4.3　音声応用システム

河　原　達　也（京大）

音声は元来，発声された場所・時刻でのみ聞くことのできる揮発的なメディアであるが，19 世紀後半の電話と蓄音機の発明によって場所や時間の壁を超えることができるようになった．20 世紀に入っても音声に関する研究開発は，記録と伝送の安定化・効率化を主な目標としており，音声の処理を行えるようになったのは 20 世紀後半のことである．しかもその大半は，音声の分析及び認識と合成の基礎研究であり，本節で紹介する様々な処理や応用システムに取り組まれるのは 1980 年代以降である．音声を信号処理のみでなく，情報処理として捉えるようになったのもその頃である[*1]．

音声に含まれる情報とその処理を分類したものを**表 2.16** に示す．記録や伝送においては，これらの情報をまとめてできるだけ忠実に扱うことが求められている．一方，認識や合成については，話者認識の研究開発は長く行われているが，パラ言語や感情の認識・合成に取り組まれるようになったのは最近のことであり，基本的には言語情報のみを対象とする音声認識・合成である．このように音声で扱われている主な情報が言語情報であるので，その応用システムは自然言語処理と密接に関係がある．

典型的な音声言語処理とそれらの関係を**図 2.82** に示す．対話・翻訳・検索は自然言語処理においても典型的な応用タスクであるが，これらを音声認識・合成と統合することによって，音声対話・音声翻訳・音声検索といった代表的な音声応用システムが実現されている．ここで音声検索は，音声入力で Web などを検索する場合（通常の「音声検索」：voice search）と，音声アーカイブを検索する場合（spoken term detection）がある．前者は音声認識と情報検索の単純な結合であるが，後者

表 2.16　音声に含まれる情報とその処理

情報の種類	認　識	合　成
言語情報	音声認識	音声合成
パラ言語（文字で表現できない言語情報）		
感情	感情認識	感情を含む音声合成（話者・属性ごとに音声合成）
話者	話者認識	
話者の属性（性別・年齢・出身）	性別認識 など	

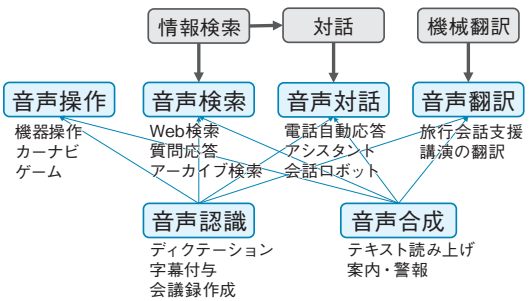

図 2.82　音声技術のアプリケーション

＊1　本会が 1987 年に「電子情報通信学会」と名称変更されたのも同じ頃である．

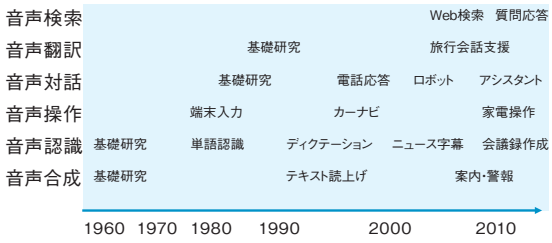

図 2.83　音声応用システムの歴史

は音声認識誤りを考慮した様々な研究開発が行われている．対話に関する研究開発は近年，音声対話が中心になっている．音声翻訳の研究開発も主に音声研究者の主導で行われている．音声認識や合成などの入出力の処理が，全体の仕様や性能において極めて重要であることを示唆している．なお，図には示していないが，要約技術と組み合わせた音声要約に関する研究も行われている．

次に，これらの応用システムの歴史を**図 2.83** に示す．前述のとおり，1980 年以前は音声認識・合成の基礎的な研究が中心で，1980 年頃からそれらが少しずつ実用化され，他の応用システムの研究開発につながっていった．しかし，本格的なシステムの実現は，連続音声認識が実用化された 1990 年代後半以降であり，音声検索や音声翻訳を含む様々な応用システムが実用化されたのは 2000 年代後半以降（すなわち 10 年以内）のことである．この技術の進展は，データベースの大規模化，そして，計算機性能や通信ネットワークの進歩に負うところも大きい．

以下，これらの処理と応用システムについて説明する．

4.3.1　音声認識応用

音声入力タイプ・ワープロは，音声認識の長年の目標であった．世界的にも初期のものに 1960 年に京大と NEC の共同で開発された「音声タイプ」[1]がある．これは単音節単位の認識であった．その後，単語単位の入力のシステムを経て，本格的な不特定話者大語彙連続音声認識が実現されたのは，HMM に基づく技術が確立された 1990 年代である．1997 年にはパソコンのディクテーションソフトとして Dragon systems の Naturally-Speaking と IBM の ViaVoice が発売された．これらは日本語版も発売されているが[2]，我が国でも NEC や東芝などによって相次いで開発されている．なお，フリーソフトの Julius の最初の版が開発されたのも 1997 年である[3]．

しかし，ディクテーションソフトは，医療や法廷などの一部の特殊用途や視覚障害者を除いて，あまり普及しなかった．その理由として，音声入力の方がかえって疲れる，周囲に迷惑をかける，内容を周囲に聞かれたくない，といったことが考えられる．

これに対して，会議や講演のように元来音声で話されているものをテキスト化する応用も考えられる．ただし，これらは自然な話し言葉であるのでかなり難しく，基礎研究が開始されたのが 1990 年代である．米国 DARPA プロジェクトでも，放送ニュースと電話会話が主要な対象とされた．我が国でも，NHK 放送技術研究所が，2000 年 3 月に世界に先駆けて，ニュース番組の字幕付与を実用化した[4]．ただし，直接音声認識するのはアナウンサの発話のみで，ほかは復唱方式を採用している．音声認識は，2005 年頃から地方議会の議事録作成においても導入されていたが，2011 年には衆議院の会議録作成システムで全面的に導入された．これは京大と NTT により開発されたもので，国会レベルで議員や閣僚の発言を直接音声認識するシステムは世界初である[5]．

4.3.2　音声合成応用

世界で初の本格的なテキスト音声合成システムは，1979 年に米国 MIT で開発された MITalk で，その後 Klattalk，DECTalk に発展している．我が国でも電電公社（現在の NTT）を中心に様々な研究開発が行われたが，その後 ATR で開発された Chatr をはじめとする素片選択方式が主流となった．1990 年代後半にはパソコンソフトとして実用化された．ただし，テキスト読み上げとしては視覚障害者以外ではあまり使われていない．2000 年代になって，自然性や明瞭性がかなり改善され，公共交通機関でのアナウンスや防災無線などで導入されている．

4.3.3　音　声　操　作

キーボードなどの入力装置が使えない状況で機器を操作するのに音声は適している．1981 年に電電公社が開発した世界初の音声認識を利用した銀行自動照会システム ANSER も音声操作と捉えられる．カーナビへの音

声認識の導入は，我が国では1990年代から先駆的に行われていた．しかし，計算資源が限られた環境で十分な性能が得られなかった．最近では，カーナビや家電機器がネットワークに接続されるようになり，クラウドサーバ型の音声認識を用いて，次項の音声検索と合わせたシステムが実現されつつある．その代表例が2014年に発売されたAmazon Echoである．

4.3.4 音　声　検　索

音声入力でWebなどを検索する「音声検索」は，大語彙音声認識とWeb検索の単純な結合であるが，画期的な応用である．2008年にGoogleにより米国でサービスが開始され，翌年には日本でも展開された．スマートフォンなどで効率良くWebや地図の検索が行えるので，幅広く使われている．

その後，単純な検索だけでなく，質問応答や音声操作を組み合わせたアシスタントソフトが開発された．その代表例が，2011年に米国でリリースされたAppleのSiriであり，その後Google NowやMicrosoft Contanaなどがリリースされている．前述のAmazon Echoもこの範疇に入る．我が国でも，2012年にSiriの日本語版と同時期にNTTドコモの「しゃべってコンシェル」[6]がリリースされた．これらは，次項の音声対話の振舞いを示しているが，基本的には一問一答の域を出ない．

一方で，大規模な音声アーカイブの検索（spoken term detection）に関しても，2000年代に入って研究開発が行われている．米国のDARPAプロジェクトでは，公安目的を想定して電話会話や（未知の）外国語会話を対象とした研究開発が主流である．我が国では2009年に裁判員制度の導入に伴って，裁判の公判の音声をテキスト化して検索するシステムが導入されている．これはNECの開発によるものである．また，民間のコールセンターでは，顧客との会話が大規模に録音・蓄積されており，その検索やマイニングのニーズも強い．

4.3.5 音　声　対　話[7]

人間と音声で対話するロボットやコンピュータは長らくSFの範疇であったが，音声認識・合成技術の発展を受けて，1990年頃から本格的な音声対話システムが構築されるようになった．その先駆けは，MITのVOYAGERである．これは街の案内を行うもので，現在スマートフォンで行われているサービスに近い．その後米国では，DARPA主導でATISプロジェクトが行われた．

我が国でも1990年代前半に，音声対話システムの研究が活発に行われた．その先駆けは東芝のTOSBURG[8]である．その後，科研費重点領域研究「音声対話」プロジェクトが行われたほか，企業等の研究所も含めて，多くの研究機関で音声対話システムが構築された．ただし，これらは基本的にタスクドメインに特化して，人手で記述した文法と対話フローに基づいて処理を行うもので，パソコン（当時のワークステーション）上でプロトタイプとしては動作したが，実用化には至らなかった．

一方米国では，NuanceやSpeechWorksが電話の音声自動応答（IVR）サービスに特化したシステムを実用化し，2000年頃から多くのコールセンターで導入された．これらにおいては，文法や対話フローなどの記述法がVoiceXMLなどとして規格化された．ただし，これらのシステムは，日本ではそれほど普及していない．その理由として，丁重なサービスに対する要求が高い半面，単純なことは早くから携帯電話のネットサービスで提供されていたことが考えられる．

しかしながら，スマートフォンの登場・普及により状況が一変した．パソコンと同様の複雑なことができるにもかかわらず，キーボードがない状況（音声入力のニーズ）が現れたのである．シーズ面からも，クラウドサーバ型の音声認識により性能が格段に改善した．その代表的なアプリは，前述のSiriや「しゃべってコンシェル」などのアシスタントソフトである．これらのシステムでは，語彙や対象ドメインが非常に大きく，様々な発話を扱う必要があるため，統計的言語モデルに基づいた音声認識と機械学習に基づく言語理解が導入されている．

以上の音声対話システムの進化と淘汰の過程を**図2.84**にまとめる．

一方音声対話は，パソコンやスマートフォンだけでなく，人間型ロボットや仮想エージェントでも必要な機能として求められている．特に人間型ロボットは我が国の得意分野であり，2005年の愛知万博では様々な展示が行われた．ただし，ロボットに装着されたマイクで音声認識を行うのは非常に困難で，限られた会話しか実現さ

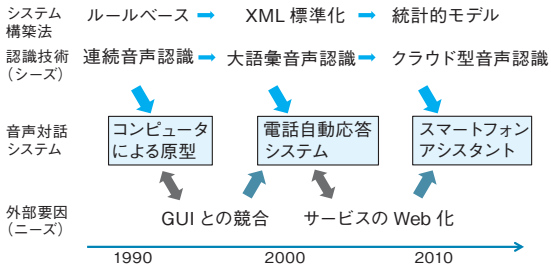

図 2.84　音声対話システムの進化と淘汰

れていない．近年は，クラウドサーバ型の音声認識の導入により改善が図られ，2014年に発表されたPepperのように店頭で接客するロボットも出現している．対話の制御を含めて技術的な課題は多いが，受付や話し相手として活躍するロボットの実現が期待されている．

4.3.6　音　声　翻　訳[9]

音声翻訳は，言語の壁を越える夢の技術である半面，音声認識と機械翻訳という不確実性の高い処理を組み合わせるという点で，非常に挑戦的なものである．我が国は世界に先駆けて，1986年のATR自動翻訳研究所設立以来，継続的に音声翻訳の研究開発を進めている．音声認識・合成と機械翻訳にコーパスベースの手法が導入され，性能が改善されるに伴って，実用的な水準になってきた．対象ドメインも，当初は会議予約やホテル予約などの限られたものから取り組まれ，徐々に旅行会話や多様な日常会話に展開してきた．この成果を基に2007年には，世界に先駆けてATR-Trekが，携帯電話向け音声翻訳サービス「しゃべって翻訳」のサービスを開始した．その後，情報通信研究機構（NICT）が研究を継承し，スマートフォンアプリVoiceTra[10]などのサービスも展開している．このように我が国の研究開発が，外国人とのコミュニケーション支援を主な目標としているのに対して，欧州では議会や会議の同時通訳を対象としたEUのプロジェクトが，米国では外国語の様々なメディアの情報分析を対象としたDARPAプロジェクトが継続的に行われており，各々の優先度が反映されているのが興味深い．我が国では2020年の東京オリンピックに向けて開発や実証実験が加速している一方，世界的にはGoogle TranslateやSkype Translatorなどのサービスも展開されており，今後一層の進歩・普及が期待される．

一方で，2004年から実施されている国際的な評価型ワークショップIWSLTでは，当初は旅行対話を対象としていたが，2010年からはTEDの講演の翻訳に取り組んでいる．このように，講演や会議の同時翻訳も遠くない将来に実用化が期待される．

参　考　文　献

（1）　T. Sakai and S. Doshita, “The Phonetic Typewriter,” Proc. IFIP Congress, pp. 445-450, 1962.
（2）　西村雅史，伊藤伸泰，山崎一孝，“単語を認識単位とした日本語の大語彙連続音声認識，”情処学論，vol. 40, no. 4, pp. 1395-1403, 1999.
（3）　鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，音声認識システム，オーム社，2001.
（4）　安藤彰男，今井　亨，小林彰夫，本間真一，後藤淳，清山信正，三島　剛，小早川健，佐藤庄衛，尾上和穂，世木寛之，今井　篤，松井　淳，中村章，田中英輝，都木　徹，宮坂栄一，磯野春雄，“音声認識を利用した放送用ニュース字幕制作システム，”信学論，vol. J84-D2, no. 6, pp. 877-887, 2001.
（5）　T. Kawahara, “Transcription system using automatic speech recognition for the Japanese Parliament (Diet),” Proc. AAAI/IAAI, 2012.
（6）　辻野孝輔，栄藤　稔，磯田佳徳，飯塚真也，“実サービスにおける音声認識と自然言語インタフェース技術，”人工知能誌，vol. 28, no. 1, pp. 75-81, 2013.
（7）　河原達也，“音声対話システムの進化と淘汰―歴史と最近の技術動向―，”人工知能誌，vol. 28, no. 1, pp. 45-51, 2013.
（8）　竹林洋一，坪井宏之，貞本洋一，橋本秀樹，新地秀昭，“不特定ユーザを対象とした音声対話システムの試作，”人工知能研資，SIG-SLUD-9201-4, 1992.
（9）　中村　哲，“音声翻訳技術概観，”信学誌，vol. 98, no. 8, pp. 702-709, 2015.
（10）　松田繁樹，林　輝昭，葦苅　豊，志賀芳則，柏岡秀紀，安田圭志，大熊英男，内山将夫，隅田英一郎，河井　恒，中村　哲，“多言語音声翻訳システム“VoiceTra”の構築と実運用による大規模実証実験，”信学論，vol. J96-D, no. 10, pp. 2549-2561, 2013.