解　説　音声処理技術とその応用

1. 花開く音声処理技術

Speech Processing Technology towards Practical Use by Katsuhiko SHIRAI, Tetsunori KOBAYASHI (Waseda University) and Ikuo KUDO (Justsystem Corporation).

白井克彦[1]　小林哲則[1]　工藤育男[2]

1 早稲田大学
2 (株)ジャストシステム

1. はじめに

近年の音声処理技術の進展は，大変目覚ましいものがあるといえるが，本稿とそれに続く各稿によってその根幹をなす技術を明らかにし，さらに，その応用の実状と将来の可能性について示したいと考える．

音声情報処理の技術分野は，音声符号化，音声認識，話者認識，音声合成などである．音声に関する研究は，当然，これらの技術開発以前から，言語学や医学の中で存在していたが，情報通信の発展につれて，音声の工学的研究が進んだ．音声の通信品質や音声符号化技術は音声通信の基本的問題として，有線通信のはじめから今日の携帯電話に至るまで長い発展の歴史をもつが，やはり，近年，加速度的進歩を遂げているように思う．

技術の飛躍的に発展した大きな理由はかなり明らかである．第1は，やはり，半導体技術の進展である．80年代にはミニコンピュータでも十分な速度で実行できなかった複雑なフィードバックを含む符号化アルゴリズムが，携帯電話の中で実時間で動作している．音声合成でも多量なメモリの利用によって格段の高品質が達成されつつある．第2は，音声に対するモデルのパラメータを大量の実データの解析に基づいて自動的に作成するアルゴリズムが開発されたことである．第3は，米国を中心に実用的な実験研究が進み，大企業，ベンチャ企業から多くの製品が販売されて，フィールドで試されて，発展していることである．

以下本稿では，音声処理技術の現状について概括し，その後の各章で技術のポイントを示す．

2. 音声処理技術が拓く世界

2.1 音声符号化

ここ数年のディジタル携帯電話(PDC)，PHSの出荷台数の伸びは著しい[1]．PDC，PHSでは，音声(64kbit/s)をディジタル信号に変えた後，信号をそのまま送信するのではなく，信号を圧縮してから送信する．PHSではADPCM(ITU-T: G726)符号化により半分(32kbit/s)に圧縮するが，この時の音声品質劣化は僅かである．日本の携帯電話(PDC)では，VSELP(音声6.7＋誤り訂正4.5kbit/s)やPSI-CELP(音声3.45＋誤り訂正2.15kbit/s)符号化によって，音声を通常の約1/6から1/10に圧縮している．このまま携帯電話の需要が伸びていけば，携帯電話に利用可能な電波の帯域が不足する事態が予想され，限られた資源を有効に利用するために，音声品質が劣化しても話の内容が分かる程度まで音声を効率よく圧縮することが行われている．

PHSや携帯電話以外にも，インターネット電話やTV会議システム，DSVD(Digital Simultaneous Voice and Data)にも，音声符号化技術が用いられる．どの符号化を用いるかは，用途(要求される圧縮率，品質，符号化に要する計算量と許容される遅延時間など)により選択される．(本特集「2. 音声符号化技術」を参照)

音声符号化技術は，符号化の対象が音声であることを前提として，音声の性質を利用し極限近くまで圧縮している．そのため，音声以外のデータの圧縮には有効でない☆．音楽用にはオーディオ符号化技術があり，MPEGオーディオ符号化と

☆ 音声と音楽では扱うべき帯域が異なる．電話音声では0.3～3.4KHz，高品質音声では8KHz，音楽帯域では20KHzまで扱う必要がある．

して標準化されている．MPEG4については98年10月頃には決定される見通しである．

2.2 音声合成

2.2.1 テキスト合成

音声合成には，あらかじめ入力した音声を情報圧縮しておき，それを復号化して再生する分析合成方式と，テキストから音声を生成するテキスト合成(Text-To-Speech: TTS)方式とがある．「ありがとうございます」のような決まりきった表現を何度も繰り返す場合には分析合成が適するが，異なる表現を作りたい場合や語彙の追加などが頻繁に生じ，そのつど音声を録音することが困難な場合にはTTS方式が用いられる．TTSの代表的な用途としては，視覚障害者用のホームページの読み上げシステムや，記事の校正のための文章読み上げシステム，カーナビゲーションにおける音声案内などがあげられる．

技術的には，ボコーダタイプのものと波形合成方式とがある．かつてはボコーダタイプが主流であったが，音の加工の度合が大きいことから明瞭性に問題があり，最近では波形合成方式が主流となっている．ATRでは大量のデータベースから選んだ音素片の単純な接続により，音声を合成する方式を提案している．音の加工の度合が小さいことから，従来の方法に比べ明瞭性を格段に改善することに成功している一方，韻律にはまだ問題があり，独特のなまりが感じられたり，感情の付与が難しいなどの問題も残されている．

2.2.2 話速変換

話速変換は，入力音声のスピードを変化させて出力する技術である．研究の発端は，放送局に高齢者からアナウンサが早口で聞き取りにくいという苦情が寄せられたことにあったといわれている．苦手な英語もゆっくり話してもらえば理解できる人も多いことから，外国語の了解性を高めるための応用や教育用システムへの応用が期待できる．逆に発声速度を上げる方向としては，早送りビデオの再生音声にも応用される．

2.2.3 声質変換

声質変換は，合成音に個人性を付与することを目標とした技術である．典型的には，Aさんの声をもとにBさんの声を合成するということが行われる．たとえば，音声翻訳システムにおいて，外国人の外国語音声を認識して，日本語合成音を出力する際に，相手の声の特徴を合成音に含めることで，あたかもその外国人が日本語を話しているように演出するなどの応用が考えられる．

最近では，AさんとBさんの声の中間的な声を合成するという音声モーフィングの研究も進められており，将来的には，特定のキャラクタをもった音声を合成するなども考えらている．

2.3 音声認識

米国では，90年代になってから音声認識のマーケットの開拓期に入っており，いくつかの製品化が進んでいる[2]．MicrosoftはWindows用の音声認識API(SAPI)を定め，98年にはOSにバンドルする予定である．パソコン分野ではApple，IBM，Dragon Systems，AT&T，L&H社，Kurzweil社，Philips社，Kolvox社などが音声認識エンジンを供給している．取り扱う言語も，英語，フランス語，ドイツ語，イタリア語，スペイン語，スウェーデン語や日本語，中国語と多言語化してきている．音声を書き起こすディクテーションシステムのアプリケーションの中心は，医療分野と法律分野が多い．とくに医療分野では，音声認識システムをレントゲンの診断データベースとリンクして使うことが行われるので(医者は，キーボードをタイプするのを嫌うといわれている)，システム開発が必要になることが多い．そこで，エンジンメーカによって提供される認識エンジンのシステムに対する組込みを請う多くのVAR業者が誕生している．また，ここ数年，CMUやSRIなどの研究機関をやめて独立する音声認識ベンチャ企業も多くなってきている．また，音声によるインターネットナビゲータも数社から開発されている．インターネットTVなどへの応用が期待されている．また，米国のような電話社会においては，オペレータにつながるまでに長時間待たされることがある．オペレータを待つかわりに音声認識によるダイヤリングシステムが実用化されている．

日本では，パソコン用に，IBMがボイスタイプディクテーション[3]を，NECがボイスリモコンやしゃべっていいめーるを販売している．パソコンがオフィスから家庭へとその利用の場を拡大するにつれ，さまざまなユーザがパソコンに接する機会が増えている．キーボードとマウスの標準的なインタフェースでは難しいというユーザも増

えてきており，音声入力が新たな入力手段となる可能性もある．音声認識が求められる1つの要因として，機器のサイズの問題がある．携帯端末機器の場合，キーボードが使えないので，音声が有力な手段と考えられている．

研究レベルでは，大語彙の不特定話者連続音声認識がここ数年急速に進歩を遂げている．米国ではDARPA主導で競争開発が行われており，6万語彙の新聞記事の読み上げ音声認識ではすでに90％を超える単語認識率を達成している．現在は，ラジオ音声の認識や，電話での自由な対話を収録したデータに対する音声認識などが試みられている．日本では大学を中心に対話の研究が進められている[4),5)]．

しかしながら，進歩したとはいっても究極的に望まれる姿とは隔たりがあり，多くの課題を残している．大語彙連続音声認識で90％を超える単語認識率を実現したとはいっても，それは学習データとテストデータの発話内容がかなり近い時の話であり，話題が変われば認識率は極端に低下する．新たな話題に適応するタスクアダプテーションは早急に検討する必要がある．

音声認識に影響を与える要素として，実環境における雑音の問題がある．雑音には2種類あり，背景雑音のような加法性雑音とよばれる雑音と電話回線を通ることによる歪みやマイクロホンの距離による伝達特性による歪みがある．背景雑音を除去するための方法として，適応サブトラクション(適応SS)方法や雑音モデルと音声モデルから雑音重畳音声のモデルを構成する方法の研究が行われている．また，マイクロホンの技術も向上している．このように実環境における雑音の問題がかなり進歩を遂げたが，それでも音楽のような複雑なノイズ環境や発話が重なったような場合における雑音処理はまだ問題を残している．

そのほか話者やマイクロホン・回線チャネルの違いに対する適応化技術の開発が進んでいる．(音声認識については本特集「4．音声認識技術」，「5．認識技術の進展」を参照．)

2.4 話者識別，言語識別

2.4.1 話者識別

話者識別とは，話し手が誰であるかを識別する技術をいう．応用としては，音声による鍵として，本人照合や本人確認への応用が期待されている．また，不特定音声認識の精度向上にもこの技術が役立つ．携帯電話などは非常に不正使用が多く，この技術を不正防止用に導入し始めた携帯電話会社もある．ただ，現在でも音声データのどの部分に個人の特徴となるデータが含まれているかは非常に難しい問題であり，時間がたってもその個人であることを認識するための研究がなされている．個人的特徴量と時間的な関係を調べるためのデータベース(同じ話者が同じ内容を1週間後，1カ月後，1年後というように発声する)が大規模に構築されることが望まれている．

2.4.2 言語識別

言語識別とは，何語で話をしているかを識別する技術である．言語によって応答を変えるようなシステム(たとえば，電話でガイダンスする言語を変えたり，自動多言語翻訳システムの入力言語の識別など)への応用が想定される．何カ国語を一度に識別するかということによっても識別率は変わってくるが，実用化にはもう少し時間が必要である．短い発声時間でも信頼性の高い言語識別の精度をあげることが課題である．(詳しくは，本特集「6．音声識別」参照)

3．音声処理技術を取り巻く環境

3.1 マイクロホンの技術革新

マイクロホンには，指向性の有無や，エレクトニックコンデンサ型，ダイナミック型などの種類の違いがあり，周波数特性やゲインが大きく異なる．音声認識などでは，専用マイクロホンを推奨している場合が多いのはこのためである．

最近では，高性能なノイズサプレッション技術をもつactive noise cancellation typeのマイク[6)]が安価(約60ドル)で発売されている．このマイクロホンは，音声認識だけでなく，コンピュータテレホニーにも用いられている．

ソニーのカーナビゲーションシステム用の音声認識システムには，アレイマイクロホンが用いられている．運転席の真上のサンバイザのところに装着する．指向性の強いマイクロホンが3個一列に並べられており，通常のマイクに比べて，15dBくらいノイズを減ずることができるという．車などの騒音の激しい場所で威力を発揮する．

また，Philips社では，マイクロホンにマウスボールを組み込んで，片手でマイクロホンとマウ

スを制御できるようなデバイスを開発している．

3.2 信号処理プロセッサ

音声処理は大量の数値演算を行う．このため音声処理では，信号処理プロセッサDSP(Digital Signal Processor)やRISCチップが用いられることが多い．PCでは，インテルがMMXアーキテクチャ[7]を発表した．従来のPentiumと比較して，音声認識が1.7倍，モデムが1.8倍になるという．これは，信号処理用に57種類の新しい命令セットを用意し，従来の乗算器が10cycleかかっていたものを1cycleで処理できるようにデータの構造を工夫し，演算器などのパイプライン化を行い，高速化したものである．MMX用の信号処理ライブラリが供給されている．ただ，実際には，MMX上での音声認識の性能も規模に依存しているという意見もある．小語彙では，高速化されるが，大語彙になるとポインタ処理が増えるのでさほど高速化されないという指摘もある．

3.3 API(Application Programming Interface)

MicrosoftがWindows95用とNT用に音声認識，テキスト合成用にAPIを定めている．これをSpeech API(SAPI)という．SAPIの基準を満たすソフトは，アプリケーションレベルとエンジン部分とを切り離すことができるわけである．ユーザは自分にあったエンジンを選択することができるというメリットがある．

Novel，IBMなどは，UNIXやOS/2，Netwareにも対応したSRAPI(Speech Recognition API)を発表している[8]．音声認識だけでなく，話者識別技術(Speaker Verification)のAPI(SVAPI)も提案している．

また，SUNを中心としたグループはJAVAによるSpeech API(JSAPI)を作成している[9]．

3.4 開発ツール

3.4.1 フリーソフト

音声研究用のソフト，および，モデル，データがFTPできるようになっているサイトがある[10]．詳しい解説や文献に関する情報も豊富である．

音声ファイルには，Apple用，Windows用，SUN用，Netscape用など各社によりフォーマットが異なる．そのために，フォーマット変換プログラムSoX[11]がある．

日本語についてはこれまで目立ったフリーの音声認識ソフトはなかったが，IPAの援助を受けて大語彙連続音声認識を対象としたフリーソフトの開発が最近始められている．できあがった成果はそのつど公開される予定である．

3.4.2 市販ツール

開発ツールとしては，Entropic Research Laboratory, Inc.からUNIX上で動くHTKと呼ばれる音声認識のツールが販売されている．HMMを用いた標準的な手法はほぼサポートされており，日本の大学でも購入しているところがある．Windows上のツールとしては，Philips，AT&T，DragonSystemなどから音声認識開発ツールが製品化されている．また，音声合成，話者識別，音声符号化などのソフトも販売されている．

3.5 データベース

音声認識合成や話者識別，言語識別の開発には，データが必要になる．米国には，ARPAと米国計算言語学会(ACL)が援助してできたLinguistic Data Consortium(LDC)[12]という組織があり，音声データベース，言語データベースをCD-ROM化し，各研究機関に配布している．このデータが流通したことにより，共通のデータ上でいろいろな手法を評価できるようになり，研究がより客観的になった．また，ビジネスの分野にも大きな影響を与えるようになってきている．LDCを経て供給されたデータが低額のロイヤリティを払うことにより商用利用可能になってきている．日本には，LDCのような組織はないが，日本音響学会，ATRなどから学術利用の目的でデータが公開されている[13)~15)]．しかしながら，商用利用可能な音声データベースは公開されていないのが現状である．(詳しくは，本特集「7．音声コーパス」参照)

4. むすび

音声処理の全般的状況について述べた．紙面の都合で詳細については述べることはできなかったが，本稿に続く各稿を参照していただきたい．

より基本的な内容については，以下の入門書を一読することをお勧めする．音声処理の全般的な教科書としては，文献16)，17)があげられる．音声の音響・信号処理に関しては，文献18)～20)が詳しい．認識の基礎に関しては，文献21)～23)があげられ，若干高度な話題は文献24)で

扱われている．1992年以前の研究動向に関しては，文献25)が詳しい．符号化に関しては，(社)電波産業会(The Association of Radio Industries and Business: ARIB: formaly RCR)[26]やPHS MoU[27]が参考になろう．

また，先端の研究については，以下の資料が有用である．国内の研究会資料としては，情報処理学会の音声言語情報処理研究会，電子情報通信学会・日本音響学会共催の音声研究会，人工知能学会の言語・音声理解と対話処理研究会などがある．国際会議としては，IEEEのInternational Conference on Acoustics, Speech and Signal Processing, International Conference on Spoken Language Processing, ESCA Eurospeechなどが重要である．また，音声メールグループonsei-mail@etl.go.jpに入ると会議の情報などが送られてくる．登録方法は，情報処理学会(http://www.ipsj.or.jp/)音声言語情報処理研究会のホームページ参照．

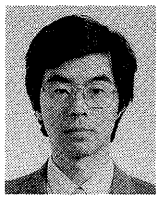
参考文献

1) 竹中豊文編：小特集「パーソナル移動通信」，電子情報通信学会誌，Vol.78，No.2 (1995)．
2) http://www.yahoo.com/Business_and_Economy/Companies/Computers/Software/Voice_Recognition/
3) http://www.software.ibm.com/is/voicetype/
4) 中川聖一他編：特集「音声言語情報処理の現状と研究課題」，情報処理，Vol.36，No.11 (Nov. 1995)．
5) 白井克彦編：特集「音声対話」，人工知能学会誌，Vol.12，No.1 (1997)．
6) http://www.AndreaElectronics.com/
7) http://www.intel.com/drg/index.htm
8) http://www.srapi.com/
9) http://www.sunlabs.com/research/speech/
10) http://www.speech.cs.cmu.edu/comp.speech/ http://www.itl.atr.co.jp/comp.speech/
11) Sound eXchange(SoX) http://www.spies.com/Sox/
12) ftp://ftp.cis.upenn.edu/pub/ldc/ http://cactus.aist-nara.ac.jp/staff/utsuro/ldc_www/
13) 板橋秀一編：小特集「出揃った音声データベース」，日本音響学会誌，Vol.48，No.12 (1992)．
14) 板橋秀一編：音声データベース，人文学と情報処理，No.12 (1996)．
15) 重点領域研究音声対話コーパス http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/
16) 北脇信彦編：音のコミュニケーション工学，コロナ社 (1996)．
17) 古井貞熙：ディジタル音声処理，東海大学出版(1985)．
18) Rabiner, L. R. and Schafer, R. W.: Digital Processing of Speech Signal, Prentice Hall (1978)．
19) Deller, J. R. Jr. et al.: Discrete-Time Processing of Speech Signals, McMillan Pub. Co. (1993)．
20) Ray, D. K. and Charles, R.: The Acoustic Analysis of Speech. 日本語訳：荒井隆行，菅原勉：音声の音響分析，海文堂 (1996)．
21) Rabiner, L. R. and Juang, B-H.: Fundamentals of Speech Recognition, Prentice Hall (1993)．日本語訳：古井貞熙監訳：音声認識の基礎，NTTアドバンステクノロジ (1995)．
22) 中川聖一：確率モデルによる音声認識，コロナ社 (1988)．
23) 北 研二他：音声言語処理，森北出版 (1996)．
24) Lee, C-H. et al.: Automatic Speech and Speaker Recognition, Kluwer Academic Publishers (1996)．
25) 田中和世他：音声の知的処理に関する調査研究，日本情報処理開発協会 (1992, 1993)．
26) http://www.arib.or.jp/
27) http://www.phsmou.or.jp/

(平成9年9月18日受付)

白井 克彦(正会員)
(特別論説「情報処理最前線：ヒューマノイド－人間形高度情報処理ロボット－」の著者紹介を参照)

小林 哲則(正会員)
(特別論説「情報処理最前線：ヒューマノイド－人間形高度情報処理ロボット－」の著者紹介を参照)



工藤 育男(正会員)
1957年生．1983年早稲田大学理工学部電気学科卒業．1985年同大学院修士課程修了．(株)CSK，(株)ATR自動翻訳電話研究所，(株)テキサスインスツルメンツ筑波R&Dセンターを経て，(株)ジャストシステム．自然言語処理，音声認識，CAI，知的財産権の問題に興味をもつ．早稲田大学客員研究員．博士(工学)．情報処理学会学会誌AWG主査．EIP研究会幹事．教育システム情報学会評議員．電子情報通信学会，日本音響学会，言語処理学会，人工知能学会，ACL各会員．e-mail:Ikuo_Kudo@justsystem.co.jp