音声合成の観点から見た言語音声の特徴

音声合成技術の進歩は目覚しい。
だが、従来の機械音から人間の声をベースにした合成音へと格段に進歩したいま、
合成音声と人間の音声との、いままで気にならなかった微妙な違いが新たに問題として生じてきた。

ニック・キャンベル
(Nick Campbell)

一　人間の言語音はどのような要素で成り立つか

情報社会と言われるこの社会の中で一番重要な情報通信は何か。それは人のコミュニケーションであろう。言語によるコミュニケーション、特に音声を用いた音声言語のコミュニケーションである。日常的に最もよく使うモダリティは声で伝える情報のやり取りである。

現在、初期の頃に比べると機械音声はかなり人間の声に近くなった。合成音声を改良する過程で人間らしい声の特徴が明らかになり、人が伝えようとする「意味」との違いも見えつつある。ここでは人間の言語音がどのような要素から成り立ち、どこに人間らしさが表れるのか、機械音声との対照も含めて探ってみよう。

本稿では、人間の声が持つ情報をテーマとし、以下、「言い方」による「意味の違い」の観点から、音声言語と文字言語の情報の違いを提案する。もちろん、コンピュータ処理の立場からそれらをどれぐらいモデル化できるか、また、音声合成の立場からどれぐらいマネができるかも示せればと思う。コンピュータはよく「計算機」とよばれるが、私はむしろ「情報器」とよぶほうが適当ではないかと思っている。近い将来、多くのコミュニケーションのアプリケーションが、情報提供サービスあるいはカスタマーケアなどにおいて、人間の代わりに人と話すことになるだろう。そのためには、人間に聞きとりやすく、十分な情報をもつコミュニケーション技術が必要である。

特に、表現豊かな音声技術の開発のためには、どのような音声データが必要であろうか。人間同士の日常的なコミュニケーションの中ではどれぐらいのバリエーションがあるか、そのバリエーションの中でどれぐらいの意味の違いをモデル化する必要があるか、といった観点から検討する。

まずは、音声合成、たとえば音声翻訳で利用されるものの進展を概観し、将来の音声合成のニーズを提案しよう。筆者は声が持つ言語的情報、パラ言語情報、非言語的情報の特徴を含めて、人間らしい言い方ができるものを期待している。従来の音声合成は読み上げ朗読のためのものであったが、朗読音声が伝えられるのは、文内容情報のみといえよう。自然対話音声は文内容情報だけでなく、話者の意図、態度、感情を示す。ここでいう朗読とは文読み上げで、発声者（あるいは音声合成器）は作成者ではなく（すなわち、自分の言いたいことを伝えているわけではなく）既にある文字列を音に変換するだけである。この際の発声は文構造や言葉の関係だけで意味を作る。自然発話の場合、発話者は言いたいことを考えながら、単語列とその発話様式を同時に生成する。多くの場合、聞き手は目の前におり、その聞き手の声の表情やあいづちなどの影響を受けながら、発話内容と言い方を決める。すなわち、インタラクティブな行為である。

二　「音」と「意味」との関係

音声言語コミュニケーションとは何か。人が互いに話をするとき、その対話を文字に置き換えるだけでは意図した「意味」を十分に伝えることはできない。声がもつ音情報は、数種類の情報を同時に伝える。言語情報のほかに、パラ言語情報、非言語情報も含んでいるためである。言語情報を発話内容とすれば、話者意図、態度、感情などはその言い方で示されている。話す内容について、どう思っているか（パラ言語情報）、発話者の個性、性別、年齢、体形など（非言語情報）が音によって表現されている。

今まさに書かれているこのページは文字列であり、二次元のものである。上下、左右、フォントの大きさ、字体、レイアウトなどのグラフィック情報から鳥瞰的に文全体の意味構造を解釈することができる。文の書き手は文構造を意識して伝えたい意味をわかりやすいかたちで表現しようとする。読

み手はページ全体という文環境も確認しながら、各文章の意味を解釈する。それに対して、声による音声言語は時間軸に依存して一次元のものである。音は瞬時性のもので、文字テキストは残るが、音声は瞬時に消失し、全体を見ることはできない。しかし、単語情報とともに、声の抑揚、発話リズム、声の調子など（以下、イントネーションと言う）によって、発話内容に構造を与える。

音声では語彙的意味の違い、たとえば「あめ」（雨）と「あめ」（飴）の区別は、文字の違いの代わりにピッチアクセントで示す。統語的意味の違い、たとえば「古い本と雑誌」（本のみが古いか、本も雑誌も古いか）の区別はフレーズアクセントで示す。強調的意味の違い、たとえば「田中さんの本」（だれの本？　田中さんの何？）という区別は文全体の抑揚によって示す。このような読みにおける相違は異なる表現形式によって区別することができる。言語的情報は単語の並びを換えれば、あるいは文構造を換えれば、意味の曖昧性を減らすことができる。合成音声では朗読するとき、文構造を与えるためのイントネーションが付与されている。しかし話し言葉のイントネーションは、文構造以上の多くの情報をもつ。

人と人との日常的な会話では、言語的情報とともに話者の意図や態度を声で示す。特に、あいづちや感動詞などの言い方によって心的状態を示す。たとえば、「そうですね」という音声は、「はい」という意味、あるいは「いいえ」という意味にも伝えられる。実際に録音した日常対話音声データを分析すると約半分以上はこういった短い語句であった。長い、内容をもつ発話は、豊富な言語情報とともに比較的イントネーションパタンが限られているが、それに比べて、感動詞などはイントネーションが意味伝達の上でより重要な役割を果たす。このようなパラ言語情報は音声言語の特徴であり、言語的意味と違って実音声を聞かなければ、話者の伝えたい意味は判断できない。なお、イントネーションの機能的役割についての詳細は『文法と音声Ⅰ』を参照されたい。

三　コーパスベース音声合成の進化

ところで、合成音声は、「人間のように喋る」人工的なコンピュータの声である。現在の合成音声はどのように喋ることができるのか、インターネットを検索すると、いくつかの合成音声サンプルが聞ける。しかしどれも人間らしく喋ってはいない。目的の違いに起因するのであろうが、多くは言語

レベルの情報のみしか表現していない。というのも、歴史的に見れば、音声合成器は話すためのものではなく、書かれたものを読み上げる朗読器だ、という発想があった。合成音声は「声で情報を伝える」が、「話す必要はない」という見方に立ち、入力情報は言語情報のみという枠組みで、ひとつの文字列からはひとつの意味しか伝えることができない。

音声合成の入力、たとえば「今日は、生田町からのニュースを放送します。」のような入力文字列は読みの曖昧性を含む。文字列「今日は」は、「こんにちは」と発声するのか、「きょうは」と発声するのかを選択しなければならない。同じく「いくたちょう」か「いくたまち」かも同様である。漢字仮名混じり文からの適切な読みを予測し、文構造による韻律パタンの予測を行う。これらの統計手法のパラメータから合成音声を作成し音を発声する。

従来は音声波形そのものもパラメータから作られていた。しかし、現在は波形接続手法による音声合成の方が多く利用されている。パラメータから作成した音声波形は音韻の特徴をはっきり実現できたが、音はあくまで機械音であるため人間らしさには遠かった。一方、近年では音声データベースから選択した音声波形を用いている。これは九十年代からの波形接続手法で録音した音声から各音素毎の実声サンプルを選択し、単位データベースを作成する手法である。そのサンプルデータベースから新たな波形を組み立てる。このことで出力音声の質は大きく改良された。しかしなお、ピッチとタイミングを予測したパラメータに合わせる必要から信号処理を行った。波形単位データベースの音声波形サンプルを分解して、違う形式で組み立てるその処理結果は、正しいイントネーションを得ることには成功したが、声の自然性が逆に阻害されていた。

信号処理によって損なわれた自然性の回復の方法として、現在提案されている大規模音声データベース波形接続音声合成は、適切な韻律をもつ音韻波形サンプルを増やすことによって、信号処理を除き、人間らしい声により一層近づくことができた。しかしこの飛躍は知覚や認識における別の問題を前面に押し出した。つまり、合成音声における機械音から人間の声への進化が人間らしさを実現したそのらしさの程度が増すほどに、今まで問題として気にかけられることのなかった人間の音声表現と合成音声とのより微細な相違が意味をもつようになってきたのである。

四　言語情報とパラ言語情報

右記の手法により、黒柳徹子さんの声（六〇分のカセットブック録音音声）から合成データベースを作成した。約三万音素の音サンプルを含む。その中にたとえば、/a/の音はおよそ五〇〇〇ある。その中に同一のものは含まれない。それぞれは、ピッチの違い、長さの違い、声の強さ、文の位置による言い方の違い、フレーズやアクセント句の環境からの違い等により、多くのイントネーションパタンが選択可能な音韻サンプルとなった。これらの合成単位を利用し、目的とするイントネーションや音韻環境を選択基準として適切な特徴をもつサンプルを並び替えるだけで黒柳さんらしい合成音声ができた。しかし、本人の声そのままであっても、かならずしも本人の喋り方にはならない。たとえば、天気予報の内容を合成音声で出した場合、元音声が悲しい声であるため、天気予報の意味が変わる。カセットブックの内容は「鮒」と「うそつき卵」（向田邦子作）であった。それを読むには悲しい声は適切であるが、その話し方は天気予報の内容には一致しなかった。「明日あめがふる」と合成音声で発声する場合、泣き出しそうな声は別の意味を示す。

言い方に焦点を絞れば、日常会話の一〇〇時間音声データの分析結果では、そのおよそ半分程度の発話が、あいづちや感動詞などの一音節語や二、三音節のくりかえしである。残り半分については文字化することにより意味がだいたい定義できるが、前者の「簡単語」は言い方を確認しなければ意味を判断することができない。

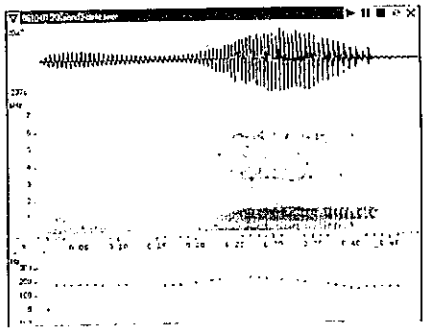

図1　「ほんま」の音声波形、スペクトログラム、ピッチ曲線

日常会話では話しながら同時に考えるため、文字言語より発話文の文構造は簡単である（チョムスキーのcompetenceとperformanceではないが）。書き言葉と話し言葉は異なる。「え」「えー」「あ」「あー」「はい」「そうか」「はいはいはい」などは多くの感情や意図・態度を表現する。このような心的情報を表すことばは、音響的あるいは韻律的に単純であるが、日常の会話において、コミュニケーションの観点から約半分程度の情報をコード化する。辞書に登録しやすい内容語のことばは、音声技術の観点から簡単に処理することができるが、この複数の意味をもつ「簡単語」の方が意味解釈において役割は大きい。図1が示す「ほんま」（関西弁）はそのひとつの例である。明るい、ソフトな声で、ピッチの変化があまりないこの例の音声を聞くと発声者の興味の深さが聞こえる。

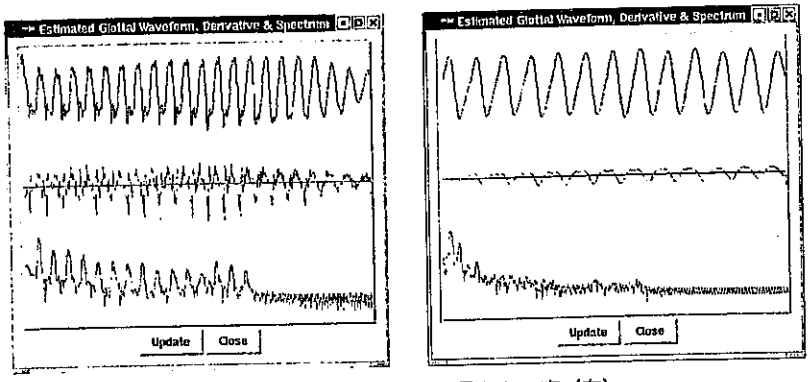

図2　「あ」の声質の違い　硬い声（左）柔らかい声（右）
（上は声門波形、中は声門波形の変化率、下はパワースペクトルを示す）

イントネーションパタンと音色との組み合わせによって、伝える意味は異なる。たとえば、驚きを示す場合はピッチの変化が大きくなる、悲しい場合はピッチが低くなる、疑問を表す場合は文末ピッチが上がるなど。しかし、イントネーションだけではなく、声の質の違いも同様に意味の違いを表現する。図2の母音/a/の声質の違いを例として示す。左は硬い、怒ったような声、右は同一発声者による、同じ母音の柔らかい、親しみやすい声である。図3は三つの異なる「ほんま」の例で、上から、同情を表す言い方、謝罪を表す言い方、驚きを表す言い方を示す。

上記の声がもつ情報を将来の表現豊かな音声合成に応用するために、大規模な日常の対話音声サンプルのコーパスが必要である。しかし、自然な音声対話データを収録するのは意外に難しい。ここで社会言語学者ラボヴが言うObserver's Paradoxの問題が浮上する。つまり、人の前にマイクを置けば、その人の話し方は自然でなくなる。その問題を解決するため、ボランティア話者が常に小型マイクを頭に装着し、一日中スイッチを入れたままという方法で、二五〇時間の対話データを収録した。最終目的時間は一〇〇〇時間のデータ

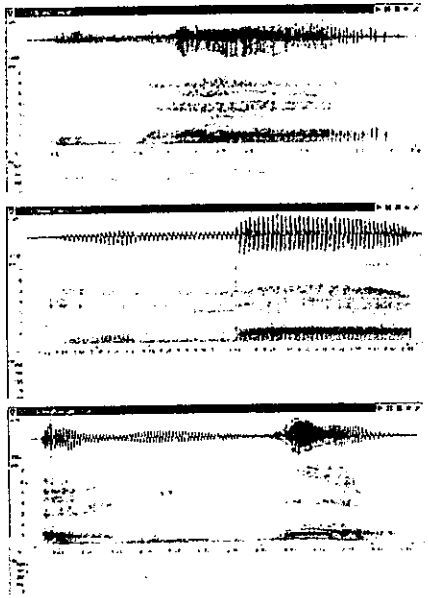

図3　3種類の「ほんま」の音声データ
（上-同情，中-謝罪，下-驚き）

であるが、その中の一〇〇時間を分析した結果を以下で示す。

五　音声データの分析について

データは、まずはじめに人手により書き起こし、発話単位ごとに区切って各単位毎に発話様式ラベルを付与する。もとの目的は喜怒哀楽を区別する感情ラベルを付与する目的であったが、日常対話の音声データにおいて、不適切であるとわかった。声が持つ情報はそんなに簡単ではない。その中の話者状態情報と発話様式情報は一致しないことが多い。たとえば、ある場面で学校の教師が、やかましくする生徒に向かって怒った声で「しずかに！」と怒鳴る。声は怒っていても、先生自身は怒っていない。逆に、先生自身が非常に怒っている場合でも、教室では怒った声を示さないこともある。一発話に対して一つのラベルを付与するのは不十分である。情報には次元がある。すくなくとも話者に対する情報、発話に対する情報、声の調子に対する情報。この三つのカテゴリーの情報を含む、声が伝える言語外情報（supra-linguistic variation）を区別しなければ、適切な自然対話音声データベースにはならない。

カテゴリー毎に、別々に3種類のラベルが必要であるとわかった上で、カテゴリー内にも数種類のラベルが必要であるとわかった。たとえば、声の調子に対して、明るさ、硬さ、エネルギーを6段階（表1）に分けラベルを付与する（0を除いたので7段階スケールは使っていない）。

また、態度、感情、意図に対して尺度が考えられる。具体的にパラ言語情報と非言語情報から示され、少なくとも以下の5種類の情報を区別する必要がある。

(ア)　発話者と聞き手との関係：年齢、性別、上下関係など文化枠における次元の差

(イ)　発話内容に対する本気度：叙述、思い出、意見、主観と客観、演技、主張など、言い方の違い

(ウ)　話者状態・話者感情：表現されたものと実際の違い

(エ)　発話意図・発話行為：言い方による話者の基本目的と発話からみえる目的

(オ)　声質（音響的情報）：硬さ、強さ、明るさ、ハスキーさ、色っぽさなど

現在、表2のように、このようなラベル付与を進めている。ラベル付与の目的は二つである。一つは音声文法のための基本単位を検討する。それとともに音声合成単位選択基準の観点からの音声データをカテゴリー化する。ラベル付与の際、判断しにくい場合は、ある発話単位が、別の発話単位の代わりに置き換えられても、同じラベルであれば発話の意味（解釈）は変らないとする。

表1　6点スケール　（ゼロなし）

	Positive	Negative
目立つほど	＋3	－3
割りに	＋2	－2
少し	＋1	－1

表2　声質、発話様式、発話者について、付与するラベルの種類

Voice	Speech	Speaker
柔かさ (softness)	種類 (type)	目的 (purpose)
明るさ (brightness)	意図 (purpose)	感情 (emotion)
活気 (energy)	本気度 (sincerity)	興味 (interest)
	気の使い方 (manner)	積極性 (confidence)
	気持ち (mood)	態度 (mood)
	好意 (bias)	

・Voice　声の要素は三つの次元で表す。声の明るさの程度、声の柔かさの程度、声の強さやパワーの程度。これらは話者の声の使い方を定義し、話者の個人性ではない。
・Speech　発話様式の情報として、簡単にいえば一発話における話し方の定義。種類カテゴリーは怒った1、怒った2、挨拶1、ぶりっこ等自由に表現のラベルを与える。意図カテゴリは、確認、挨拶、承諾等すでに対話分析に使われている発話アクトラベルを利用する。本気度は演技性、経験からの判断、主張性などを含む度合い。気の使い方は、その話し方から受け止められるもので、丁寧さや、言い方の表情のようなものである。気持ちは、一発話毎の話者態度の判断。好意は、一発話から分かる話し手と聞き手の関係に関わった受け止め。親近度や、皮肉、甘え等など。
・Speaker　話者状況の情報として、話者自身の態度・状態を示すため、以下のカテゴリーを試している。発話アクト（目的）、話者の気分（態度）、喜怒哀楽の四種類のみ、狭義の感情（感情）、対話への関心度（興味）、話者の期待度（積極性）という観点から区別し、また、これらの判断は、対話の流れから得る。

六　「感情音声」と「環境音声」

感情音声合成が次の段階であるという見方がある。しかし必ずしもそうとは言えない。「感情音声」というより、「環境音声」合成と考えるほうが適当ではなかろうか。パラ言語情報や非言語情報の分析から、話し手が聞き手とどのような関係にあるか、どのように話し手が発話内容と関係しているか、という二つの観点でまとめよう。

聞き手との関係によって、つまり誰に話すかによって、もちろん語彙的選択も変わるが、同時に発話様式も定義される。その枠内でやさしい言い方、硬い言い方などの選択の自由がある。発話様式は相手とのその時の関係を示し、社会的文化にも依存する。

発話内容との関係は、本気度のように、命題にどれぐらい関与しているか、どれぐらい感じているか、どれぐらい重要か、などの程度を示す。演技であるか、ただの情報伝達か、教えたい情報か、相手の考えを変えることを要求するほどの主張的情報であるかなどの次元である。

この二種類の関係から発話環境の基本構造が得られる。親しみやすい音声合成、つまり人間と同じように音声言語を利用する技術の開発のため日常的喋り方も含めて、話者間の関係、話者とその内容の関係、両者を含めるものとしての「環境音声」を考えるべきであろう。

七　人間らしい声の使い方

現在の音声合成は言語情報のメディア変換技術である。情報社会のため、人間の代わりに機械が声をだすようになれば、音声言語の特徴も含む必要がある。音声合成の歴史のあゆみの中で、韻律が音韻の特徴であるとわかったことで大きな改良ができた。その時、韻律の役割として単語の区切りや単語アクセント、強調などしか考えられていなかったが、今発話様式のバリエーションを考えてみると、韻律による意味の表現が人間らしさのコミュニケーションと大きく関わることが分かった。つまり心的情報まで韻律によって定義されているのである。

音声データベースの質により、未来の研究が決定される。現在日常対話音声データベースの設計・収集・分析の問題の解決手法はみえてきた。今後、パラ言語情報を含む音声文法の構築が必要になる。上記の五・六節でその基本単位の案を定義したが、音声合成を考慮して、これらの特徴をパラメータとして人間らしい音声合成の単位選択による有効性を確かめる必要がある。

一般に言語教育のカリキュラムにはこういう韻律の使い方までは入らない。これらは当然のことと思われ、言語レベルの韻律役割しか考えない。しかし、われわれの音声言語の重要な部分であることはかわりない。人間の情報コミュニケーションには、言葉の構造だけではなく、声の「色っぽさ」も入るのである。

【参考文献】

Nick Campbell (1997)「プラグマティック・イントネーション」『文法と音声 I』くろしお出版

Nick Campbell (1999)「韻律解釈における基本単位」『文法と音声 II』くろしお出版

Nick Campbell (1995) "Synthesising Spontaneous Speech" in *Computing Prosody*, Sagisaka, Campbell & Higuchi (eds), Springer Verlag.

山本誠一 (2000)「コーパスベース音声翻訳技術」電気情報通信学会論文誌 J83-D-II 8号

JST/CREST 表現豊かな音声処理技術の研究, www.isd.atr.co.jp/esp

（ATR国際電気通信基礎技術研究所／人間情報科学・音声言語処理）