指す。

1.3 論文の構成

第2章、第3章では、原初的聴覚情景分析に焦点をあてる。原初的聴覚情景分析における系列的統合過程として、新しい連続性の評価の導入の必要性を述べる。第2章では、心理物理学的実

験を通して、従来研究では否定的な見解が主流であった聴覚の「外挿過程」の存在について検証する[38]。第3章では、この心理学的知見の計算モデルを構築し、さらに、計算機による原初的聴覚情景分析の実現を通して、計算機モデルの妥当性を検証する[39]。第4章では、音声スキーマに基づくトップダウン処理を導入した、騒音下における音声知覚モデルを提案する。これを一般に普及している音声認識システムと融合させることにより、工学的な意義が認められるシステムとして実現する[40][41]。第5章では、音声言語の本質的な役割である音声コミュニケーションの形態である「音声対話」を通して、いったん受容した音声信号からでは取得できなかった情報の獲得を可能にするシステムを実現する[42]。第6章で本論文を総括する。