9. 発話音声の特徴

―音声文法からの観点―

ニック　キャンベル

要旨

本稿では、発話信号が複数の層にわたる意味情報をどのように伝えているのかを示したい。また、従来、発話処理とは「文字テキストを音声に変換したり、音声を文字テキストにしたりするといった『メディア変換』処理のことだ」という考えが広く浸透しているが、それは単純すぎ、あるいは発話における情報表現の問題を過小評価しすぎているということを示したい。対話音声を処理するには、新たな技術の開発が必要である。その技術とは、対話参加者どうしがやりとりした言葉だけでなく、情報のやりとり、つまり談話の文脈や話し手の意図、態度、発話アクトをも含んだやりとりの技術である。そこで本稿では言語情報やパラ言語情報を伝える音声信号のバリエーションを詳細に論じる。さらに、さまざまな層にわたる情報が、どのように相互に影響し合って話し手の意図を表し、メッセージを聞き手が解釈しやすいようにしているかを示す。

1. はじめに

話し言葉は、多層の意味情報をもち、それらは相互に影響し合いながら文脈の中で生きた発話をつくる。それぞれの発話は、言語情報（すなわちその内容）、パラ言語情報（すなわちその目的）、言語外情報（すなわちその話者と文脈について）から構成されている。現在の音声処理技術は、主として言語情報の抽出と生成に焦点が当てられており、発話者とその文脈に関する言語外情報や、意図された発話の解釈に関するパラ言語情報は、あまり重要視されていない。これはあたかもメッセージを担い伝達するのは言葉だけだと

考えられているようなものである。音声を使ったやりとりでは、発話者の意図を伝達するためには、常に、言葉だけで十分とはいえない。

人間が他者とやりとりする場合、重要になるのは、「何が」話されているかということだけではない。「どのように」話されているか、あるいは「誰が」話しているのかということも、否応なしにわれわれに影響する。将来の音声処理インターフェース、特に人間の対話の翻訳に使われるインターフェースも、それが言語間翻訳機器として用いられるにせよ、情報提供機器として用いられるにせよ、「どのように」「誰が」といった情報を利用可能とする必要があろう。音声Aによって発話された場合も、音声Bによって発話された場合であっても、伝えたいことはそれなりに伝えられるかもしれない。しかし、単語の連鎖がどう解釈され、どう翻訳されるかは、韻律Aか、あるいは韻律Bであるかによってかなり変わってくる。

以下の節では、発話における韻律の役割を詳細に位置づけ、韻律が表す情報の諸レベルを定義する。また、韻律の違いが信号化されるメカニズムも示す。以下では「韻律」の定義は伝統的な定義から拡張し、それがパラ言語情報を伝える以上、音色をも含むとする。

2. 発話のメカニズム

発話は、時間に依存するものである。発話全体を瞬時に知覚することは不可能であり、発話各部は発話されると順次消えていく。発話は時間進展に密接に関わる信号である。発話を聞いて処理する場合、我々人間は、短期記憶を働かせて、メッセージ内容のイメージを心内に組み立てる。つまり、その発話が、どういう単音をどう連鎖させたもので、どういう韻律特徴を持っているかを短期記憶し、知覚されたそれらのパタンを相互に影響させ合って、話し手の意図はこうであろうとイメージを作り上げる。発話音声がすぐさま消えてしまうのに対して、韻律は、韻律パタンにより構造を明確なものにし記憶に残りやすくすることを助け、メッセージの組み立てに役立っている。

図1に発話様式によって4つの異なる意味を伝える発話例を示す（坊農2001）。文字で表せば同じ「あ」であるが、それらは何か理解したときの「あ」であったり、忘れ物にでも気づいたときの「あ！」であったり、夏の

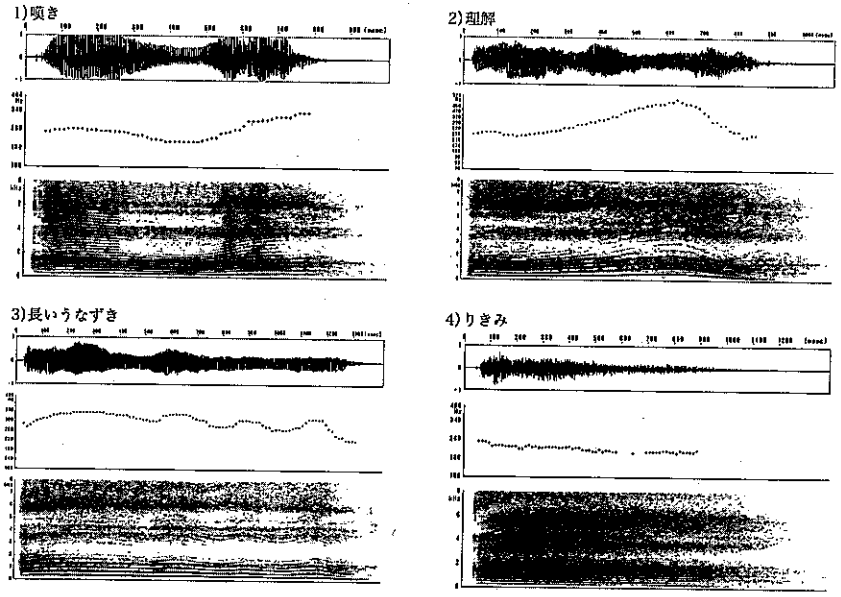

図1　発話 /a/ のバリエーション（坊農 2001 からの引用）

冷たいビールに「あー」であったりなど。ここに示した表現を一例として、はるかに多くの異なった意味の「あ」を音声は表現しうる。

それに対して今、読者が見ている文字は音声言語ではない。

聴覚よりもむしろ視覚に適合するよう構造化されたメッセージである。文字テキストを処理する際には、読者にはメッセージの、全体ではないにせよ、かなりの部分が見えているのがふつうである。書き手の意図を推測する場合に、読者は文章の前に戻ったり先へ進んだり、ページを行ったり来たりすることができる。テキストの意味は形式的な文法によって表現される。この形式的な文法は、語彙項目を最大限構造的で視覚的に組み立てるために発達してきたものである。テキストは時間の流れとは独立して、1つの固まりとして読むことができる。一方、テキストと違って発話は、時間が経つにしたがって減衰していく短期記憶の保管庫に合うよう、より小さく記憶しやすい単位に区切られなければならない。言語コードは、書き手の意図を形どおりに読者に伝える慣習だが、発話コードは言語コードより解読がむずかしい。言語コードの場合と似た慣習が発話に関しても発達してきているが、それら

の慣習は、メッセージの構造を伝えるのに、視覚情報にたよらず、その代わりに、聴覚信号のもつ柔軟さを最も効果的に利用している。

2. 1. 発話の基本要素

図２が示すように音声言語は階層的な構造の中に複数の情報を持っている。音素が組み合わさって音節を作り、音節が組み合わさってフットを作る。フットが組み合わさればフレーズができ、フレーズが組み合わされば発話ができる。したがって、音声信号に含まれている情報をうまくモデル化するには、発話解釈の理論は、韻律構造と単音連鎖の構造のどちらも考慮しなければならない。このモデルがなくては、発話処理は、発話形式と発話意図が一対一で対応したきわめて単純な発話しか対象にできなくなってしまう。しかし、音声言語の階層性の解釈には、リズム構造もイントネーション構造も関わってくる。

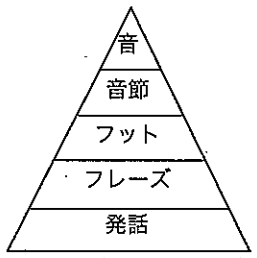

図２　発話音声の階層的構造

発話音声の基本単位（音素）には、聞こえ度の高い音声（母音）から中程度の音声（流音・わたり音・鼻音）、さらに聞こえ度の低い阻害音（摩擦音や破裂音）まで、さまざまなものがある（Ladefoged et al 1996）。阻害音の影響で音節どうしは切れたり重なったりするが発話信号は基本的に母音の連続として構成されていると見ることができる。音節の核に聞こえ度の高い母音がなければ、発話は知覚不能なものになってしまうだろうし、これらの組み合わせの変化がなければ、意味は構成できない。発話において、語彙情報は主にこうした音節のレベルで伝達されているが、語彙のどういう解釈が意図されているかは、この音節の組み合わせだけではなく、それらの韻律のバリエーションによって表現されている。

音連続を韻律レベルで変調させるには、主として３つの方法がある。それらは、音の高さ（ピッチ）を変えること、タイミングを変えること、強さを変えることである。音声の高さは主に、聞こえ度の高い、音節の核が担っている。リズムは、聞こえ度の高い音が阻害性の音で中断されるというタイミング構造で表される。

以下で述べるように、韻律バリエーションで伝えられる情報は、音連続で伝えられる語彙項目の情報よりも複雑な場合がある。会話における発話にはこのことが特によく当てはまる。会話の発話では、たとえば「え？」「あ」のような単一音素でできている音声が、談話の進展を表すのによく使われるからである。

2. 2. カテゴリーによる知覚

Fant（1960, 1991）以来、発話音声は、声道や音源要素に起因すると考えられてきた。音源である声門での声帯振動を変えれば、音声の高さが変わる。また声道（つまり唇から喉頭まで）の形を変えれば、発話の音響スペクトル上の特徴が変わり、これらのおかげでわれわれはさまざまな聞こえ度の高い音声を発することができる。

例えば、/a/ という母音は、あごを下げて声道を広げることで産出される。/i/ と /u/ は、あごの位置が比較的高いが、両者の違いは、舌の盛り上がりが前寄りか（/i/）、後寄りか（/u/）ということで、これによって、声道を通る気流のパタンが変わる。/e/ と /o/ は、あごの位置は中ほどで、舌の盛り上がりがどの程度後寄りかによって区別される。英語のような中舌シュワ母音[ə]や、他の言語で観察されるさらに極度な母音の偏倚を持たない日本語の母音は、これだけですべて説明できたことになる。しかし、この日本語の５母音が不変のものであって、それぞれ声道の形態はいつもまったく変わらないと考えるのは誤りである。

声道のさまざまな場所で気流が阻害されると、子音が生み出される。阻害が強いほど、阻害性の高い音が生じる。鼻音性は、気流が鼻腔に流れることによって生じ、摩擦音は、気流がなかば阻害されて乱れる結果生じ、閉鎖音は短時間完全に気流を遮断することで生じる。声道は、前の音声が産出されている段階で次の音声を産出する形に変化し、複数の音声が同時に産出されるようになるので、結果として、さまざまな音声どうしの間で相互作用が引き起こされる。例えば、/k/ という軟口蓋閉鎖音は、直後に前舌母音 /i/ が来る場合と、直後に後舌母音 /u/ が来る場合とでは、舌が触れる軟口蓋の位置

が異なる。文脈に置かれれば、あらゆる発話音がそのような変化をこうむるものである。だが、我々は、それらの発話音を、カテゴリーとして知覚する。たとえば、/i/ の直前の /k/ と、/u/ の直前の /k/ は、音響スペクトルはかなり違っているにもかかわらず、同じ子音として知覚される。

聞こえ度の高い音声を生み出すのに必要な声帯振動の源は、喉頭を通る気流であり、これが声帯をふるわせる。喉頭の高さや緊張度を変えたり、喉頭を通る空気の圧力を変えたりすると、音声信号のピッチや振幅が変わる。音声のピッチの上下は、意味を表すこともある。例えば日本語では、下降ピッチで発話された母音 /e/ は、同意の意を示すものと解釈され、逆に上昇ピッチで発話された /e/ は、疑問、あるいは驚きを表現するものと解釈される。

こういった声道の影響と音源の影響は、大体のところ、互いに独立していると考えることができる。例えば、母音 /a/ の知覚は、その基本周波数を変えても大きく変わらない。高い /a/ の基本周波数を低いものに変えても低い /a/ になるだけで、/i/ や /u/ に知覚されるということはない。ピッチが変わればスペクトルの質は異なり得るが、知覚的な質（どの母音であるか）は変わらない。同様に、ピッチはそもそも /a/ や /i/ などの母音によって本来的に違ってくるものだが、それは音響レベルでの違いであり、聴覚のレベルでは母音が変わっても発話の基本周波数は、変わっていないと知覚される。

発話音声がある程度安定しているという知覚は、錯覚であり、これはカテゴリー知覚の理論で説明されている（Stevens 1972, 1980, Kuhl et al 1992）。音響レベルでは、発話音声のスペクトル上の特徴は、前後の音素や、音素列にかぶさってくる韻律に大きく影響される。前後の音素による影響の多くは、声道が、前の音声が産出されている段階で次の音声を産出する形に変化することや、共調音（異なる音声を同時に産出すること）によって説明がつく。しかし、我々が自動音声処理がうまく行くように発話特徴を適切にモデル化したいのであれば、韻律についても考慮しなければならない。

2. 3. 発話の韻律

類似の音素構成を持つ語彙項目は、多くの言語では、語彙的な韻律特性によって区別される。例えば、英語の語“import”の強勢を、第１音節に置くか第２音節に置くかという違いは、名詞的意味（[輸入]）か動詞的意味（[輸入する]）かという言語的意味を区別している。さらに日本語においても、（あめ＝雨・あめ＝飴のように）語彙的なアクセントの違いが多くの語を区別している。これらの語は同じ音素連鎖をもっているが、母語話者は別々の違った語として知覚する。

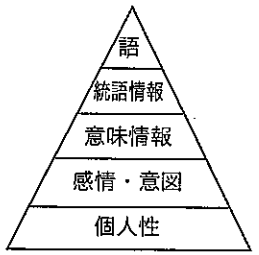

図３　発話韻律の階層的構造

語彙的な韻律だけでなく、句レベルの韻律も、同一フレーズ内の語どうしの関係を示すのに用いられる。１つのフレーズはふつう、波のように上下する発話のピッチ曲線の中で、１つのピークとして現れる。フレーズは、１つのピークと両端の２つの（ふつうは低ピッチの）境界でマークされる。たとえ語１つでできているとしても、フレーズは、発話の流れの中で「意味ある単位」であることを示す韻律的特徴を依然として持つ。

例えば、語「ママ」の２つの音節「マ」は音韻は同一だが、句内での位置が相対的に違うため、これらは同じようにはまず発音されない。最初の「マ」は句頭に位置しているので、比較的高く、比較的短く発音する。後に位置する「マ」は比較的低く、下降調で、比較的長く発音する。自然に発音された語「ママ」に、信号処理をほどこしたり、あるいは録音テープを切りつないだりして、２つの音節「マ」を入れ換えてしまうと、新しくできあがる「ママ」は外国語のように感じられるだろう。「ママ？」という問いかけとして発話される場合は、２番目の音節のピッチは下降調ではなく上昇調になるが、長さは依然として最初の音節よりも長いのが普通である。

いま述べた、語「ママ」の２つの音節の、句レベルでの韻律特徴は、予測可能なものであり、そこには規則がある。句内での位置がスペクトル上の特徴に及ぼす影響は大きい。しかしほとんどの母語話者は、語「ママ」を聞けば、音響的には２つの音節はきわめて違ったものであるにもかかわらず、同じ音節「マ」を２度繰り返して聞いたと言うだろう。というのは、韻律特徴は知覚フィルターにかけられるからである。句の層の情報と語彙の層の情報のように異なった層の情報は大脳外皮での高いレベルの処理において組み合

わされる。

音素の連鎖が同じで、句内での位置も同じだとしても、フォーカスが違えば発音は違ってくる。例えば、“Mama went to Osaka.”という文における“mama”という語のピッチの高さは、話者がこの発話の中で“Mama”と“Osaka”のどちらを新情報としているか(つまり「誰が行ったのか」にフォーカスがあるか「どこへ行ったのか」にフォーカスがあるか)によって変わる。聞き手は、このようなピッチの高さの違いを「意味のあるもの」として処理することができる。身体の状況や個性は人間一人一人みな違うので、さまざまな話者が、さまざまな速度やピッチレンジで発話することになるが、それにもかかわらず、その言語の母語話者は、たとえ話し手のことを知らなくても、意図された意味の違いを直観的に理解できる。

上の文の“Mama”を低いトーンで発話すれば、さらに一つの意味効果が生まれる。つまり話し手は、母親ではなく他の人が大阪に行くと思っていたというような驚きの表現となる。韻律は、どの語とどの語がまとまって句になっているのかというフレージングや、語と語の統語関係や、問いかけなどの強調や、文内でのフォーカスといった特徴を示しながら、同時にまた、強調、驚き、皮肉、熱心さなどのパラ言語的特徴も表す。韻律はさらに、話し手の体格、年齢、性別、個性、気分、感情などといった言語外情報も表現する。

2.4. 韻律と意味

語彙の連鎖はすでにあるものとしよう。これらにかかっていく韻律的特徴を、われわれはどのようにして解釈できるのだろうか。第1に挙げられる最も明らかな韻律効果はチャンキング効果である。つまり、文の中で語どうしがどのように結びついているのかは、韻律上の境界で示される。たとえば、非常によく引用される例だが“old men, and women”という単語の連鎖を見てみよう。この場合、“old”が“men”を修飾しており、“women”の年齢を特定していないということは、文字テキストから明らかである。これはカンマという句読法によって表されている。発話にはカンマなどないが、句末が長くなったり下降したりする(さらに発音もそこで区切られがちである)。ということで、単語の切れ続きが表される。“old men and women”のように1単位として発話される場合は、中間に韻律境界は存在せず、単一のピッチの山が存在するのみであり、形容詞のスコープは男性と女性の両方を含むものとして理解される。このような場合、テキストと発話には1対1の対応があり、統語的な構造や切れ続きは、韻律上の違いで表されている。

韻律は意味のスコープを変えるだけでなく、意味自体を変えることもあるということを、別の例で見てみよう。単純な下降ピッチで話される英語“yes”は、肯定の意を表す。これに対し、上昇ピッチで話された場合、この語は、疑問、つまり確認を求める表現になる。さらに、上昇－下降－上昇というピッチで話されると、それはためらいの標識となり、「あなたの言っていることは理解できるが、私は完全に同意するわけではない」という、ほとんど否定に近い意味を表す。日本語の「そうですね」なども同じような振る舞いを示す。こういった韻律の使用はどんな言語でも共通しているのかもしれない。つまり、上昇調は(たとえば返答を期待する疑問のように)期待を表し、下降調は発話の完了を表し、そして、上昇でも下降でもない中レベルのトーンは継続を表す。人間が円滑にコミュニケーションしていけるのは、このような慣習があるからである。

抑揚のない平坦な韻律で発話された文でさえ、ふつうと違った音色で発話されれば、「強調」、あるいは「対比」として知覚されうる(郡 1989)。例えば、りきんだ声が強調を表すことは、日本語でも英語でもよく行われることである。日本語でも英語でも日常会話では、りきみのような韻律的あるいは発声的な調節は、聞き手にはそれとはっきりわかるもので、複雑な意味を持っており、単語連鎖がどのようなものであれ、その語彙的な意味としばしば相互に影響を与え合う。

どのような語の連鎖なのかがわかっているとしても、単語の連鎖だけで韻律を欠いた音声転写では、非常に多くの解が生じ、理解に時間を要する。発話の韻律的要素を含まずに音素のみを拾う形で記述された自然発話は、書かれた文と比較して、くりかえしや、言い淀み、言い直しなどを含み構造的に欠けていることが多い。しかしその欠落に構造を与えるのが、韻律なのである。例えば、

「　... どうして これ が 出てき あ こういう その 技術 が あのー 考え ら

れてきたのかっていうのは当然そのコーパスがいろいろとその一使えるようになったとかあるいはその計算機があの一非常に能力高くなってですねそういう　...　」(前川 et al, 2001)

3. 0. 音声工学と研究

コンピュータによる音声言語処理の、最も一般的な応用技術は、音声認識、音声翻訳、音声合成などがあげられる。発話を他言語に翻訳するための話し言葉処理には、音声を認識する技術と、翻訳を音声化する音声合成技術の2つの技術がともに必要となる。これらの技術は、電話や自動車内での情報提供サービスに今後ますます利用されていく傾向を持つ。

携帯電話所有人口の爆発的な増加でわかるのは、たいていの人は、個人間でコミュニケーションをとる手段として、音声発話に非常に満足しているということである。音声合成技術の利用がなかなか伸びないのは、この技術がまだ、人々が満足するような形で発話を産出できていないということを示している。おそらくこれは、音声合成が文字情報のみを重視し、言語的内容の再産出に重点を置き、韻律情報をないがしろにして発展してきたためであろう。

3. 1. 音声認識

現在までの音声認識の研究では、韻律情報にほとんど注意をはらわず、韻律情報は統計的モデリングで排除されるべき、信号内の雑音とみなされてきた。音声認識で問われるのは、話し手が意図した意味を韻律から解釈することではなく、つねに、語彙の形式のみであった。

もちろん「車を発進させろ」「ラジオを切れ」といった単純な命令や、「A地からB地まで最短時間で行くにはどうしたらいいか」「このソフトのダウンロード代をどうやって支払えばいいのか」といった情報要求だけに対応する、限られたアプリケーションなら、音声認識は単に単語連鎖を認識するだけで十分である。現在の技術では、このレベルの処理はすでに可能になっている。しかし、(声の抑揚による)肯定の意を示す“yes”と、ためらいの標識としての“yes”を区別することは今もなお不可能である。対話においては、2つの“yes”の違いは決定的なものになる。これを間違って解釈してしまえば、コミュニケーションは失敗に終わってしまいかねない。

3. 2. 音声合成

音声合成は、一般的には未だに「テキスト読み上げ音声合成“text-to-speech synthesis”」と同意だと考えられている。本節では音声合成装置といっても、テキストを読み上げる朗読エンジンと、人間の聞き手に合った形で情報を教えるよう設計されたしゃべるエンジンとでは、いくつかの重要な点が異なっていることを示す。

朗読エンジンとしての音声合成は、書かれた文字テキストを入力とし、単語連鎖を処理して音素連鎖と韻律連鎖を作る。この韻律連鎖は文脈とは無関係なもので、音素連鎖を発話にするために、個々の音声を調節する。音素連鎖を発話に変換する際には、かんたんな統語的情報や意味的情報だけでなく、辞書的な情報も用いられる。しかし文境界を超えるような試みは、2、3の例外(e.g., Hirschberg 1990, 1993)を別とすればほとんどおこなわれていない。話し手の意図に関してはほとんど情報が利用できない。したがって、フォーカスは恣意的に付与されている。入力テキストのフォーカスが、なんらかの事情で明示されている、あるいは(たとえばレーマの前のテーマのような(Halliday 1963, 1967, 1994))フォーカス付与のデフォルトパターンと幸運にも一致するといったことがなければ、朗読エンジンは単語連鎖を発話に変換する際、([ママ、おおさかにいった]は、いくつかの発話意図が考えられる。その1、だれが大阪にいったか、その2、ママはどこにいったか、その3、ママはまだ大阪にいるのか、など)いくつかの可能な変換案を区別するというようなことはない。

先に指摘したように、テキストは発話とは違った形で構造化されている。実際のところ、書き言葉を読み上げ装置のように読み上げる人はほとんどいない。アナウンサーは数年間トレーニングをしたおかげで、テキストの真意を伝える「正しいフレージング」をおこなえるようになっている。また、人間であるからコンピュータなら今後も当分アクセスできないであろう想像や記憶、世界知識の資源にアクセスできる。このような点から、「朗読は音声

合成器でおこなわねばならない」という考えには筆者は強い疑問をもっている。

入力テキストをフレーズに分けたり、或る部分を強調したりすることが正しくおこなえるのに十分な情報を、読み上げエンジンが仮に何らかの方法で得たとしても、テキストの大部分は視覚的な表現に合うよう作られており、テキストの言葉づかいや言い回しは日常会話の構造とほとんど無縁のものである。文字テキストとは、音声の関わらないメディア用に調整されているものである。

ところで発話は、単なるテキストの断片ではないし、「貧弱な文法」しか備えていないわけでもない。発話は発話自体のメディアに合うよう、うまく作られている。テキストを発話に適した形に変換しようと思えば、発話されるメッセージの時間的制約を考慮して、語句を大きく言い換えたり、短縮したり、繰り返したり、単純化したりする必要があるだろう。

一方、しゃべるエンジンは人間の代わりをする音声合成装置である。このエンジンは人間に情報を伝えるためのものである。例えば、駅のアナウンス、カーナビゲーションシステム、天気予報、時報サービス、ユーザーサポートなどが音声合成機器を使っているのは、文章を読み上げるためではなく、しゃべるためである。これらのシステムの目的は、情報を、聞き手がかんたんに理解でき、自分のものにできるような形で提供することであり、それらのシステムはふつう、発話する単語の適切な意味を計算する必要はない。

ほとんどの音声合成アプリケーションは、文字テキストに情報が書き加えられたものが入力として使えるようになっている。文字テキストに書き加えられる情報とは、フレージングやフォーカスやプロミネンスなど、単語どうしの関係を表すものであり、これらの情報は文字テキストの単語列に注記という形で書き加えられている。カーナビ向けに作成されているオンラインニュース配信サービスシステムなどでも、人間が介在して、たとえば名前が正しく発音されているか、ストレスが間違って置かれていないかといった確認をおこなっている。注記付きテキスト（単語読み情報を含んだテキスト）を発話に変換するために必要な、より高次の情報タイプを特定するために、XML ベース (e.g., www.VoiceXML.org) の書き加え構造を国際標準レベルで用いる申し合わせが現在、提案されている。

したがって、将来の音声合成の仕事は、文字テキストを発話に適した形に処理するということよりもむしろ、聞き手が最小の努力で情報を理解し、自分のものにできるような、与えられた単語列と与えられた意図とを結びつけるしゃべり方を見つけだす。つまり、人間がやっているのと同じように音声文法の発話構造を考慮する技術を開発することである。

4. 0. 発話コード

では、音声合成という観点からすれば、意味の違う発話とはどのようなものなのだろうか。我々は、どのようにしてその発話コードを解読できているのであろうか？　あるいは、音をつないでいく音声合成システムが、人間が言えることはすべて言えるようになるには、発話のさまざまなバリエーションをどのような発話単位でとらえたらよいだろうか？

この問いに答えるにはまず、意味の違う人間の発話を数え上げていく必要がある。この作業は、テキストが与えられているものと仮定すればいくらか縮小できる。つまり、与えられた或る文字列に対して、異なる意味を表す言い方を数え上げる必要がある。文字列に依存しないようにこの作業をおこなうには、個々の単語列を抽象化して、意味の違うさまざまな発話イベントのタイプをまとめる、つまり発話コードの基本要素を定義する必要がある。

どんな言語であれ音素の数は有限である。また、どんな話し手であれ韻律のレパートリは有限である。さらに、パラ言語的な意味を表せる音色タイプもごくかぎられている。そこで、発話イベントはこれら３つの軸（音素・韻律のレパートリ・音色）に沿ってばらついていると考え、発話イベントを数え上げる作業のポイントをしぼることができる。

4. 1. 音声のバリエーション

音声認識でも音声合成でも、利用されるのは音韻的にバランスのとれたコーパスである。音声認識の場合、広く用いられている隠れマルコフモデルの必要不可欠な学習材料となるのは、「３連音声」（これは或る音声を、直前の音声、直後の音声と組み合わせたもの。その言語に現れ得る組み合わせがすべてそろっている）に現れるさまざまな音声である。現在最も広く受け

入れられている方法「接続型連続音声合成」(Sagisaka 1988, Hirokawa 1989))の場合も、その言語に現れ得る音声のバリエーションがもれなく再産出できるように、同様の文脈から音声が切り取られてつなぎ合わされる。

さまざまな音声をもれなくカバーすることに関しては、2つの問題がある。それは、発話信号の冗長性の問題と、合成された音声の自然度の問題である。これら2つの問題は互いに無関係というわけではない。音響パラメータを操作して音声を合成すれば(つまり録音された音声をそのまま使うのではなく、声質を変化させて別の音声を作ってしまえば)、できあがった発話音声には最小限の冗長性しかない。これは静かなところで短時間聞くぶんには理解可能だし自然に聞こえるが、騒音のあるところでは品質が落ち、少し長く聞くと聞き手には我慢できないものとなる。

基本的な発話単位のバラエティを限定してしまうと、3連音声文脈はたやすくそろえることができるが、日常会話中に話される音声の自然なバラエティとは(連続音声合成の場合でさえ)合わなくなる。結果的には、スペクトル上のバリエーションや韻律のバリエーションは、主なものしかモデル化できない。自然発話でさえ、録音し、何度も繰り返して聞くと、単調に聞こえだしてくるものである。もしわれわれが自然の対話の中で何かを繰り返すとしたら、聞き手が理解しやすいように、繰り返すたびごとに違った発音をするだろう。

自然さを考慮し、冗長性を最小にするには、発話信号をモデル化する際、音声の自然環境でのバリエーションを最大限考慮する必要がある。大規模なデータベースから音声波形を切りとってつなぎ合わせる音声合成方式(例えば　Campbell 1992, Campbell & Black 1996)を用いれば、微細な韻律の違いを区別し、繰り返しを最小限に止めるための、発話単位の十分なバラエティが利用できるようになる。但し、或る音声のさまざまなバリエーションのうち、どれをどの環境で用いればよいのかは、まだ完全には知られていない。

長年の間、音声合成は、「声道に関するバリエーションと、音源に関するバリエーションは相互に独立している」という仮説に基づいて発展してきた。ほとんどの場合、2つのバリエーションは別々に扱うことができるが、互いに完全に独立しているというわけではない。例えば、高いピッチで発話された母音は、同じ母音が低いピッチで発話された場合と比べると、スペクトル上の特徴が違っている。基本周波数が異なる場合、さらに声の大きさや音源パルスの開放指数、(音楽的な意味での)攻撃性、母音の発声、発話中の減衰パターンといった、他の音響的な変項にも影響が及ぶものである。

規則に基づく音声合成は、自然な発話を再産出するために必要となる、上述のような音響特徴どうしの相互作用を同時にモデル化することは未だにできていない。しかし、自然な発話サンプルに合わせるパラメータ設定作業を手作業でおこなえば、自然性の高い音声合成は可能である(Stevens & Bickley 1991)。

大規模なコーパスに基づく音声合成システムは、自然対話に見られる発話単位のバリエーションをもれなくカバーするために、十分な発話サンプルを取り込むことができる。だが、すべての音響特徴どうしの相互作用をうまくとらえる方法は現在のところ得られていない。このシステムがどの程度適切な発話単位をデータベースから選び出してくることができるかは、音響特徴のモデルがどこまで得られるか、これしだいである。

したがって、「コーパスは音韻的なバランスがとれたもの」という考えは限定的なものであるが、バランスをとるのに必要な発話サンプルの数は、韻律的バランスも含めて考えても、有限だと考えられている。有限であれば、計算処理でき、確実にもれなくカバーできる。

4, 2. 韻律のバリエーション

音声的そしてスペクトル的にはいろいろバリエーションがあるにもかかわらず、人間は音素というカテゴリーを知覚する。それと同様に、発話韻律をカテゴリーとして知覚することもあるだろう。発話における韻律のバリエーションをどう転写するかについては多くのシステムが提案されているが、それらはすべて、ごく限られた数の韻律記号の用法記述にとどまっている。

韻律情報のラベリングシステムどうしの違いは、主に、「ターゲットポイント」を仮定してそれに注記を加えるのか、それとも「ピッチ変動」の曲線を表示するか(Hirst & de Cristo 2000, Crystal 1975)の違いである。韻律予測システムどうしの違いも同様で、基本周波数の曲線をターゲットポイン

トがつながったものとしてモデル化するのか、それともさまざまな層の韻律情報どうしの重なりとしてモデル化するのかの違いである。しかしながら、韻律要素の目録としてごく少数のものしか考えない点ではすべてのシステムが一致している。

現在、世界中のさまざまな言語で広く試用されている韻律記述方式は、ToBI (Tones and Break Indices) システムと呼ばれるものである。このシステムが前提にしている考えは、「大きく３つのレベルを考え、それぞれのレベルの中で２つの目録を考えれば、韻律信号の（少なくとも言語的な使用に関する）バリエーションはとらえることができる」というものである。ToBI システムは発話単位として語彙レベル・フレーズレベル・発話レベルの３レベルを考えており、各レベルで（高い－低いという）２つのトーンを考えている。これとは別に、単語どうしが結びついている度合いは、韻律的な境界指標によって、５段階で表示される (Beckman & Ayers 1993)。

韻律的に意味のある発話タイプが以上のように少数に限られているのであれば、韻律的にバランスのとれたコーパスを設計するために、あらゆる韻律環境におけるあらゆる組み合わせの音韻連鎖を完全に計算処理することは可能と考えることができる。この作業は、始まったばかりであり、韻律的にバランスのとれた発話コーパスの録音作業も、始められつつある。

4. 3. 音色のバリエーション

さまざまな意図で発話される発話バリエーションをとらえる第３の軸は、音色である。先にかんたんに述べたように、平坦な韻律での発話でも、りきんだ声で発すれば、ふつうとは違った意味を表す。例えば、“That's fantastic!” という発話は、本来的には賞賛の意を表すが、あるべきピッチの山がなければ皮肉や失望を表していると知覚される。しかし、発声スタイルによっては、つまり音色を変えれば、素直に感心している意味になる。

発話態度は一種のパラ言語特徴であり、韻律で表されることも多いが、音色で表されることも多い。発声スタイルの不適切な使用は、コミュニケーション上のいさかいのもとにもなっている。

音声合成においては、このような音色の使用に関する研究はほとんどなされていないが、技術が発展し、特にマーケティングや、ユーザーサポートの場面での利用が増えてくれば、この領域の研究は大いに進む可能性をもっている。声のかたさ、やわらかさなど音色が、もたらす情報のイメージは付加要素として重要となろう。

5. 0. 人はどのように意味を伝えるのか？

要約すると、人は言語的意味を表すのに、３つの要素を組み合わせている。それらは音声的要素・韻律的要素・発声的要素である。スピーチアクトの背後にある発話意図を表すパラ言語的意味も、これら３つの要素によって表される。したがって発話理解は（そして発話表現も）、言語的情報・パラ言語的情報・言語外的情報の絡み合いを必然とする、多層的な処理ということができる。

われわれは発話を解釈する際、誰が発話したのか、どこで発話されたのか、誰に発話されたのかという環境場面を解釈の手がかりとする。それと同じように、何が発話されたのか、つまり発話された語句や、どのように発話されたのか、つまり発話スタイルということも、解釈の手がかりとする。パラ言語的な意味、つまり当該発話に対する話し手の態度と、話し手に関する言語外的な事実、たとえば、そのときの気分や、感情、身体の状態などの境界線は明確にひけるものではなく個人の解釈しだいという部分もある。しかし、「最低限」の言語外的情報は、発話を解釈するうえで有効とみなすことができよう。

5. 1. 音声言語処理

現在のところ、音声言語翻訳アプリケーションの用途といえばたいてい、外国旅行用の「旅行会話集」に載っているような対話か、講義やニュース番組のように型が決まっている独話の翻訳である。このような比較的中立な談話文脈では、個人的な情報の交換や、発話態度の表出などは、一般的ではない。しかし、この技術をビジネスに向けていこうとするのなら、談話の中に「交渉」というものが入り込んでくるにつれて、もっと細かな情報の翻訳が必要になってくるだろう。

旅行会話やニュース番組を翻訳するための音声翻訳と、ビジネスの交渉を翻訳するための音声翻訳との間には、単なる難度の違い以上のものがあるかもしれない。そして、会話の対人関係的な側面が重要になってくるにつれて、音声翻訳アプリケーションの開発も、文字テキストの翻訳から、発話意図の翻訳（もっと具体的にいえば、発話の語用論的力を伝えること）へと、パラダイムシフトする必要があるかもしれない。

たとえ音声翻訳アプリケーションの用途が、発話態度が希薄な発話に限定されるとしても、翻訳が満足すべき標準的なものになったことを確認する必要がある。極めて単純な文章であっても、不適切な韻律や音色で発せられれば、誤解をまねく可能性がある。

この種の発話生成と発話合成には、２つのレベルの情報が必要である。それは、どういう単語が連続しているのかという情報と、どういう意味が意図されているのか、つまり、どの単語がどの単語とどういう関係を持って発話されるのかという情報である。上のような場合、単語列は同じでも韻律の違いによって、形容詞のスコープ、統語的な切れ続き、疑問文か平叙文かの区別、新情報か旧情報かの区別、発話の意図がすべて違ってくる。これらの情報は単なる飾りではない。むしろ発話の意味の中で本質的な部分をなすものである。これらを何らかの形で表すことができないのなら、発話の意味はうまく伝わらないだろう。

単に、入力された発話の文字テキストの部分を翻訳するだけでは十分でない。発話がどのようになされたのかが、何らかの形で表示されていることが必要とされなければならない。

5. 2. 表情豊かな発話

音声処理技術が将来さらに必要になってくるという予想に基づき、表情豊かな発話の分析のためのデータ収集と、表情豊かな発話を処理するためのインターフェース設計が、最近始まっている（www.isd.atr.co.jp/esp）{CREST「高度メディア社会の生活情報技術」という一連の研究プロジェクトで、JSTの助成を受けている}。この多方面からの研究プロジェクトによって、大量の自然な日常会話のコーパスが得られることになるだろう。そして、その

コーパスには、さまざまなタイプのパラ言語情報をさまざまな度合いで表すのによく用いられる、種々の音色や発話スタイルのサンプルが含められることになる。

人々は、発話の背後にある意図を表したり、テキストだけでは伝わらない情報を付け加えたりするために音色や発話スタイルを変化させる。そのさまざまな変化に適応する発話インターフェースアプリケーションの基礎として、このコーパスは用いられるだろう。

このコーパスには感情的発話のサンプルも含まれるが、中心になるのは（丁寧さ、ためらい、親密さ、疑い、皮肉などの）発話態度の表現である。このコーパスは、社会的距離や話し手―聞き手の関係から生じる発話スタイルのバリエーションを、具体的に示すものになるだろう。

このようなコーパスから派生した技術を用いて産み出されるもっともわかりやすいアプリケーションは、音声合成にあるだろう。しかし、発話のラベリングと注記を自動化する目的で、このプロジェクトは音声認識技術に関する研究も含んでいる。結果として得られた技術を、現実のインタラクティブな対話に使えるようにすることも計画されている。これは発話翻訳の領域で特によく使用されるだろう。というのは、入力発話のパラ言語情報に対して敏感にならなければ、会話を適切に翻訳するということはあり得そうにないからである。

6. 考察

以上に述べたことから、「発話処理を効果的におこなうには、システムに発話の意味を理解させなければならない」と結論づける必要はない。発話スタイルの違いや韻律パラメータ、音色のタイプを感知できる技術が利用できるようになりさえすれば、入力から出力へのマッピングは、音声認識や音声合成の技術にいま使われているような統計的手法で実現できる。

しかしながら、上記で述べてきた情報が処理されなければ、誤解が生じやすいということは意識しておくべきである。発話は、単語だけでは伝わらない情報も含んだ、多層の情報から構成されている。発話情報を効果的に表現するためには、その発話情報に関連する層はすべて、処理の対象にするべき

である。

7. おわりに

本稿では、情報伝達という観点から発話産出メカニズムの概要を明らかにした。発話の基本要素が、音声的成分・韻律的成分・発声的成分から作り上げられていることを示し、これら３層の情報の影響し合いにこそ、音声言語の「意味」のすべてが表されているということを強調した。

また、本稿では現在の発話音声技術がバランスに欠けていると主張した。すなわち、音韻情報や語彙情報に比重を置きすぎており、だからこそ、人間が表現や通信のメディアとして音声言語のしなやかさを十分に活用しているその方法を、モデル化できないのだと主張した。

文字テキストは、単語だけで適切に意味を伝えることができる代わりに、通常の日常会話にはないような構造化・組織化が施されている。われわれが生み出したい技術が、発話処理の技術であるなら、われわれは音声言語が文字テキストよりもはるかに多くの層の情報を持つように進化してきているということに気づかねばならない。音声言語の文法は、理解され始めたばかりであるが、すでに見たように、音声言語の文法は、文字言語の文法とは、大きく異なっているということができる。

Acknowledgement:

The author wishes to acknowledge the assistance of the JST/CREST, Prof. Miyoko Sugito, Shinji Karita, Kumiko Hayakawa, and the students at Kobe University's Graduate School of Cultural Studies and Human Science in the production of this paper, which is an extended version of an invited plenary lecture given at the 7th annual meeting of the Association for Natural Language Processing, in Tokyo.

参考文献

Beckman, M. E. and Elam, G. A. (1997) "Guidelines for ToBI labeling (version 3, March 1997)". Copyright (1993) The Ohio State University Research Foundation.

坊農真弓（2001）「音声対話における感動詞・応答詞の感情的意味機能―「ああ」を手がかりに」第７回社会言語科学会研究大会予稿集，pp.113-118.

Crystal, D. (1975) *The English Tone of Voice*. Edward Arnold, Ltd.

Campbell, W. N. (1992) "Labelling an English speech database for prosody control", 1-P-8, Proc *ASJ*, Spring, 1992.

Campbell, W. N. and Black, A.W. (1996)「CHATR:自然音声波形接続型任意音声合成システム」『信学技報』SP96-7.

Fant, G. (1960) *The Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G. (1991) "What can basic research contribute to speech synthesis?", *J. of Phonetics* 19, pp.75-90.

Halliday, M. A. K. (1963) "The Tones of English", In WE. E. Jonas and John Laver (eds.) *Phonetics and Linguistics*. London: Longman.

Halliday, M. A. K. (1967) *Intonation and Grammar in British English*. The Hague: Mouton.

Halliday, M. A. K. (1994) *An Introduction to Functional Grammar* (2nd edition). London: Arnold.

Hirschberg, J. (1990) "Using discourse context to guide pitch accent decisions in synthetic speech". In *ESCA Workshop on Speech Synthesis*, pp.181-184, Autrans, France, September 1990. ESCA.

Hirschberg, J. (1993) "Pitch accent in context: Predicting intonational prominence from text". *Artificial Intelligence*, 63.

Hirst, D., Di Cristo, A. and Espesser, R. (2000) "Levels of representation and levels of analysis for the description of intonation systems", In Merle Horne (ed.) *Prosody: Theory and Experiment*, Dordrecht: Kluwer Academic Publishers.

Hirokawa, T. (1989) "Speech synthesis using a waveform dictionary", pp.140-143, Proc *Eurospeech* 1989.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. and Lindblom, B. (1992) "Linguistic experience alters phonetic perception in infants by 6 months of age", *Science*, 255, pp.606-608.

郡 史郎（1989）「強調とイントネーション」『講座日本語と日本語教育２ 日本語

の音韻と音声（上）』明治書院.

Ladefoged, P. and Maddieson, I. (1996) *The Sounds of the World's Languages*. Oxford: Blackwells.

前川喜久雄（2001）ワークショップ「『話し言葉工学』の科学と工学」，平成13年2月28日(水)～3月1日(木)，東京工業大学.

Sagisaka, Y. (1988) "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. *IEEE-ICASSP* 88, pp.679-682.

Stevens, K. N. (1972) "The quantal nature of speech: Evidence from articulatory-acoustic data", In P.B. Denes and E.E. David Jr. (eds.), *Human Communication: A Unified View*. New York: McGraw-Hill.

Stevens, K. N. (1980) "Acoustic correlates of some phonetic categories", *J. Acoust. Soc. Am.* 68, pp.836-842.

Stevens, K.N. & Bickley, C.A. (1991) "Constraints among parameters simplify control of Klatt formant synthesizer", *Journal of Phonetics*, 19, pp.161-174.