論文 / 著書情報
Article / Book Information

論題(和文)	音声の信号処理
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本信号処理学会誌, Vol. 2, No. 6, pp. 384-390
Citation(English)	Journal of Signal Processing, Vol. 2, No. 6, pp. 384-390
発行日 / Pub. date	1998, 11

展望

音声の信号処理
Speech Signal Processing

古井　貞熙*
Sadaoki Furui*

1．まえがき

音声の信号処理に関する研究は半世紀以上の長い歴史を有しているが [1]，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。音声信号の特徴は，言語としての性質を有することであり，このため音声信号処理の研究も，言語あるいは知識処理としての色彩が強い。これが画像信号処理との極めて大きな違いである。音声信号処理は，音声分析（speech analysis），音声符号化（speech coding），音声合成（speech synthesis），音声認識（speech recognition），話者認識（speaker recognition）などに分類できる。

図1は，音声認識，音声符号化，および音声合成の相互関連を示したものである [2]。図の縦軸は，各階層において音声情報を表現するのに必要な典型的な情報量（符号化速度）を表す。音声波形を保存しながら情報圧縮することを波形符号化（waveform coding），波形を特徴パラメータに変換することによって，一層の情報圧縮をする方法を分析合成（analysis-synthesis）と呼ぶ。これを更に言語的知識を用いて音素や単語などの言語符号に変換するのが

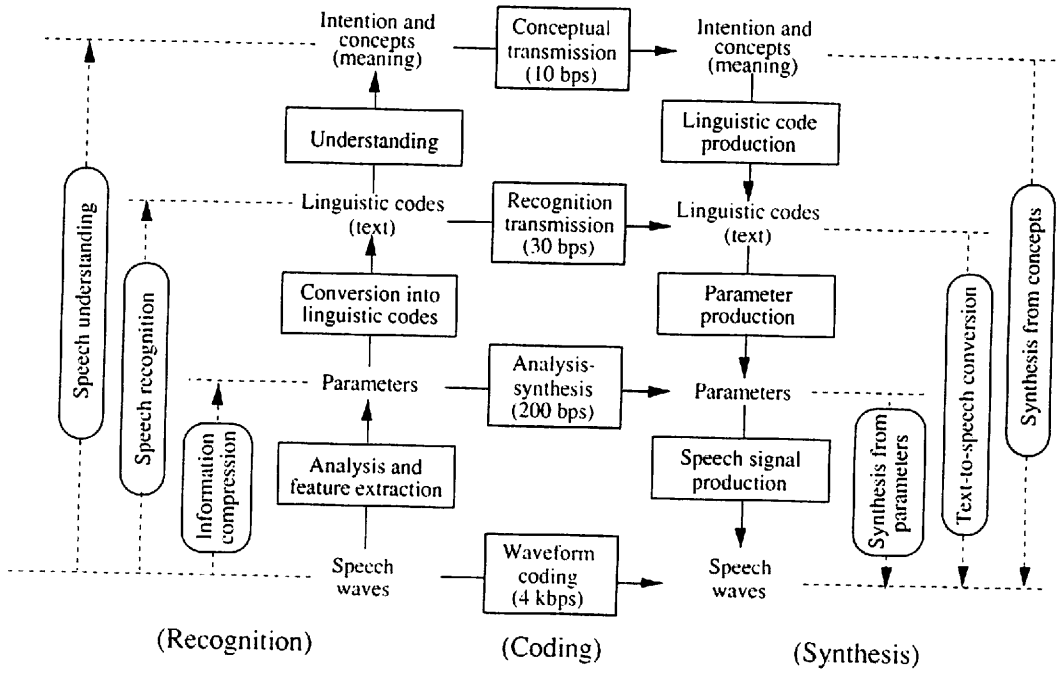


図1 音声認識，音声符号化，および音声合成の相互関連

Fig. 1 Relationships between principal speech recognition, coding and synthesis technologies

* 東京工業大学大学院情報理工学研究科計算工学専攻
〒152-8552 東京都目黒区大岡山2-12-1
Department of Computer Science, Tokyo Institute of Technology,
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: furui@cs.titech.ac.jp

音声認識，この言語符号の形で音声を伝送することを認識符号化と呼ぶ。そのレベルにとどまらず，意味や概念を抽出するのが音声理解，この形で伝送することを概念符号化と呼ぶ。一方，意味概念から文を自動的に生成し，さらに音声波形にまで変換するのが概念からの音声合成，文字で書かれた文を音声に変換するのがテキスト音声合成である。

以下では，音声分析，音声符号化，音声合成，音声認識，話者認識の各分野の最近の技術的発展を紹介し [2]，今後の新しい技術について展望する [3]。

2．音声分析

音声波に含まれる情報は，音源に関する情報と調音に関する情報に分けられる。音源は，声帯の振動（母音など有声音の場合），あるいは空気が口の中の狭いところを通ったり（「す」などの子音），せき止められた空気が急に開放される（「ぱ」などの子音）ことによって生成される。調音は，声道（声帯から唇まで）の形を変えることによって，音源に「あ」や「い」の音韻性を付けることである。

音声波の（周波数）スペクトルは，図2に示すように，スペクトル包絡成分と微細構造に分けて考えることができる。大ざっぱに言えば，スペクトル包絡は調音に対応し，微細構造は音源に対応する。このため音声認識では，もっぱらスペクトル包絡だけが使われる。音声符号化では，まずスペクトル包絡情報を取り出して，これで音声波を逆フィルタすることにより音源を抽出し，符号化する方法が広く使われている。元の音声波よりも，スペクトルのダイナミックレンジが小さくなるので，効率的に符号化ができる。音声合成では，合成すべきテキストを解析して，音源と調音に関する制御パラメータを生成し，これを用いて音声を合成する。

スペクトル包絡を抽出する方法としてよく使われているのは，FFT（高速フーリエ変換）とLPC（線形予測分析法）である。LPCには，少数のパラメータで音声のスペクトル包絡を効率的に表現できる特徴があり，音声分析の基本的方法として，広く用いられている。目的に応じてLPCに基づく種々のパラメータが用いられている。音声のスペクトル包絡を表現する方法として広く用いられているもう一つの方法に，ケプストラム（cepstrum）がある。これは対数スペクトルの逆フーリエ変換で定義される。対数スペクトルの包絡を比較的滑らかな曲線で表現できる特徴がある。cepstrumという言葉は，spectrumの「逆」フーリエ変換という意味で人工的に作られたものである。LPCもケプストラムも，パラメータの数（次数）を制限することによって，表現するスペクトル包絡の滑らかさが制御できる。LPCによるスペクトル包絡を信号のスペクトルと考えて求めたケプストラムを，LPCケプストラムと呼ぶ。聴覚的

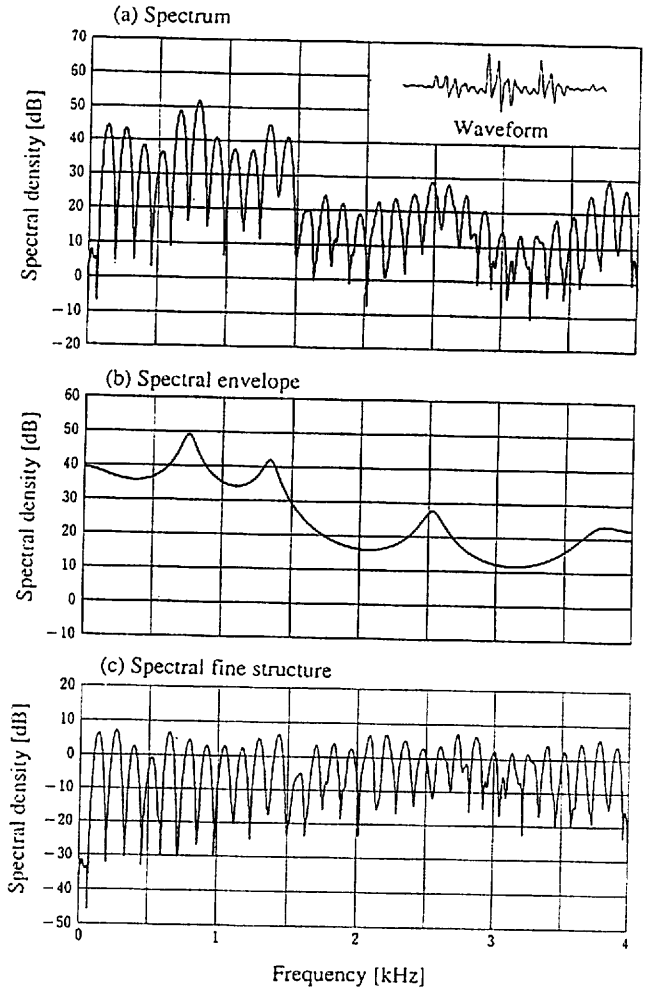


図2　男性母音 /a/ の短時間スペクトルの例 [4]

Fig. 2 Structure of short-time speech spectrum for male voice when uttering vowel /a/ [4]

に重要なスペクトルピークを重視しながら，なめらかな包絡が求まる特徴があり，計算量が比較的少ない特徴もある。このため音声認識などでしばしば用いられている。図3に，短時間スペクトルと，種々のスペクトル包絡の比較を示す。

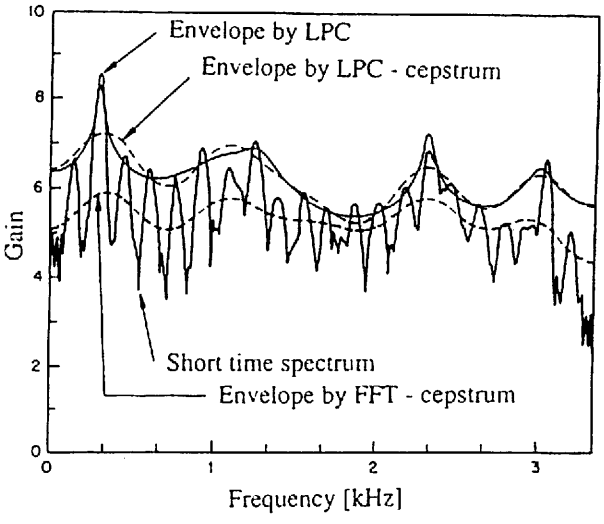


図3　短時間スペクトルと，種々のスペクトル包絡の比較

Fig. 3 Comparison of spectral envelopes by LPC. LPC

音声に含まれる重要な情報にピッチ（基本周波数，基本周期）がある。声帯振動の繰り返し周期（周波数）であり，声の高さを表し，図1のスペクトルの微細構造の繰り返し周期に対応する。音声波をLPCで表したスペクトル包絡の逆フィルタに通した残り，すなわち残差波形の繰り返し周期を求める方法（変形相関法）あるいは，ケプストラムを用いる方法が広く用いられている。

3．音声符号化

音声符号化の方法は，まえがきで述べたように，波形符号化方式と，分析合成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，両者の長所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。代表的な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。日本のディジタル携帯電話などで，最近最も広く用いられているのが，ハイブリッド符号化の一つであるCELP（code-excited LPC）方式である。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。

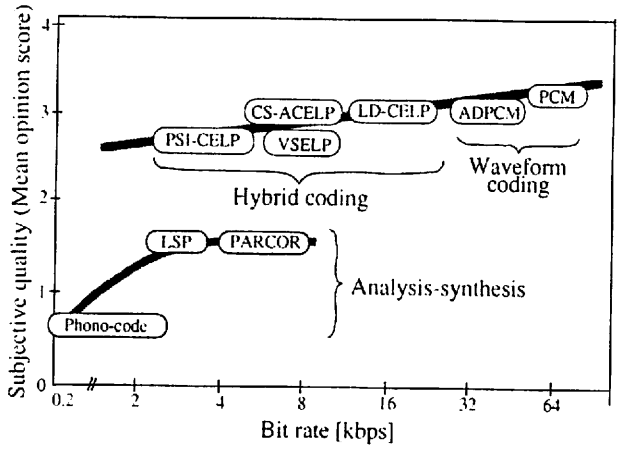


図4 種々の音声符号化法のビットレートと符号化品質

Fig. 4 Coded speech quality versus transmission rate

4．音声合成

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のが音声

合成である。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，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。

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（スペクトル包絡情報と音源情報）を生成し，これにより音声を出力する。どんな言葉でも合成できるが，録音編集方式に比べると，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。最近では，音素や基本周期に相当する短い音声波形を大量に蓄えておき，それらをつなぎ合わせて音声を合成する方法を用いることが多い。

5．音声認識

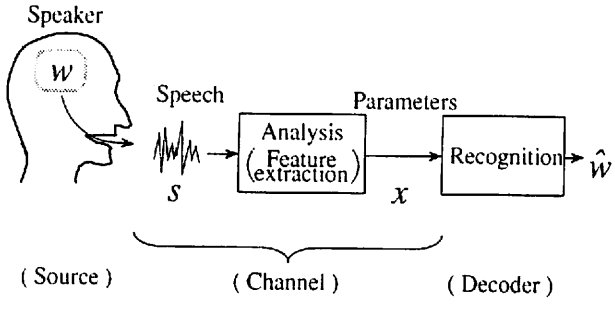
音声認識技術は，次のように分類することができる。

(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。

最近では図5に示すように，情報源，通信路，復号部からなるシステムに対する情報処理論・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多い。すなわち，発声者（情報源）が頭の中で考えた内容（通常は単語列）をwで表すと，wは発声者の発声器官によって音声波形sに変換される。この音声波形には通常，話者の個人差，付加雑音，伝送歪みなどが重畳している。音声認識システムでは，これをマイクロホンで電気信号に変換した後，波形の数値列xとしてコンピュータに取り込む。この数値列は，スペクトル包絡を表す特徴パラメータの時系列に変換される。パラメータとしては，ケプストラムと，その時間変化の線形回帰係数の組み合わせを用いることが多い。この時系列から，元の発声者の頭の中にあった単語列を推定（復号化）する。この際，パターン認識の理論から，事後確率を最大にするwを選べば誤認識を最小にすることが分かっているので，図の式に従ってwを決定する。通常，事後確率を直接計算することは困難なので，wは，ベイズ則によって展開した右辺の式を最大化するように推定される。ここで，条件つき確率$P(x|w)$は音素などの音響モデルから特徴パラメータ列が出現する確率として計算され，単語列wが発声される事前確率$P(w)$は言語モデル（文法）によって得られる。



$$\hat{w} = \underset{w}{\operatorname{argmax}}\, P(w|x) = \underset{w}{\operatorname{argmax}}\, P(x|w)P(w)$$

図5　確率モデルに基づく音声認識システムの構成

Fig. 5 Structure of speech recognition system based on stochastic model

$P(x|w)$の計算には通常，隠れマルコフモデル（HMM; hidden Markov model）を用いる。HMMは確率状態遷移機械であり，図6に示すように，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。音声の時間的変動を考慮した確率計算は，動的計画法 (DP) を用いることによって能率よく行なうことができる。

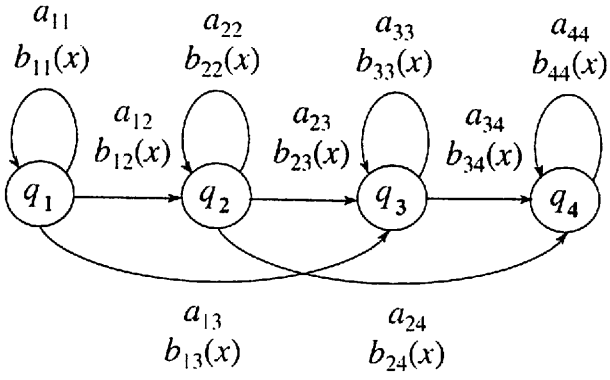


図6　隠れマルコフモデル（HMM）の構成；q_i：状態，a_{ij}：遷移確率，$b_{ij}(x)$：スペクトルパターンxの出現確率

Fig. 6 Structure of HMM (hidden Markov model); q_i: state, a_{ij}: transition probability, $b_{ij}(x)$: observation probablility for spectral pattern x

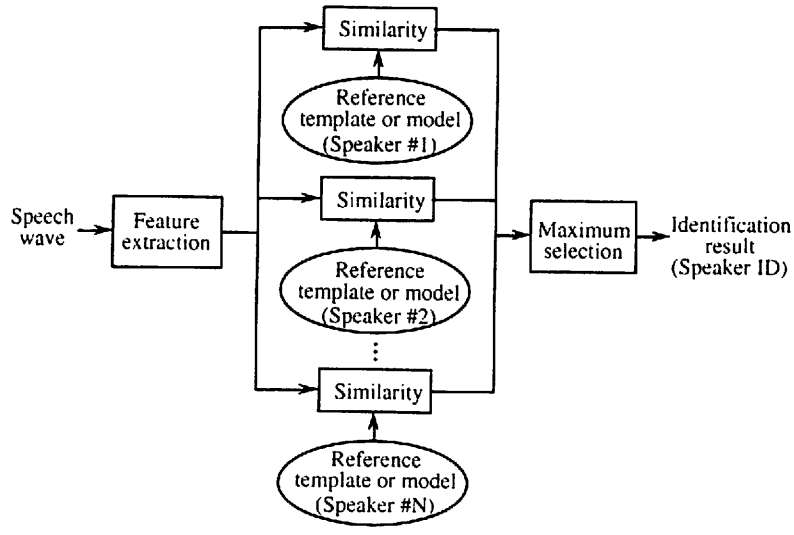
$P(w)$の計算には通常，単語の連鎖確率を用いる。大量のテキストデータベースを用いて，各単語が直前の単語に依存する確率（バイグラム；bigram）あるいは直前の2単語に依存する確率（トライグラム；trigram）を求め，これを掛け合わせることによって，多数の単語からなる系列の確率$P(w)$を計算する。

最近，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，大語彙連続音声認識の技術も大きく進歩している。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。

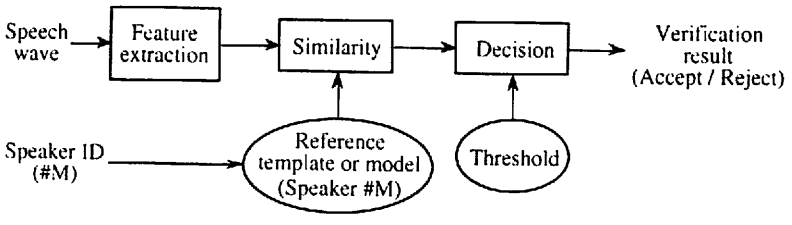
ディクテーションについては，放送ニュース音声のディクテーションの研究が活発に行なわれており，20000ないし65000語を対象として，80ないし90%の単語認識率が得られている。対話音声認識については，航空旅行案内システム，電話番号案内システム，ホテル予約の音声翻訳システムなどの研究が行なわれている。自由に発声した話し言葉を対象とするため，言い間違い，繰り返しなど難しい問題が多数含まれている。どこまで難しい音声を対象とするかによって異なるが，90%程度の単語認識率が得られている。

6．話者認識

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という。話者認識は，話者識別（speaker identification）と話者照合（speaker verification）に分けることができる。図7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（詐称者）



(a) Speaker identification



(b) Speaker verification

図7　話者識別と話者照合の基本的構成

Fig. 7 Principal structure of speaker identification and speaker verification

の音声であると判定する。話者照合の誤りには，他人の音声を受理する誤り（詐称者受理率：false acceptance: FA）と，本人の音声を棄却する誤り（本人棄却率：false rejection: FR）があり，判定のしきい値と2種の誤り率にはトレードオフの関係がある。

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，任意の言葉で判定する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。話者照合の応用においては，キーワードをそれぞれの人について固定できる場合が多いが，テープレコーダによって話者認識装置がだまされる可能性がある。これを防ぐ目的で，テキスト指定型話者認識法が提案されている。この方法では，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。

音声と指紋の違いは，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である。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。

7．音声信号処理の今後の課題

音声信号処理の今後の課題としては，次のようなものが考えられる。

まず音声符号化に関しては，現在よりもさらに低ビット・高品質で，しかも小型かつ廉価（低演算量）な端末で実現できる技術の開発が求められるであろう。音声合成に関しては，合成音声（特に韻律制御）の自然性向上，種々の発声スタイル（朗読調，会話調など）の制御と個性のある声質の実現，テキスト解析の精度向上が課題である。音声認識に関しては，声の個人差，雑音，歪みのある音声に対する精度向上と，これらに対する自動適応機能の実現が必須である。また，言い直し，未知語などを含む会話音声の言語モデルの構築，認識対象（応用場面）が変わっても，容易に言語モデルを適応させる方法の開発，音声から自動的に意味を抽出する方法の開発も重要な課題である。話者認識に関しては，雑音が重畳したり，電話を通したり，風邪を引いたりして声が変わっても安定な認識方法の実現が求められている。

これらの個々の課題の他に，分野間を横断的に扱ったアプローチが必要になると予想される。例えば，音声の個人差に関する研究に，話者認識，音声認識における話者適応，音声合成における声質変換，低ビット音声符号化における声質による性能差への対処などがある。これらの研究はこれまで独立に行なわれることが多かったが，問題の本質的解決を図るためには，従来以上に横断的かつ多角的なアプローチが求められるようになるであろう。

今後の新たな技術的展開としては，人間の音声知覚や音声生成モデルに基づく信号処理が期待される。これまでの聴覚末梢系や発声器官の機能の数学モデルに関する研究を一層進めることも勿論であるが，それに止まらず，脳の機能に立ち入った研究が求められる。特に，音声理解，高度な音声合成などの実現のためには，人間がいかに音声を理解しているか，生成しているかという研究を避けて通ることはできないであろう。

8．むすび

音声信号処理の最近の動向を紹介し，今後の展望を述べた。これまで、信号処理の基本技術の多くが、音声信号を対象として開発されて来た。最近の音声信号処理技術の進歩は著しく，種々の新製品が登場しているが，まだ今後解決しなければならない課題も多い。種々のマルチメディアサービスの展開が期待されているが，その中でも知的音声信号処理の果たすべき役割は大きい。画像・映像・テキストなどの種々のメディアとの協調・融合が重要になって来ると思われる。

参考文献

[1] B. H. Juang, et al. : Speech processing: Past, present and outlook, IEEE Signal Processing Magazine, 1998

[2] 古井貞熙：音響・音声工学，近代科学社，1992

[3] S. Furui : Future directions in speech information processing, Proc. 16th ICA/135th ASA, 1aPLa1, 1998

[4] 東倉洋一：偏自己相関係数を用いた音声分析合成系における音声品質向上の研究，東京大学学位論文，1980

古井貞熙　1968年東京大・工・計数卒。1970年同大大学院修士課程了。同年日本電信電話公社（現NTT）研究所入所。1978年～1979年ベル研究所客員研究員。1986年日本電信電話公社基礎研究所第四研究室長，1989年NTTヒューマンインタフェース研究所音声情報研究部長，1991年同研究所古井特別研究室長，1994年東京工業大学客員教授，1997年同大教授，現在に至る。工学博士。この間，音声認識，話者認識，音声知覚などの基礎研究に従事。IEEE，電子情報通信学会，日本音響学会などから論文賞など受賞。著書「ディジタル音声処理」（東海大学出版会），「Digital Speech Processing, Synthesis and Recognition」（Marcel Dekker），「音響・音声工学」（近代科学社），「音声情報処理」（森北出版），監訳「音声認識の基礎（上・下）」（NTTアドバンステクノロジ）など。