17 スーパーヒューマン音声対話コミュニケーションシステム

俵　直弘 | 日本電信電話（株）　塩田さやか | 首都大学東京
音声言語情報処理研究会（SLP）

人間は話し好きな動物である．彼は食べ物とほぼ同じくらいニュースや情報，エンタテインメントを要求する．そして，驚くべきことに，感覚喪失実験が示すように彼は情報よりも食物（ときには水でさえも！）がないときの方が遥かに長く生き残ることができる．

—アーサー・C・クラーク

電話 100 周年を祝う通信に関する講演 [1] *より*

コンピュータインタフェースとしての音声対話コミュニケーション

他者とのコミュニケーション，とりわけ音声を介した音声コミュニケーションを行いたいという欲求は，人間の知能そのものに深く根付いた根源的な欲求の 1 つである．多くの人は意識することなく音声によるコミュニケーションを日常的に行っているが，音声を理解し，音声を生成する枠組みをコンピュータで実現することは知能そのものへの挑戦となる困難な挑戦であった．

音声コミュニケーションデバイスが描かれた有名な作品として，冒頭で引用した 2001 年宇宙の旅 [2]（1968）がある．劇中において，木星探査船ディスカバリー号の David Bowman 船長は，宇宙船に搭載された人工知能コンピュータ HAL 9000 とともに木星への調査へ向かう．作中において，HAL 9000 とのやりとりはすべて自然言語を用いた完全な音声対話インタフェースにより行われている．現実世界の 2001 年では HAL 型コンピュータは間に合わなかったが，近年ではスマートフォンやスマートスピーカ，ロボットなど音声入力に適した機械の普及もあり，HAL 9000 のような音声対話システムの実現が現実味を帯びつつある．特に，音声対話システムを構成する個々の技術を見ても，深層学習を始めとした多彩な技術を導入することで，いくつかの分野において人間の能力に匹敵する能力を獲得しつつある．さらに，急速な技術の進展に伴い，分野によっては人間を超える（＝スーパーヒューマンな）機能の実現も夢ではなくなってきた．そこで本稿では，コンピュータの耳と口である音声認識と音声合成に焦点をあてて，これまでの技術を概観し，そして未来の音声コミュニケーションデバイスの実現に向けて，各分野における今後の発展について考える．

音声認識の歴史と課題

音声認識の目的は，システムに入力された音声信号から，その発話内容を文字列として出力することである．その歴史は古く，1950 年代には米国ベル研究所にて，孤立数字の音声認識に関する研究が行われている．最初期の音声認識では，単音節や 1 単語単位で音声を認識する，いわゆる孤立単語認識が対象であったが，人の認知プロセスと同様に，文法に基づき次の単語を予測する言語モデルと，音声波形から現在の単語を予測する音響モデルとを組み合わせることで，複数の単語からなる連続音声認識認識が可能となった．その過程で動的計画法や隠れマルコフモデルといったさまざまな技術が提案され，着実に音声認識性能は高められてきた．そして近年の深層学習技術の発展に伴い，まだ議論の余地はあるが，限られた条件下であれば「人間と同等」の音声認識性能が実現されつつあるとされ

る．しかし，これはあらゆる環境に適用可能な音声認識技術がすでに完成したという意味ではない．たとえば，身近で試すことができる音声認識として，YouTubeで提供されている音声認識機能を見てみると，未だ多くの認識誤りが発生することが分かる．特に，ニュースキャスターの音声のように，雑音が比較的少ない環境下で，正しい文法で明瞭に発話された音声に対しては，ほぼ完璧に近い音声認識結果が得られる一方で，普段我々が過ごしているような，雑音環境下での日常会話に対しては，かなりの頻度で認識誤りが発生してしまう．実際，雑音環境下声認識の代表的なコンペティションであるCHiMEチャレンジ[3]の結果を見ても，2019年現在において最も高性能なシステムを用いても，単語誤り率が50%前後であり，完全な音声認識技術の完成にはまだ道半ばであといえる．このような極端に難しい環境下における音声認識を実現するためには，音声に含まれる言語情報を認識するだけでなく，伝達過程において失われてしまっていたり，そもそも存在しない情報を高度に推論する枠組みが新たに必要であると考えられる．さらに近年では，音声想起時の脳波信号から，発声前の単語認識が行える可能性が示唆されており，音声にすら用いない認識法の発展が期待されている．30年後にはこのような人間でも実現不可能なスーパーヒューマンな音声認識技術が開発されていることを期待したい．

音声合成の歴史と課題

音声合成とはコンピュータの声を作る技術である．これまで，さまざまな技術が提案され，バスや駅の案内など身近な部分での導入が進んできている．しかし，これまでの合成音声は聞けば韻律の不自然さなどからすぐに気が付かれる程度の品質であった．ところが，2016年に登場したWaveNetと呼ばれる深層学習を用いた音声合成法により合成音声の品質は飛躍的に向上し，人間が聞いても人間の発声か合成音声かの判別が難しい高品質な音声生成が可能となりつつある．もちろん音声合成技術が完成したというわけではないが，合成音声の可能性が飛躍的に向上したと言っても過言ではない．一方，音声合成の一端を担う技術に歌声合成がある．歌声合成といえばVOCALOIDで一躍有名になった技術であるが，人間が可能な歌い方を超えるような楽曲も公開されており，合成技術によって生成された音声や歌声が人間を超える未来がやってくることを示しているといえる．現在の技術では，限られた環境や条件においてのみ高品質な合成音声が生成可能となっているが，今後はデータ量やリアルタイム性，個人性，感情など実環境での使用や多様性に関連した課題をクリアしていくと期待できる．その上で，人を超える合成音声は今後，福祉やエンタテインメントなど活躍の幅を広げ日常社会により溶け込んでいく，そんな30年後を期待したい．

30年後に寄せて

昨今の加速度的に発展し続ける科学技術に対して，未来の技術，ましてや30年も先の技術を予想することは，まったくもって不可能であろう．本稿で言及した内容はあくまでも現在存在する技術の延長としての予想であり，今からでは想像もつかない新たな技術や概念が登場しているかもしれない．そのときのためにも，ここはHAL 9000の言葉を借りて拙文を締めたい．

申し訳ありません，デイブ．ご期待に添えなくて残念です．

参考文献

1) Arthur, C. C. : Communications in The Second Century of the Telephone, In Paleotronic Magazine (1997).
2) Arthur, C. C. : 2001 : A Space Odyssey, Hutchinson (1968).
3) Barkeretal, J. : The Fifth 'CHiME' Speech Separation and Recognition Challenge : Dataset, Task and Baselines, In Proc. Interspeech, pp.1561-1565 (2018).

(2020年1月17日受付)

■俵　直弘（正会員）　naohiro.tawara.ex@hco.ntt.co.jp

2016年早稲田大学基幹理工学研究科情報通信学専攻博士課程満期修了退学．2017年同大学にて工学博士取得．その後同大にて助教，講師を経て2019年より日本電信電話（株）コミュニケーション科学基礎研究所．日本音響学会，IEEE，ISCA各会員．

■塩田さやか（正会員）　sayaka@tmu.ac.jp

2012年名古屋工業大学創生シミュレーション工学研究科博士課程修了．同大学にて特任研究員，統計数理研究所での特任助教を経て2014年より首都大学東京システムデザイン学部助教．日本音響学会，電子情報通信学会，IEEE，APSIPA，ISCA各会員．