# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 音声情報処理技術 |
| Title | |
| 著者(和文) | 古井貞熙 |
| Author | SADAOKI FURUI |
| 出典(和文) | Advanced Image Seminar '98, Vol. , No. , pp. 27-34 |
| Journal/Book name | , Vol. , No. , pp. 27-34 |
| 発行日 / Issue date | 1998, 4 |

# 音 声 情 報 処 理 技 術

古井 貞熙
東京工業大学大学院情報理工学研究科
〒152-8552 東京都目黒区大岡山 2-12-1
furui@cs.titech.ac.jp

## 1．まえがき

音声情報処理の研究は長い歴史を有しているが、近年、統計的枠組みによる新しい音声情報処理技術の発展、コンピュータの高性能化・低価格化、データベースの拡充などを背景に、大きな技術的進歩を遂げている。種々の新しい応用も展開されるようになってきているが、その一つの重要な分野がマルチメディアコンテンツである。音声情報処理は、音声認識、話者認識、音声符号化、音声合成などに分類できる。ここでは、これらの各分野の最近の技術的発展を紹介し、マルチメディアコンテンツと音声情報処理技術の関連に視点をおいた新しい技術応用について展望する。

## 2．音声認識技術

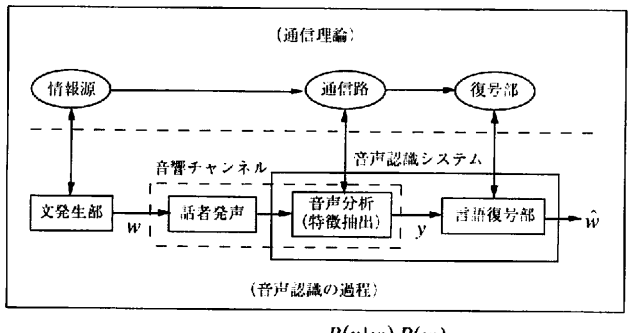
### 2.1 音声認識の原理

音声認識技術は、次のように分類することができる。

(1) (孤立) 単語音声認識：区切って発声した単語を認識する。

(2) 単語スポッティング：連続音声中のキーワードだけを認識する。

(3) 連続音声認識：単語を連続して発声した音声を認識する。文音声認識、会話音声認識などとも呼ばれる。




$$P(\hat{w}\mid y)=\max_{w}\frac{P(y\mid w)P(w)}{P(y)}$$

図1　確率モデルに基づく音声認識システムの構成

音声認識システムは、図1に示すように、音響処理部とそれに続く言語処理（復号）部から構成され、情報処理論・通信理論の問題として、確率モデルを用いて定式化することができる [1] [2]。この際、話者による発話と音響処理部を、一つの音響チャネルとしてモデル化する。話者は情報源に対応する文 $w$ を頭の中で組み立て、それを音声波に変換する。その音声波には通常、話者の個人差、付加雑音、伝送歪みなどが重畳している。音響処理部はスペクトル分析・特徴抽出を行なって、時系列データ（特徴パラメータ系列）系列）$y$ を得る。音響チャネルは雑音を含んだ通信路に対応し、言語復号部に $y$ を送る。言語復号部は $y$ から元の文 $w$ を推定する。$w$ は、ベイズ則によって展開した次式の事後確率を最大化するように推定される。

$$\hat{w}=\underset{w}{\operatorname{argmax}}\,p(w\mid y)=\underset{w}{\operatorname{argmax}}\,p(y\mid w)p(w) \qquad (1)$$

条件つき確率 $p(y\mid w)$ は音素などの音響モデルとの照合によって得られ、文 $w$ が発声される事前確率 $p(w)$ は言語モデルによって得られる。

人が音声を聞いて理解しようとするとき、音としての情報だけでなく、語彙に関する情報（どのような単語があるかという情報）を始め、文法、意味など種々の情報を組み合わせて音声を聞き取っている。多くの場合、状況や話題の展開から無意識に単語や句を予測したり、文脈から推し量って認識している。その一つの理由は、音素情報のあいまい性である。すなわち、通常の音声の中では各音素は必ずしもはっきりと発声されず、その部分だけを聞いたのでは何の音素か分からないことがめずらしくない。このため、コンピュータによる音声認識においても、語彙、文法などの情報を組み合わせて用いることが不可欠である。

$p(w)$ として、単語音声認識の場合は各単語の事前確率だけを、単語スポッティングの場合は連続音声中にキーワードが存在する事前確率を、連続音声認識の場合は単語列の事前確率を用いる。単語毎に区切って発声した文音声を認識するディクテーションでは、音響処理は単語音声認識と同じに、$p(w)$ は連続音声認識と同じになり、単語境界のセグメンテーションの必要がないだけ認識が容易になる。ただし、ユーザにとっては単語に区切った発声の負担はかなり大きく、特に日本語では単語の定義が明確でないため、このような発声は難しい。

### 2.2 音声認識の基本技術

#### 2.2.1 音響モデルと音響処理

条件つき確率 $p(y\mid w)$ を求めるためには、音声の基本的な単位の標準パターンあるいはモデルを蓄えておき、認識すべき音声をそれらの単位と音響的に照合する [2]。音声の基本単位としては、音素から単語まで種々のレベルが考えられる。調音結合、すなわち各音素の物理的特性が前後の音素の影響によって変形する現象があるため、単語を単位とした方が、その変形を自然に取り込めるが、認識対象語彙数が大きくなると、記憶容量と認識のための計算量が膨大になってしまう。このため語彙数が大きいときには、音素

を基本単位として用いて、前後の音素に依存した複数のパターンあるいはモデルを用意することにより、調音結合に対処する。音素の数が数十程度であっても、当該音素の直前あるいは直後のいずれかを考慮する場合（このような音素単位をダイフォンと呼ぶ）で数百、前後両方を考慮する場合（このような音素単位をトライフォンと呼ぶ）には千種類以上の単位が必要となる。なお、音素と単語の中間である音節（日本語の「かな」に対応）や、それを半分に割った半音節と呼ばれる単位を用いる方法も研究されている。

音声の特徴パラメータ系列と、各音素や単語の基本単位との類似の度合いを計算する際に重要な課題の一つは、音声の時間的な伸縮にどう対処するかである。例えば、同じ音素でも、それが含まれる言葉、話者、発声速度などによって長さが変化する。しかも、その変化はゴム紐のように線形（一様）ではなく、部分的に伸びたり縮んだりする非線形な伸縮である。このように非線形に伸縮するパターンを、そのずれを調整しながら照合する方法として、動的計画法（DP; dynamic programming）を用いた方法（DPマッチングと呼ばれる）が広く用いられている [2]。

音声には、時間軸だけでなく、スペクトルの形そのものが、個人差を始めとする多様な要因によって変化する性質がある。この問題に対処するため、近年、隠れマルコフモデル（HMM; hidden Markov model）による方法が広く用いられている [2]。HMMは、確率状態遷移機械であり、各状態遷移確率と、状態遷移に伴う特徴パラメータの観測（出現）確率によって特徴付けられる。これにより、音声の確率的変動が、極めて効率よく表現できる。音声の各基本単位に対応するHMMをあらかじめ作成しておき、入力音声が各HMMから生成される確率（尤度）を計算する。この計算も、DPを用いることによって能率よく行なうことができる。

認識すべき音声に対して、一つの基本単位を選択すればよい場合（典型的には、単語を単位として単語音声を認識する場合）は、これで条件つき確率$p(x|w)$が求まる。しかし、複数の基本単位の連続で音声が表現される場合には、あらゆる種類の基本単位の組み合わせ（並び）の候補と入力音声とを照合する必要がある。例えば7桁の数字の組み合わせは$10^7$の膨大な数となり、膨大な計算が必要となる。しかしこの場合も、DPを基本単位の中とその並びの2段階に用いることによって、大幅に計算量を減らすことができる。語彙数がさらに大きい場合には、言語モデルによって候補単語列を効率的に絞ることが必須になる。

### 2.2.2 言語モデルと言語処理

単語列の事前確率$p(w)$としては、単語の連鎖確率に代表される統計的言語モデル（stochastic language model）がよく用いられる [3]。単語列

$$w_1^k = w_1 w_2 \cdots w_k \tag{2}$$

の生起確率は

$$\begin{aligned} p(w_1^k) &= \prod_{i=1}^{k} p(w_i \mid w_1 w_2 \cdots w_{i-1}) \\ &= \prod_{i=1}^{k} p(w_i \mid w_1^{i-1}) \end{aligned} \tag{3}$$

と表される。ここで、$N(w_1^i)$を言語モデルの学習に用いるテキストデータベースに現れる$w_1^i$の頻度とすると

$$p(w_i \mid w_1^{i-1}) = \frac{N(w_1^i)}{N(w_1^{i-1})} \tag{4}$$

となる。ここで、単語数が2000種類で、$i=4$、すなわち4単語からなる文の場合を考えると、$w_1^4$の可能な異なり単語列の数は16兆（$2000^4$）通りとなり、かなり大量の学習用テキストがあったとしても、有効な統計量を得ることは極めて難しい。このためマルコフ過程を導入して、統計的に可能なモデルとして、マルコフモデル（バイグラム; bigram）や2重マルコフモデル（トライグラム; trigram）を用いる。バイグラムモデルでは、

$$p(w_i \mid w_1^{i-1}) = p(w_i \mid w_{i-1}) \tag{5}$$

と仮定し、トライグラムモデルでは、

$$p(w_i \mid w_1^{i-1}) = p(w_i \mid w_{i-2} w_{i-1}) \tag{6}$$

と仮定する。しかしながら、トライグラムモデルでの異なり単語列の数は80億（$2000^3$）通りとなり、多くのトライグラムが学習用テキストから全く観測できなかったり、わずかしか観測できないということが頻繁に起こる。この問題に対処するため、Katzのバックオフ平滑化（backoff smoothing）法 [4] が広く用いられている。この方法では、観測回数の少ないトライグラムの回数をさらに少なく見積もり、余った確率値の総和を、観測されなかったトライグラムに、バイグラムの値に基づいて分配する。4グラム以上の統計量も同様に推定することができるが、有効性は一般に小さい [5]。

品詞（part of speech）のマルコフモデルを用いることも試みられている [3]。また、単語でなく音素のバイグラムやトライグラムを用いる試みも行なわれている [6]。統計的言語モデルとしては、隣接関係だけでなくやや広い範囲の統計量として、単語の共起関係を用いる方法も用いられている。これらの統計的言語モデルの有効な利用のためには、大量な言語データベースと、それからモデルを推定するための大量な演算が必要であるが、文書の電子化、コンピュータの高性能化などによって、比較的容易に行なえるようになってきた。これにより、従来の人手に頼ることを前提としていた文法を用いるよりも認識性能が大きく向上したが、統計的言語モデルで扱える知識には限界がある。このため、従来からの文脈自由文法、有限状態ネットワーク文法などの中に統計量を取り入れる試みも行なわれている [3]。

### 2.2.3 大語彙連続音声認識システムの構成

典型的な大語彙連続音声認識システムの構成を図2に示す[7]。大語彙連続音声認識では、ボトムアップに処理を行なったのでは、文仮説（認識結果の候補）の数が爆発的に増えてしまう。このため、認識対象タスクに応じた言語モデル（単語辞書、意味情報、構文情報、文脈情報など）をトップダウンで用いて文仮説を生成（予測）し、それを音素系列で表わしたものを逐次的に入力音声と照合して、検証していく方法をとる。

連続音声認識の研究の流れは、ディクテーションと対話音声の認識・理解を対象とした研究に分けられる。前者はいわゆる音声ワープロに相当し、書き言葉を発声したものを文字に変換するものである。そこで認識の対象とされる言葉は、書き言葉であるから、文法的にはほぼ正確であると想定することができる。そのかわり、認識しなければならない語彙や文体が極めて多様であるという点に難しさがある。後者の対話の場合には、対象（タスク）をある分野の情報案内などに限定すれば、ある程度語彙を制限することができるが、話し言葉特有のあいまいな文や、文法的に誤った文も対象としなければならない難しさがある。

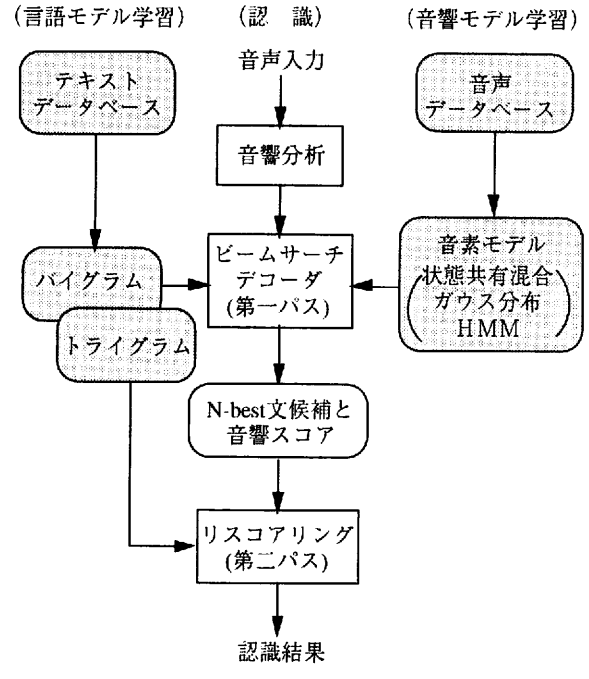



図2　典型的な大語彙連続音声認識システムの構成

### 2.3 連続音声認識技術の進歩

単語音声認識や連続数字音声認識は実用化あるいはそれに近いレベルに達しており、連続音声認識も最近急速に進歩している。世界で最大のプロジェクトは、米国のARPA（Advanced Research Projects Agency）によるものである。その内、ディクテーションについては、Wall Street Journal新聞を中心として、北米の種々の経済新聞（NAB：North American Business News）の記事を読み上げた文を認識するプロジェクトが強力に進められた。音素認識にはトライフォンのHMMを用い、文法には統計的言語モデルを用いて、65,000語の語彙を対象にした場合、93%の単語認識率が得られている[8]。同じくARPAの対話音声の認識・理解プロジェクトとしては、ATISシステム（Air Travel Information System）が、やはり強力に進められた。これは、米国内の航空路線を用いて旅行をすることを想定して、コンピュータによる案内システムに、音声で問い合わせや予約をするシステムである[8][9]。自由に発声した話し言葉を対象とするため、言い間違い、繰り返しなど難しい問題が多数含まれている。どこまで難しい音声を対象とするかによって異なるが、90～95%程度の単語認識率が得られている。ARPAのプロジェクトは、最近ではHub4と呼ばれる放送ニュース音声のディクテーションと、Switchboardデータベースや Call Homeデータベースと呼ばれる日常的な電話対話の認識へと展開している[10]。

日本では、自動翻訳電話[11]、電話番号案内システム[12]、Wall Street Journalに似ていると考えられる日経新聞の記事を読み上げた文を認識する研究[13]などが行なわれている。最近ではNHKの放送ニュース音声を認識する研究も進められている[14]。電話番号案内タスクでは、極めて多数の人の名前や住所を、かなり自由な文体で発声した音声の認識ができるように、約80,000語を語彙とし、文脈自由文法を用いた認識システムの実験が行なわれている。

日経新聞タスクでは、約5年分の新聞記事680万文を用いて、頻度の高い7,000語を対象としたトライグラムに基づく言語モデルを作成し、トライフォンのHMMを用いた認識実験で、90%程度の単語（形態素）認識率が得られている。ニュース音声のタスクでは、約5年分のニュース原稿50万文を用いて、頻度の高い20,000語を対象としたトライグラムを作成し、アナウンサーの音声に対して80%程度の単語認識率が得られている。表1に、ニュース原稿、日経新聞、およびNABの各学習テキストの単語数とカバー率の比較を示す[7]。20,000語でも、ニュース原稿と日経新聞では、約70%しか単語が重複していない。

表1　各学習テキストの単語数およびカバー率の比較

| | ニュース | 日経新聞 | NAB |
|---|---|---|---|
| 学習テキスト(単語数) | 24M | 180M | 237M |
| 異なり単語数 | 114k | 623k | 476k |
| 5kカバー率 | 91.5% | 88.0% | 90.6% |
| 7kカバー率 | 93.7% | 90.3% | - |
| 20kカバー率 | 98.0% | 96.2% | 97.5% |
| 30kカバー率 | 99.1% | 97.5% | - |
| 65kカバー率 | 99.7% | 99.0% | 99.6% |

上述のATISシステムや電話番号案内システムでは、基本的にシステムへの入力は音声認識で行い、システムからの情報は画面上の表や文で受け取る形をとっている。マルチモーダル、あるいはマルチメディア環境でのコンピュータとの情報の授受の方法として、これが人にとって最も容

易かつ便利な方法と考えられるためである [15]。入力手段として音声と画面のタッチ入力を併用して、使い勝手を向上させる研究も行なわれている [16]。

### 2.4 マルチメディアコンテンツへの音声認識技術の応用

最近の大語彙連続音声認識技術の進歩により、書き言葉に近いきちんと発声された音声であれば、ほぼ80%程度の単語が正しく文字に変換（ディクテーション）できるようになってきた。この技術を用いて、ニュースなどの音声を文字に変換し、そこから主たる話題（トピックス）を表す言葉や句を自動的に抽出したり、ほしいニュースを検索したりする研究が、最近活発に行なわれている [17]。ニュースをデータベース化し検索を容易にする上で、このような技術が役に立つと期待される。音声、音楽、雑音などを自動的に判別する研究も行なわれており、これを用いれば音声データやビデオデータから、音声区間だけを検出できると期待されている。

CMU の "Informedia Digital Video Library Project" では、自然言語理解、画像処理、音声認識、および情報検索の技術を組み合わせて、膨大な映像と音声データベースを検索するシステムを構築している [17]。音声認識技術は、音声データベースのディクテーションと、音声による検索の両方に用いられる。

対話音声認識・理解技術の利用としては、WWW ブラウザにマウスやキーボードに加えて、音声による選択や呼び出しなどの操作を可能にすることがある。この発展として、コンピュータ中の仮想的な人物がエージェントとなって、インターネット上の情報を自在に検索してくれるシステム（visual agent system）が考えられる。バーチャルモールの中の店員と対話しながらのショッピングがその一例である [18]。

## 3．話者認識技術

### 3.1 話者認識の原理

音声波に含まれる個人性情報を用いて、誰の声であるかを自動的に判定することを話者認識（speaker recognition）という [2] [19]。話者認識の形態は、話者識別（speaker identification）と話者照合（speaker verification）に分けることができる。図3に示すように、話者識別とは、入力音声が、あらかじめ登録されている $N$ 人の内の誰の声であるかを判定するものであり、話者照合とは、入力音声と同時に自分が誰であるかのIDを入力して、その音声が本当にそのIDに対応する人の音声であるか否かを判定するものである。音声を本人確認の一種の鍵として用いるケースは、話者照合に対応する。いずれの場合も、入力音声と登録されている各話者の標準パターンとの類似度、あるいは各話者のモデルに対する入力音声の尤度を調べ、その値に基づいて判定を行う。話者識別の場合は、多数の登録話者の内から最も類似度（尤度）の高い話者を選び、その話者の音声であると判定する。話者照合の場合は、本人の標準パターンとの類似度（モデルに対する尤度）が、一定のしきい値よりも大きければ本人の音声であると判定し、それ以外の場合は他人の音声であると判定する。

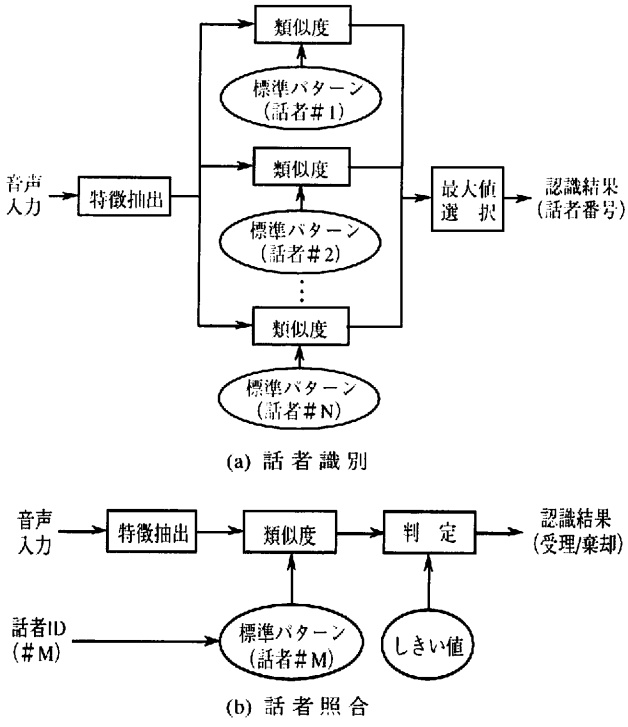



図3 話者識別と話者照合の基本的構成

話者識別の性能は、登録話者の内の本人以外の話者が選択される誤り率で評価する。当然ながら登録話者の数が多くなればそれだけ難しくなるので、話者識別の誤り率は、登録話者の数が増えるにつれて単調に増加する。一方話者照合においては、未知音声が本人の音声である状態（$s$）とそうでない状態（$n$）、その音声を本人の音声と判断して受理する場合（$S$）とそうでないと棄却する場合（$N$）の4つの組み合わせがあり、対応して4種類の確率が定義できる。$p(S|s)$ は正しい受理、$p(S|n)$ は他人（詐称者）の音声を受理する誤り（詐称者受理率：false acceptance：FA）、$p(N|s)$ は本人の音声を棄却する誤り（本人棄却率：false rejection：FR）、$p(N|n)$ は正しい棄却の確率である。

しきい値と2種の誤り率にはトレードオフの関係があり、実用的には、2種の誤りの相対的な重要性に従って設定する（FR の誤りが多少大きくなっても、FA の誤りがある一定値を越えないようにするなど）。しきい値をある値に固定したときのこれらの誤り率は、話者識別と異なり、登録話者の数には依存しない。母集団が同じであれば、登録話者数の増加によって、似た声の話者も増えるが異なる声の話者も同じ割合で増えるからである。

### 3.2 テキスト依存性による話者認識の分類

話者認識の方法はさらに、発声すべき言葉（キーワード）をあらかじめ決めておくテキスト依存型（限定型）と、任意の言葉を発声すればよいテキスト独立型に分類できる。テキスト依存型の場合は、標準パターンと入力音声に含まれる物理特徴を直接比較すればよいので、音声認識の音響

処理部とほぼ同様の方法を用いることができる。キーワードを話者によって違えておけば、その違いも手がかりとして用いることができるので、それだけ認識性能を高めることができる。これまでに実用化されている方式はすべてこの型である。

話者認識、特に話者照合の応用の多くにおいては、キーワードをそれぞれの人について固定することが可能であるが、どのような言葉を発声しても話者を認識することができれば、応用の分野は一層広がると期待される。犯罪捜査などにおいては、必ずしも同じ言葉どうしを比較できず、テキスト独立型が要求される場合もある。人間は通常、発声内容に拘らず相手が誰であるかを判定することができるので、コンピュータを人間に近付けるという意味でも、テキスト独立型の研究の意味があると考えられる。ただし、音声に含まれる個人性情報は、一般に発声内容すなわち音素によって変化し、個人性情報と音韻性情報の間には交互作用があるため、両者を明確に分離するのは極めて難しい。このため、テキスト独立型で話者を認識するには種々の工夫が必要である。一般に、テキスト依存型では認識に2～3秒程度の音声を用い、独立型では、学習に10～30秒程度の音声、認識に5～30秒程度の音声を用いる場合が多い。

テキスト依存型、独立型に共通して、詐称者（本人以外の話者）が本人の音声を録音して装置の前で再生すれば、本人になりすまして装置を容易にだますことができるという問題がある。このためテキスト依存型において、数字や決められたキーワードの発声順序を、認識のたびに装置の側から指定するという方法が試みられている。しかしこの方法でも、最近の電子化された録音装置を用いれば、比較的容易に任意の単語系列を再生することができるので、万全ではない。このためテキスト指定型話者認識法 [20] が提案されている。

テキスト指定型話者認識法では、認識装置を用いるたびに、装置の側から新しいテキストを、文字あるいは合成音声でユーザに提示し、ユーザが本人の声でそのテキストを正しく発声したと判定できる時のみ、話者照合の受理判定をする。その言葉としては、初めてでも比較的発声しやすい、短い文章を用いる。決められたキーワードの発声順序を変えるのではなく、発声する言葉をその都度根本的に変えてしまう。予想のできない言葉が文字表示あるいは音声合成によって指定されるので、テープレコーダによってだますことは極めて難しくなる。

### 3.3 話者認識に用いる音声の物理特徴

音声による個人識別には、鍵、カード、指紋、サイン、ID番号などを用いる方法に比べて、忘れたり紛失したり、他人に盗まれる心配がない、電話などを用いて遠隔地からでも判定ができる、等の特徴がある。しかし、指紋などと決定的に違う点として、同じ人が同じ言葉を発声しても、音声の波形やスペクトルはそのつど変化してしまうということがある。このため、ある程度時期がたっても変化しにくい物理特徴を探索して用いたり、変化をキャンセルできるような正規化を行うことが必要である。

話者認識に用いる物理特徴の持つべきその他の条件としては、音声波から自動的に抽出しやすいこと、詐称者が真似しにくいこと、伝送系やマイクロホンの変動の影響を受けにくいことなどがある。実際に用いられる物理特徴としては、

・スペクトル包絡（スペクトル概形）……声の質に関連
・ピッチ（基本周波数）……声の高さ、アクセント、イントネーションに関連

などがある。この他に、発声レベル（声の大きさ）、発声速度なども個人差に関係しているが、物理特徴として抽出するのは難しい。ピッチは有用な特徴ではあるが、正確な測定が難しい（特に雑音がある場合）、他人が真似しやすい、ストレス等による話者内変動が大きいなどの問題がある。スペクトル包絡に関しては、他人が真似することは極めて難しく、これによる問題は少ないことが確認されている。これらの特徴を用いる具体的方法としては、テキスト依存型の場合は、これらの特徴の時系列を直接用い、テキスト独立型の場合は、細かい時間毎の特徴を独立に取り扱ったり、統計量に変換してから用いることが多い。

### 3.4 マルチメディアコンテンツへの話者認識技術の応用

電話やインターネットによるバンキングや買物サービス、クレジットカードコール、コンピュータのリモートアクセス、高度な情報提供サービスなどにおいて、ユーザが正しい本人であることを確認することは極めて重要である。このようなサービスでは、指紋などを用いることは不可能なので、話者認識技術により、音声で確認できることが期待されている。車や家や特殊な場所への非登録者の侵入防止のための音声キーなども検討されている。上で述べたように同じ人の声でも変化するので、それを正規化する技術を追及するとともに、実用的には、本人しか知らない内容を発声させて音声認識し、発声内容と声の個人差の両方を組み合わせて用いるなどの方法を用いるのが現実的であろう。

## 4．音声符号化技術

### 4.1 音声符号化の原理

音声波のもつ冗長性を利用して、少ない情報量で表現することを音声符号化（speech coding）あるいは音声情報圧縮と呼ぶ [2]。その基本は、音声信号のもつサンプル値の統計的相関、言い換えればスペクトルの分布の偏りを利用して情報圧縮することであるが、冗長性の背景には、声道による音声生成の物理的メカニズム、聴覚器官の性質から人間が聴くことができる音のダイナミックレンジや周波数分解能に制限があること、言語的構造など、種々の要因がある。音声符号化は、音声のディジタル伝送、蓄積などに必須の技術である。

音声符号化の方法は、波形符号化（waveform coding）方式と、分析合成（analysis-synthesis）方式に分けることができる。波形符号化方式は、波形をできるだけ忠実に表現しようとするものであり、分析合成方式は、音声の生成モデルに基づいて、スペクトル包絡情報と音源情報に分け、音源をパルスと雑音でモデル化する方法である。最近では、

後に述べるように両者を組み合わせたハイブリッド符号化（hybrid coding）方式、すなわちスペクトル包絡情報と音源情報に分けるが、音源情報を波形符号化方式で符号化する方法が広く用いられている。これらの方式の比較と代表例を表2に示す。これらの方式によるビットレートと符号化品質の関係を図4に示す。

表2 音声符号化の3つの方式の比較

| 項　目 | 波形符号化 | ハイブリッド符号化 | 分析合成 |
|---|---|---|---|
| 符号化情報 | 波　形 | スペクトル包絡情報<br>（短期予測係数）<br>音源情報<br>（長期予測係数、振幅、波形、符号帳） | スペクトル包絡情報<br>（短期予測係数）<br>音源情報<br>（ピッチ、振幅、有声／無声） |
| 量子化ビット数（kpps） | 16～64<br>（中・広帯域） | 4～16<br>（狭・中帯域） | 1～8<br>（狭帯域） |
| 対象とする音 | 任意音 | 単一話者の音声 | 単一話者の音声 |
| 問　題　点 | 量子化雑音の限界から符号化速度を下げるの難しい | 処理が複雑 | 音声品質の限界 |
| | | 周囲雑音や符号誤りによる劣化 | |
| 代表的方式 | [時間領域符号化]<br>PCM, ADPCM,<br>ΔM, APC<br>[周波数領域符号化]<br>SBC, ATC<br>[組み合わせ]<br>APC-AB | CELP, MPC, RELP,<br>VELP, ベースバンドボコーダ | LPC(最尤, PARCOR, LSP)ボコーダ<br>ケプストラムボコーダ<br>チャンネルボコーダ<br>ホルマントボコーダ<br>位相ボコーダ |

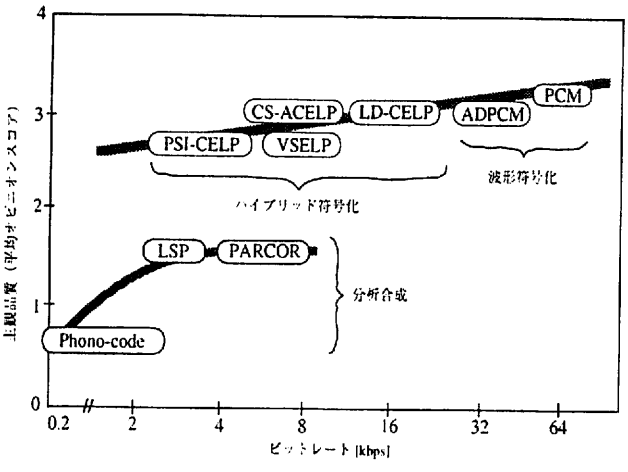



図4 種々の音声符号化法のビットレートと符号化品質

波形符号化方式では、16kbps (kbit/s) 程度よりもビットレートを下げると、急激に品質が低下する。一方、分析合成方式では1 kbps程度までビットレートを下げられるが、逆にビットレートを上げても、音源のモデル化による品質の限界がある。波形符号化と分析合成の両方式の長所を組み合わせることより、4～16kbps程度で高い品質を実現することができる。数100bps以下の究極の符号化方式としては、認識符号化（知的符号化）などが研究されている。

一般に音声符号化の種々の方式は、情報量（ビットレート）、符号化音声品質、装置の複雑さ、および符号化による遅延の4つの要因によって評価することができ、これらの要因の間にはトレードオフがある。音声品質の中には、音声に雑音が重畳したり、伝送路に符号誤りが生じたときの品質の劣化の度合なども含まれる。一般に同じ品質でビットレートを半分にしようとすると、1桁複雑になる傾向がある。装置の複雑さは、それを実現するためのコストに密接な関係があり、一般に復号化器よりも符号化器の方がより複雑になる。音声を伝送せず単に蓄積する場合には、遅延を生じても問題はないが、伝送に用いる場合には、遅延があるとエコーを生じたり快適な通話を妨げる要因となる。

代表的な6種類の音声符号化方式に用いられている基本技術と、4つの評価要因の概略を表3に示す。この内特に携帯電話用符号化などに最近広く用いられているCELP方式の原理を、図5に示す。

表3 代表的な6種類の音声符号化方式の性能比較

| 符号化方式 | 用いられている符号化技術 | 情報量 [kbps] | 複雑さ [MIPS] | 遅延 [ms] | 音声品質 |
|---|---|---|---|---|---|
| PCM (Pulse Code Modulation) | 非線形量子化 | 64 | 0.01 | 0 | 高い |
| ADPCM (Adaptive Differential PCM) | 適応量子化<br>予測量子化 | 32 | 0.1 | 0 | 高い |
| APC-AB (Adaptive Predictive Coding with Adaptive Bit Allocation) | 適応量子化<br>予測量子化<br>時間・帯域分割符号化 | 16 | 1 | 25 | 高い |
| MPC (Multipulse LPC) | 分析合成における残差のマルチパルス化 | 8 | 10 | 35 | 中 |
| CELP (Code-excited LPC) | 分析合成における残差を雑音でベクトル量子化 | 4～8 | 100 | 35 | 中 |
| LPCボコーダ | 分析合成 | 2 | 1 | 35 | 低い |

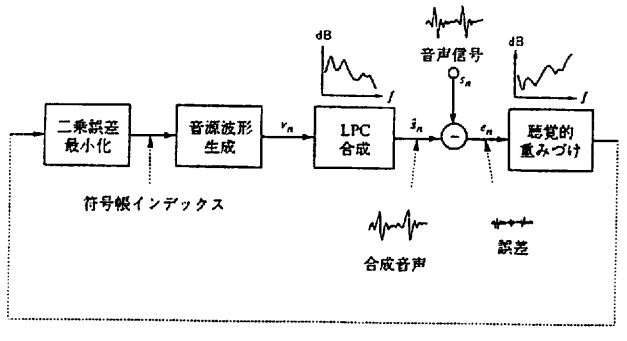



図5 CELP符号化方式の原理

### 4.2 マルチメディアコンテンツへの音声符号化技術の応用

インターネット音声・オーディオサービスへの関心が、最近大いに高まっている [18]。そのサービスは、蓄積再生型と実時間再生型に分けられる。前者はファイルをサーバからダウンロードし、その後に再生を開始するもの、後者はデータの受信と再生が平行同時進行で処理されるので、転送が始まるとすぐに再生が開始されるものである。実時間再生型は更に、サーバにあるデータをユーザがオンデマンドで聴取する放送的なサービスと、ユーザ・ユーザ間の直接通信を行う電話的なサービスに分類できる。いずれにおいても音声・オーディオの符号化技術の性能が、サービ

後に述べるように両者を組み合わせたハイブリッド符号化（hybrid coding）方式、すなわちスペクトル包絡情報と音源情報に分けるが、音源情報を波形符号化方式で符号化する方法が広く用いられている。これらの方式の比較と代表例を表2に示す。これらの方式によるビットレートと符号化品質の関係を図4に示す。

表2 音声符号化の3つの方式の比較

| 項　目 | 波形符号化 | ハイブリッド符号化 | 分析合成 |
|---|---|---|---|
| 符号化情報 | 波　形 | スペクトル包絡情報<br>（短期予測係数）<br>音源情報<br>（長期予測係数、振幅、波形、符号帳） | スペクトル包絡情報<br>（短期予測係数）<br>音源情報<br>（ピッチ、振幅、有声／無声） |
| 量子化ビット数<br>(kpps) | 16～64<br>(中・広帯域) | 4～16<br>(狭・中帯域) | 1～8<br>(狭帯域) |
| 対象とする音 | 任意音 | 単一話者の音声 | 単一話者の音声 |
| 問　題　点 | 量子化雑音の限界から符号化速度を下げるの難しい | 処理が複雑 | 音声品質の限界 |
| | | 周囲雑音や符号誤りによる劣化 | |
| 代表的方式 | [時間領域符号化]<br>PCM, ADPCM,<br>ΔM, APC<br>[周波数領域符号化]<br>SBC, ATC<br>[組み合わせ]<br>APC-AB | CELP, MPC, RELP,<br>VELP, ベースバンドボコーダ | LPC(最尤, PARCOR, LSP)ボコーダ<br>ケプストラムボコーダ<br>チャンネルボコーダ<br>ホルマントボコーダ<br>位相ボコーダ |

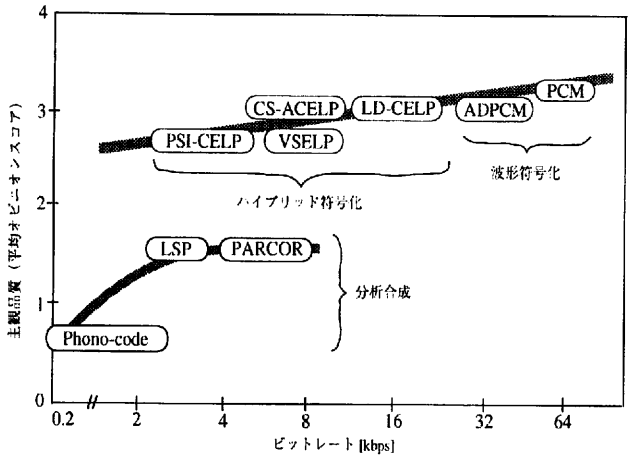



図4 種々の音声符号化法のビットレートと符号化品質

波形符号化方式では、16kbps (kbit/s) 程度よりもビットレートを下げると、急激に品質が低下する。一方、分析合成方式では1 kbps程度までビットレートを下げられるが、逆にビットレートを上げても、音源のモデル化による品質の限界がある。波形符号化と分析合成の両方式の長所を組み合わせることより、4～16kbps程度で高い品質を実現することができる。数100bps以下の究極の符号化方式としては、認識符号化（知的符号化）などが研究されている。

一般に音声符号化の種々の方式は、情報量（ビットレート）、符号化音声品質、装置の複雑さ、および符号化による遅延の4つの要因によって評価することができ、これらの要因の間にはトレードオフがある。音声品質の中には、音声に雑音が重畳したり、伝送路に符号誤りが生じたときの品質の劣化の度合なども含まれる。一般に同じ品質でビットレートを半分にしようとすると、1桁複雑になる傾向がある。装置の複雑さは、それを実現するためのコストに密接な関係があり、一般に復号化器よりも符号化器の方がより複雑になる。音声を伝送せず単に蓄積する場合には、遅延を生じても問題はないが、伝送に用いる場合には、遅延があるとエコーを生じたり快適な通話を妨げる要因となる。

代表的な6種類の音声符号化方式に用いられている基本技術と、4つの評価要因の概略を表3に示す。この内特に携帯電話用符号化などに最近広く用いられているCELP方式の原理を、図5に示す。

表3 代表的な6種類の音声符号化方式の性能比較

| 符号化方式 | 用いられている符号化技術 | 情報量 [kbps] | 複雑さ [MIPS] | 遅延 [ms] | 音声品質 |
|---|---|---|---|---|---|
| PCM<br>(Pulse Code Modulation) | 非線形量子化 | 64 | 0.01 | 0 | 高い |
| ADPCM<br>(Adaptive Differential PCM) | 適応量子化<br>予測量子化 | 32 | 0.1 | 0 | 高い |
| APC－AB<br>(Adaptive Predictive Coding with Adaptive Bit Allocation) | 適応量子化<br>予測量子化<br>時間・帯域分割符号化 | 16 | 1 | 25 | 高い |
| MPC<br>(Multipulse LPC) | 分析合成における残差のマルチパルス化 | 8 | 10 | 35 | 中 |
| CELP<br>(Code-excited LPC) | 分析合成における残差を雑音でベクトル量子化 | 4～8 | 100 | 35 | 中 |
| LPCボコーダ | 分析合成 | 2 | 1 | 35 | 低い |

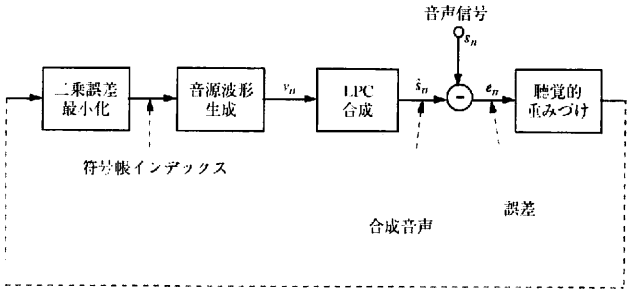



図5 CELP符号化方式の原理

### 4.2 マルチメディアコンテンツへの音声符号化技術の応用

インターネット音声・オーディオサービスへの関心が、最近大いに高まっている [18]。そのサービスは、蓄積再生型と実時間再生型に分けられる。前者はファイルをサーバからダウンロードし、その後に再生を開始するもの、後者はデータの受信と再生が平行同時進行で処理されるので、転送が始まるとすぐに再生が開始されるものである。実時間再生型は更に、サーバにあるデータをユーザがオンデマンドで聴取する放送的なサービスと、ユーザ・ユーザ間の直接通信を行う電話的なサービスに分類できる。いずれにおいても音声・オーディオの符号化技術の性能が、サービ

スの品質を大きく左右する。特に電話的なサービスの場合は、遅延時間の制約が極めて大きい。

ビットレートの低い音声符号化技術では、音声特有の冗長性を用いるため、符号化を行う前に音声と非音声を区別する機能が重要である。一方、オーディオに対しては単純なモデル化はできないので、音声のような高い圧縮率を達成することはできないが、音声よりも時間的な変化がゆるやかなことが多いので、時間分解能を落とし、時間方向の相関を利用して情報圧縮をはかることができる。オーディオ向けに最近開発された圧縮アルゴリズムにTwinVQがある。実時間再生版と蓄積再生版があり、後者ではCDに迫る品質が得られる。

モデムで音声とデータを同時伝送するDSVD（Digital Simultaneous Voice and Data）用の音声符号化方式として、ITU-Tの国際標準方式のG.729（CS-ACELP、8kbit/s）の処理量を減らしたG.729Aの標準化が決まっている。インターネットアプリケーションでの利用が見込まれている。

## 5．音声合成技術

### 5.1 音声合成の原理

人間が音声を発声するしくみをコンピュータや電気回路で模擬し、音声を人工的に生成することを音声合成（speech synthesis）という [2]。音声合成は、あらかじめ録音した音声を適宜接続して再生する録音編集方式と、文字列（テキスト）から任意の音声が合成できるテキスト音声合成方式に分けられる。

録音編集方式では、人が発声した単語、文節、定型文などをアナログまたはディジタル的に録音して、テープ、ディスク装置、LSIなどに蓄えておき、合成したい言葉に応じて、蓄えられた音声をつないで再生する。あらかじめ人が発声した言葉の組み合わせしか合成することができないが、装置が簡単で、高い品質の合成音が容易に作れるので、語彙が限られる駅の案内などの用途に広く用いられている。音声波形をそのまま蓄えるのではなく、ADPCMなどの波形符号化方式により、圧縮して蓄える方式も検討されている。

録音編集方式では、蓄積単位間の接続部における音声の音響的特徴（スペクトル包絡、振幅、基本周波数、発声速度など）の連続性の良否が文音声の品質を左右する。蓄積する単位を、文節、文というように大きくとれば、作り出される音声の品質（明瞭性、自然性など）は良くなるが、合成できる語彙や文の種類が限られる。一方、蓄積単位を音節、音素というように小さくすると、合成できる語彙や文の自由度は増すが、音声の品質が低下する。録音編集方式で現在実用化されているものは、すべて単語以上の大きな単位を蓄積の単位としており、決まり文句の文の中に単語や文節を挿入する形式をとっている。このとき、同じ単語でも、文中の位置によってピッチパターンが異なるので、上りピッチ、平坦ピッチ、下りピッチなどの変形をそろえておく必要がある。

一方テキスト音声合成方式では、単語より小さい音素、音節などの基本単位を用いて、これらの結合法や韻律情報（アクセント、イントネーションなど）に関する規則により、音声合成器の制御パラメータを生成し、音声を出力する。どんな言葉でも合成できるが、録音編集方式に比べると、装置が複雑化、高度化し、合成音の明瞭性、自然性などの品質がまだ十分とは言えない。

### 5.2 テキスト音声合成

#### 5.2.1 テキスト音声合成の原理

テキスト音声合成は、人間がテキストを朗読する能力をコンピュータで実現しようとするものであるから、究極的には、人間における文理解の過程の解明とそのプログラム化が必要である。テキスト音声合成の基本的構成を図6に示す [2]。具体的手順は、次の通りである。

(1) 書かれたテキスト（文）の構文および意味解析（単語辞書の検索を含む）
(2) 各単語の読み仮名、構文・意味情報、および音韻規則による音韻記号への変換
(3) 各単語のアクセント位置、構文情報、韻律情報および韻律規則による継続時間長、アクセント、イントネーション、ポーズなどの決定
(4) 音声合成器の制御パラメータへの変換（合成基本単位の結合、ピッチパターンの生成など）
(5) 音声合成器による音声の生成

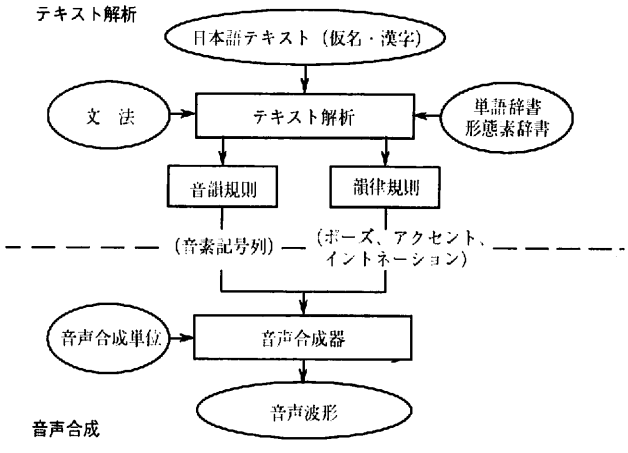



図6 日本語テキスト音声合成システムの基本的構成

#### 5.2.2 テキスト解析

日本語は英語のような言語と違って、句読点を除いて、単語がつながって書かれるので、前項 (1) では、単語に関する知識（辞書）に基づいて、単語境界を検出することが重要なステップとなる。実際には、形態素（語幹、接頭辞、接尾辞など意味をもつ文字系列の最小単位で、単語よりやや小さい）を単位とする解析、すなわち形態素解析を行うことが多い。一般に自立語は漢字で、助詞、助動詞や活用語尾は仮名で書かれることが多いため、このような手掛りも単語（形態素）境界を見出すのに用いることができる。ただし、例えば「米国産業界」という言葉が、「米／国産／業界」、「米国／産業・界」などと区分化できるように、多くの場合複数の区分化の可能性があり、必ずしも左から右へ

の最長一致による解析では正解が得られない。このような場合は、単語の統計的性質や連続関係、意味などを考慮する必要がある。

区分化された各単語の読み仮名やアクセントは、辞書から読み取ることができるが、通常の辞書に書かれている読み仮名は、必ずしも発音に対応しないし、語が連なると読みやアクセントが変化する。このとき、連なった後ろの語の語頭音節が濁音化する現象は、連濁と呼ばれる。例えば、「犬」と「会社」がつながると、後ろの「会社」の読みが／kai ∫ a／から／gai ∫ a／に変わる。前項 (2) の音韻規則では、このような問題に対処できることが必要である。その他の音韻規則としては、「は」を／wa／と読む、語頭以外のガ行音は鼻音化する、／a ∫ ika／（あしか）の／i／のように、無声子音に狭まれた母音／i／と／u／は無声化する、などがある。母音の無声化は、日本語の合成音に自然性を与える上で欠くことができない。

一般に、複合語や付属語がついた語のアクセントは、後につく語の性格によって規定される。このため通常、付属語、接辞類（～性、～学など）の辞書にアクセント結合の型を記述しておき、これに基づいて複合語や文節のアクセントを決定する。

発音記号とアクセントが決まると文節単位の合成はできるが、文として自然にするには、まとめるべき文節はまとめ、切るべきところにはポーズ（間）を入れることが必要である。人間の呼気の量は限られていて、あまり長い文を一息で発声するのは不自然なので、適当な長さまでで切り、しかも文の構造上の切れ目にポーズを入れるのがよい。ポーズとポーズに狭まれた呼気段落が、イントネーションの設定単位となる。

こうして音声合成の制御パラメータが定まると、あらかじめ蓄えられている合成単位と制御規則により、音声波形を合成する段階となる。

### 5.2.3 音声合成器による音声の生成

音声を生成する合成単位としては音素が基本であり、30～50種類ですむので記憶容量は少なくてすむが、この結合規則はかなり複雑で、良い品質を得るのが難しい。このため、各音素に対して前後の音素を考慮したバリエーション（異音）を複数用意したり、音素よりやや大きい単位を用いることが多い。音素より大きい単位を用いることにより、音素を単位としていた場合に規則で表現していた音素間の結合特性を、実音声のデータとして単位内に取り込むことができ、これによって品質が向上する。日本語の場合は仮名文字に対応する100～130個の音節（CV単位ともいう；C：子音、V：母音）が用いられることが多い。よりよい合成音を得るために、CVCを単位とする方式も検討されている。ただし、日本語に可能なすべてのCVCを準備すると5000～6000と膨大な数になるため、高頻度の約1000種類のCVCと、CVおよびVC単位が用いられている。VCV単位を用いる方法も検討されており、この場合は単位の数は700～800となる。たとえば、「さくら」という単語をこれらの各合成単位に分解すると、CV単位：sa+ku+ra、CVC単位：sak+kur+ra、VCV単位：sa+aku+uraとなる。

実際に音声を生成する方法としては、スペクトル近似の原理に基づく線形予測分析法などの音声合成法と、多数の短い音声波形を蓄えておいて接続する方法が用いられる。前者では、パラメータ補間特性のよいLSP法や、メルケプストラム法がよく用いられている。

最近ではコンピュータの処理速度が上がりメモリの価格が下がったため、自然音声から音素長程度の音声波形を大量に抽出して蓄えておき、これを接続して音声を生成する方法が広く用いられている。発声された音素環境や音素内ピッチ形状、振幅、時間長等の情報を波形辞書に登録しておき、テキスト解析の結果として設定された韻律情報と音韻情報に従って、最適な単位波形を選んで接続する。このため、ピッチ単位に波形を切り出し、合成ピッチに合わせて重ね合わせたり、ピッチ波形の繰り返し使用や間引き処理により、韻律情報の制御を行う方法が開発されている。

### 5.3 マルチメディアコンテンツへの音声合成技術の応用

近年、優れた音声合成方法が開発されたこと、LSIやコンピュータの発達によって比較的安価に装置が作れるようになったことなどにより、電話による種々の案内サービスなど、広い範囲で音声合成が用いられるようになってきた。インターネットサービスにおいても、WWWブラウザに音声合成機能を持たせ、テキストや画像を表示するだけでなく、合成音声やオーディオで案内することが考えられる。HTMLの使用を拡張して、HTMLで指定した区間の音声合成を行うことは容易であり、合成方式あるいはデータ形式の標準化がされれば広く使われるようになるであろう。インターネットのコンテンツ制作における音声処理技術も重要性が増してくるであろう [18]。

最近注目されている一つのアプリケーションとして、電子メールの読み上げがある。外出先や車の運転中、あるいは他の用事をしながら電子メールを聴くことができる。視覚障害者にも有用である。ただし、ヘッダの処理や、電子メール特有の視覚的な表示の処理などの問題を解決する必要がある。音声認識と組み合わせて、「次」「もう一度」などの音声で読み上げの制御ができれば、より有用であろう。

## 6．あとがき

音声情報処理技術の最近の動向を紹介し、マルチメディアコンテンツへの技術の適用について述べた。今後ますます画像・映像・テキストなどの異なるメディアや、情報検索・自然言語処理などの異なる分野と、音声情報処理技術との融合が重要になってくるであろう。

### 参考文献


[1] 古井貞熙：" 音声認識"、信学誌、Vol. 78、No. 11、pp. 1114-1118 (1995)

[2] 古井貞熙：" 音響・音声工学"、近代科学社、東京 (1992)

[3] 北、中村、永田：" 音声言語処理"、森北出版、東京 (1997)

[4] S. M. Katz : "Estimation of probabilities from sparse data for the language model component of a speech recognizer",

IEEE Trans. Acoust., Speech, and Signal Process., Vol. ASSP-35, No. 3, pp. 400-401 (1987)
[5] 吉田、松岡、大附、古井：" 単語trigramを用いた大語彙連続音声認識"、信学技報、SP96-82 (1996)
[6] 山田、松永、川端、鹿野：" 音声認識における仮名・漢字文字連鎖確率に基づく統計的言語モデルの利用"、信学論、J77-A、2、pp. 198-205 (1994)
[7] 古井貞熙：" 大語彙連続音声認識の現状と展望"、音学講論、1-6-10 (1998)
[8] Proceedings of the DARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, Inc., San Francisco (1996)
[9] 松岡、Hasson、Dal、Barlow、古井：" テキストコーパスを用いた音声理解のための言語モデル自動獲得"、信学論、Vol. J79-D-II、No. 12、pp.2070-2077 (1996)
[10] Proceedings of the DARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, Inc., San Francisco (1997)
[11] ATR 国際電気通信基礎技術研究所編：" 自動翻訳電話"、オーム社、東京 (1994)
[12] 南、山田、鹿野、松岡：" 番号案内を対象とした大語い連続音声認識アルゴリズム"、信学論、J77-A、2、pp. 190-197 (1994)
[13] 松岡、大附、森、古井、白井：" 新聞記事データベースを用いた大語い連続音声認識"、信学論、Vol. J79-D-II、No. 12、pp.2125-2131 (1996)
[14] 大附、松岡、松永、古井：" ニュース音声を対象とした大語彙連続音声認識と話題抽出"、信学技報、SP97-27 (1997)
[15] S. Furui : "Prospects for spoken dialogue systems in a multimedia environment", Proc. ESCA Workshop on Spoken Dialogue Systems, Vigso, Denmark, pp. 9-16 (1995)
[16] 山口、古井：" 音声対話を用いた情報検索システムの検討"、音学講論、3-6-19 (1998)
[17] Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Morgan Kaufmann Publishers, Inc., San Francisco (1998)
[18] 大室、西、野田、嵯峨山：" インターネットと音声・オーディオ処理技術"、音学誌、52、8、pp.631-636 (1996)
[19] 松井、古井：" 話者認識研究の現状と展望"、テレビジョン学会技報、Vol.20、No. 41、pp. 19-24 (1996)
[20] 松井、古井：" テキスト指定型話者認識"、信学論、Vol. J79-D-II、No. 5、pp.647-656 (1996)