Vol.2014-SLP-100 No.7
2014/2/1

テーマセッション：音声合成は今後こうなる！

徳田 恵一[†1]　峯松 信明[†2]　戸田 智基[†3]
額賀 信尾[†4]　平井 啓之[†5]

情報処理研究会 音声言語情報処理研究会（SIG-SLP）第 100 回記念シンポジウムにおいて，音声合成研究の流れを俯瞰し，今後の目標・応用や方法論を探ることを目的としたテーマセッションを実施する．本稿は，そこでの発表内容の概要を，登壇者がそれぞれ執筆したものである．

1. 音声合成技術の現状と未来（徳田）

1.1 最近の技術動向

テキストから音声を生成するテキスト音声合成の研究開発に関しては，数十年に渡る長い歴史があり，様々な手法が時代ごとに提案されてきたが，ここ 10 年ほどは，データに基づいてシステムを構築するコーパスベースと呼ばれる方式の台頭が著しい．合成音声の品質は，条件によっては自然音声と区別がつかないレベルに達している．また，Blizzard Challenge による大規模な評価試験において，自然音声と同等の明瞭度を達成したシステムもある．以下，コーパスベースの代表的な手法についてまとめる．

【単位選択型音声合成】大量の音声データを収録しておき，ランタイムにターゲットコストおよび接続コストと呼ばれるコストの総和が最小となるように音素等の音声単位を音声データから選び出し，接続する方式である．スタイルの安定した読み上げ音声に関しては，丁寧に設計・実装された単位選択方式により高品質な合成音声が生成可能であることがわかっている．

【統計的パラメトリック音声合成】同じくデータに基づいた手法であるが，音声波形を分析することにより得られたメルケプストラム，LSP 等の音声特徴列を隠れマルコフモデル(hidden Markov model: HMM) 等の統計モデルによりモデル化し，システムには統計モデルのパラメータのみを保持するものである．統計モデルに HMM を用いる場合には，HMM 音声合成と呼ばれる．音声認識で用いられる話者適応等の手法を拡張することにより，多様な話者性，感情表現等をもった音声を容易に生成できる利点がある．ボコーダに基づいた方式であるため，音声品質には限界があるとされていたが，STRAIGHT 他，様々な改良手法により，その品質は格段に向上している．

また，HMM 音声合成システムをベースに波形の素片を接続するハイブリッド型と呼ばれる方式により，単位選択型音声合成に相当する，もしくは上回る高品位の音声を合成可能であることがわかっている．

1.2 実社会での利用

一般の人々が合成音声を利用したサービスに触れる機会も増えてきている．カーナビ，電話音声案内，音声ポータル，ボイスサーチ，音声翻訳，PC によるウェブページ読み上げ，鉄道・バス等の構内・車内案内，オーディオブック，テレビゲーム・携帯ゲーム，通信カラオケ，ウェブ上の試験サービスやキャンペーンなどが例として挙げられる．Siri，しゃべってコンシェル等のスマートフォン上の音声対話インタフェースや VOCALOID 等の歌声合成が評判となったこともあり，音声合成技術に対する一般の人々の認知度も格段に向上しつつあるが，多くの人々に日常的に利用されるほど広く普及しているとは言えない段階にある．

また，現在，実用化されているシステムの多くは，読み上げスタイルのみを対象としたものがほとんどである．研究レベルでは，多様な話者性を実現する手法，言語を越えて話者性を再現する手法，感情表現を可能とする手法などが実現されているが，実際のサービスで利用されている例は未だ少ない．

1.3 今後の技術開発

今後，計算機資源の質・量の拡大を背景に，音声合成における肉声感の向上，多言語化，多様な話者性や発話スタイル，感情表現を実現する表現能力の向上などを目指して，統計的なアプローチを基盤とした様々な手法が益々深化していくものと予想される．また，これまでになく大規模なデータを取り扱うことの可能な技術的基盤を確立することも課題となろう．以下，著者が重要視する近い将来の技術開発課題について列挙する．

【波形レベルの統計モデル】現在，統計ベースの手法では，何らかの音声分析合成系（ボコーダ）を利用しており，統計的にモデル化されているのはメルケプストラム等の音声特徴である．今後は，音声特徴抽出，F0 抽出，励振源モデル等を統計モデルに統合した音声波形レベルのモデル化手法が確立されていくものと思われる．また，その過程でハイブリッド方式との関係も更に整理されてくるものと思われる．

【テキスト解析部との統合】現在，テキスト解析部と HMM 等の音響モデル部とボコーダ部はそれぞれ独立に構成されているが，上記のように音響モデル部とボコーダ部だけでなく，形態素解析，構文解析，テキスト正規化等からなる

†1 名古屋工業大学
†2 東京大学
†3 奈良先端科学技術大学院大学
†4 日立製作所
†5 エーアイ

テキスト解析部も統合したモデル化手法が発展していくと思われる．更に，段落等の長い範囲のテキスト情報や，対話制御部からの情報を利用あるいは統合したモデル化手法も重要性を増していくと思われる．

【ユニバーサル音声モデル】究極的には，あらゆる話者性や発話スタイル，感情表現を自在に実現できるユニバーサルな音声モデルの構築手法の確立が望まれる．SAT，FA-HMM, CAT, 加算モデル等を統合することにより，そのようなモデルを構築することが可能になると思われる．但し，次節で述べるとおり，人間の音声の多様性を十分に表現可能な大量の音声データを組織的に集積することが対となる重要課題である．

【新しいモデル化手法】DNN (Deep Neural network), GPR (Gaussian Process Regression) 等の新しい機械学習の手法を適用した方式が提案されている．このようなトライアルの中から次世代方式が現れる可能性がある．

1.4 超大規模音声データ

話者性，発話スタイル，感情表現等を自在に実現できるユニバーサルな音声モデルの構築のためには，それらを分離してモデル化できる手法の確立が必須だが，その一方で，人間の音声の多様性を十分に表現可能な大量の音声データを組織的に集積することがもうひとつの重要な課題となってくる．このため，音声データを常に収集・蓄積し続け，大きくなり続けるデータを永続的に維持・共有できる社会的基盤を確立する必要があると思われる．

医療応用分野では，ボイスバンクプロジェクトとしてそのような活動が始まっている．その他，エンターテイメント分野，商業分野等で音声合成システムの構築が行われてきているが，音声データは個別に収録され，そのデータベースの管理・維持も個別に行われてきた．各分野において適切にデータ収録･収集･共有のインセンティブを設定し，わかりやすい共通のライセンス形態を定義することで，音声データを集積する社会的な仕組みを形成できるのではと考えている．また，ポッドキャストやオーディオブック，あるいは様々な動画等，インターネット上にほとんど無尽蔵にある音声データを活用することも重要な課題である．

1.5 今後の音声合成

今後，合成音声の品質，自由度が向上するにつれ，様々な形で音声合成が活用される局面が増え，多くの人々の日々の暮らしの中に溶け込んでいくものと思われる．特に普及が予測される眼鏡型，腕時計型等のウェアラブルデバイスにおいては，音声インタフェースの役割がますます重要となろう．人間のように生き生きとした合成音声が人々の日常の彩りとなることを期待したい．

参考文献

1) K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, K. Oura, "Speech synthesis based on hidden Markov models," Proceedings of the IEEE, Vol. 101, No. 5, pp. 1234-1252, May 2013.

2. CALL における音声合成の今後（峯松）

音声認識・合成といった要素技術の高精度化に伴い，それらを用いたアプリケーションの品質も向上してきた。様々な音声技術を外国語の教育・学習，特に音声教育・学習に導入し，学習者の独習を支援し，教師のタスクを軽減する試みが古くから行なわれている[1]。音声合成技術の CALL での利用を考える場合，[2]では三種類のケースに分類している。1)電子化辞書など，初めてその単語と遭遇する学習者に呈示する音声（モデル音声）を合成音声に置き換える，2)ある教科書（パラグラフ）を母語話者に音声化してもらう代わりに合成音声を利用する，3)インタラクティブな対話型教材で，会話エージェントの音声としての合成音声利用，である。前者の方ほどより高自然性の音声が必要となる。近年の音声合成の品質向上に伴い，2)の利用例が広く見られるようになった。例えば，シャドーイング（外国語音声を聴取すると共に同時に追唱する。認知タスクの高い訓練法）時の母語話者音声資料としての利用や，英語での口頭発表原稿を音声合成で音声化し，それを聞きながら練習する，などの利用法である。本節では現在の音声合成品質が語学教師にとってどの程度のものであるのか，について知るところを述べ，今後の音声合成技術の CALL 利用について，私見を述べる。

母語話者がその言語の合成音声を聞くと，「時々変なところがある。どこが変と聞かれても答え難いが・・」というのが正直な感想だと思う。語学教師の多くは，音声学は「かじった」程度の知識しか持ち合わせおらず，母語話者教師でも類似した反応を示す。一方，非母語話者教師，更には，学習者に聴取させると「え，私より上手い・・」というのが正直な感想であろう。10 年前だと非母語話者教師であってもその利用に躊躇したのであろうが，現在の合成音声の品質は「実際に教えている学習者よりも上手に読み上げる」，即ち（超？）上級学習者相当の読み上げ能力を有する機械と受け止められており，これが，ケース 2)の利用例が近年増えてきた理由の一つであると考えている。以下，今後の合成技術の CALL 利用について私見を述べる。

淡々と読みあげるモノローグ形式の音声が求められる場では，音声合成の利用は一段と広がると予測する。「英語のオーラル発表を，音声合成器で行なう」という文化が定着するかどうかは分らないが，そういう発表が出てきても不思議ではない。既に PowerPoint の plug-in は存在する。かつて論文を手で書いていた時代では，最終的に「字の奇麗な」秘書さんに清書をさせていた。それを機械に置き換えた（？）のがワープロであり，それの音声版が英語上級者による代理発表，そして，合成音声技術，である。

機械に読み上げさせることに躊躇する場合でも，読み上げを支援してもらいたいと考える学習者は多い。テキストを読み上げる場合，母語話者はテキストに明示されない事

項を無意識的に処理しつつ音声化している。この処理が学習者にとって難しい。音声合成の場合，前処理として実装されている各種音韻情報，韻律情報の推定がこれに相当するが，これを学習者に視覚的に呈示し，読み上げを支援するインフラが一般化すると思われる。日本語学習に関しては[3]がその一例であるが，世界中の教師・学習者に歓迎され使われており，開発者として驚いている。

現在の音声合成は，モノローグ形式は得意だが，語学教育ではダイアローグ形式の音声が求められることも多い。昨今の語学教育はコミュニケーションを重視しており，初級者の教科書でも会話を題材とすることが多い。その結果，適切な焦点制御，イントネーション制御などが自ずと教育項目に入ってくるが，現在の音声合成技術で，これらを全自動かつ高品質で行なうことは難しい。一方（受験の為ではなく）生活のために外国語を学ぶ場合，例えば日本で日本語を学ぶ留学生などは，目上の人に対して行なう丁寧な発声方法を学ぶ。例えばバイト先の上司に「話し方が生意気だ」と嫌われることが現実に起きている。話す方は意図しなくても，聞き手に「ぞんざい，無礼な喋り方」と受け止められた例である。客相手にこのような喋り方をされてはたまらない。読み上げ調から会話調へと，合成音声の技術開発は進んでいるが，その進展が生きる応用場面はCALL には常に存在する。この場合，完全に母語話者のような発声が出来なくても，上級学習者程度での制御が実装できれば,それを使い始める教師は少なくないと思われる。しかしその一方で，音声合成技術が，初級学習者でもしない誤りを時として犯すことも事実である。応用分野に応じて許容できる誤りと，起きて欲しくない誤りがある。実用システム構築にはこのような配慮も必要である。

以上は，上級学習者，更には母語話者に匹敵するくらいの発話機械としての音声合成技術の利用であるが，学習言語を国際語である英語とした場合，上記とは少し違った応用場面が見えてくる。英語の場合，学習者の母国語に起因する訛りを積極的に認める動きが教育界にある。世界諸英語[4]と呼ばれ，英語，米語も英国訛り，米国訛りとして捉える。この場合訛りを identity として捉え，尊重することになる。しかし，この訛りに起因するコミュニケーションエラーも否めず，聞き取り易い訛りとそうでない訛りがあるのは事実であり，かつそれは聞き手に大きく依存する。世界中の英語利用者が「自分の」英語を話す場合，例えば，二者がネットを使って会話する場合に，相手の英語を聞き取り易い自分の訛りに変換して聴取するインフラや，海外のホテルやレストランで自分が聞き取りやすい英語を話す従業員をすぐ見つけられるインフラなどは，国際社会での英語の使われ方／話され方を考えると，実用価値のあるインフラであると思われる。このようなインフラ作りのためには，各自がどのような英語を話すのかを登録し，管理する枠組みが必要である。欧米では CEFR(Common European Framework of Reference for Languages)[5]と呼ばれる言語パスポートが普及しつつあり，各自がどの言語をどのくらい操れるのかを示す情報源として社会的に機能している（例えば就職面接の際に参照される）。このような枠組みに発音を登録し，多様な英語が使われるグローバルコミュニケーションを支援するインフラが求められるだろう。

なお，音声言語を対象にした自動翻訳技術が人間レベルの水準に達成すれば，そもそも人間が外国語を学習する必要性は無くなるのか？と聞かれることがある。これは，ブレイン・マシン・インタフェースが発達すれば，それを装着すれば（言語を明示的に使わなくても）あるイメージを想起しただけで，機械が適切に（音声）言語化して対応してくれるから，人間は母語を獲得する必要はないのか？という問いとも少なからず関連する。筆者の私見ではあるが，自動翻訳技術が今度どこまで発達しても，何某の自然言語を国際共通語として採択し，少なくともそれは利用できるよう子供を教育する慣習は，今後も継続されるだろう。

参考文献

1) 河原他，“音声情報処理技術を用いた外国語学習支援”，電子情報通信学会，J96-D，7，1549-1565，2013
2) Z. Handley et al., “Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning,” Language Learning & Technology, 9, 3, 99–120, 2005
3) 峯松他，“日本語アクセント・イントネーションの教育・学習を支援するオンラインインフラストラクチャの構築とその評価”，電子情報通信学会，J96-D，10，2469-2508，2013
4) B. Kachru, et al. *The handbook of World Englishes*, Wiley-Blackwell, 2009.
5) http://www.coe.int/t/dg4/linguistic/CADRE1_EN.asp

3. 音声変換技術の今後（戸田）

音声合成技術の一つに，入力された音声に基づいて所望の音声を合成する音声変換技術がある. 入力音声に対して，言語情報を保持しつつ，所望の非言語情報やパラ言語情報を持つように変換処理を行う技術である．代表的なものとして，音声モーフィングや声質変換と呼ばれる技術が挙げられるが，それら以外にも様々な技術が考えられる[1]．音声分析合成技術および統計的手法による変換・合成技術の発展により，その性能は日々着々と改善している．それに伴い，1990 年頃から研究されている話者変換のみでなく，異なるモダリティ間の変換や劣化音声に対する変換など，より高度な変換処理を必要とする応用技術も研究されるようになってきている．

個人的な話で恐縮であるが，2014 年 2 月現在の時点で 37 歳であり，今後約 30 年間は現役研究者生活を送ることができる（と信じている）．今後 30 年で音声変換技術をどう発展させていけるかについて，私見（抱負）を述べる．

3.1 音声生成機能拡張のための音声変換技術

テキスト音声合成に代表される他の音声合成技術と比較して，音声変換技術だからこそ実現可能となる応用例の

一つは,人対人の音声コミュニケーションへの適用である．言語情報を必要としない音声変換技術では，入力された音声を瞬時に処理することが可能である．リアルタイム音声変換処理を用いることで，物理的・身体的制約を超えて，所望の声質を持つ音声の発声が可能となり，音声生成機能を拡張することができる．例えば，喉頭摘出者などの発声障害者に対してより自然な音声での発声を可能とする代替発声装置や,体内伝導音声を用いた秘匿性の高い通話機器，所望の声質での発声や歌唱を可能とするボイス／ボーカルエフェクターなど，我々の生活をより豊かにする様々なシステムを構築できる可能性を秘めている．

人と人の音声コミュニケーションにおいて有効な音声変換技術を実用化するためには，解決しなければならない問題は山積みである．高度な変換処理を行う上で，統計的手法に基づく音声変換処理は極めて有効であるが，学習時に用いる音声データと使用時における音声データの不一致に対しては虚弱である．日常生活で音声変換を使用する上で，外部雑音の混入や発話様式の変動は不可避である．また，発声障害者補助応用においては，例えば術後の時間経過に伴う調音の変化なども生じる．これらは容易に上記の不一致を引き起こすため，音響特性を自動的に補正する技術の実現が必要となる．一方で，変換技術の使用を通して使用者がシステムに対して最適な発声法を習得していく可能性も考えられる．システムから人への歩み寄りと人からシステムへの歩み寄りの両方を考慮して，これらを相補的に活用できる数理的な枠組みの実現に取り組んでいきたい．

実環境で使用する上で，様々な要因により変換音声の品質劣化が生じると予想されるが,その際に最低限の品質（例えば変換前の入力音声の品質）を確保できる仕組みについても検討する必要がある．要求される品質は応用例により大きく異なると考えられるため，個々の応用例に特化した音声変換技術の構築が必要である．また，日常生活で音声変換技術を使う上で，使用しやすい音声入出力デバイスを設計することも重要である．

発声障害者補助などのように，重要な情報が欠落している音声を入力とするような応用例では，本質的に変換音声の品質は限定される．このような場合，音声に限らず，他のモダリティ信号の情報も入力として併用して，より高品質な変換音声を実現する技術が有用であると考える.また，音声生成過程における物理的制約を統計的変換処理に組み込むことで，発声器官動作パラメータ操作に基づく音声変換機能も実現できると考えている．最終的には，パラ言語情報も制御できる発声補助装置の構築を目指していきたい．

3.2 創作支援のための音声変換技術

音声合成技術の発展に伴い，専用ソフトウェアを用いて，所望の話者・歌手による音声および歌声を創作する活動が活発になっている．このような創作活動において，ユーザがイメージする音声・歌声を如何に容易に合成できるようにするかが重要になると考えられる．音声変換技術の併用により，自身の音声を用いて合成音声・歌声を制御する技術が発展していくと予想される．その場合，任意のユーザが容易に使用できる音声変換技術の構築[2]も重要である．さらに，特定の話者や歌手による音声・歌声の合成のみでなく，所望の特徴を持つ音声・歌声の合成技術の需要が高まる可能性が考えられる．そのため，直感的に音声・歌声をデザインするための技術の構築に取り組んでいきたい．

音声創作支援を行う音声変換・合成技術は，これまでに存在しなかった発声や歌唱を可能とする．これにより，ユーザの創作意欲はかきたてられ，新たな特徴を持った音声・歌唱データが生まれる可能性がある．統計的手法は大量のデータを扱うのは得意であるが，未知のデータを生み出すのは苦手である．近年急激に広まっている二次創作活動は，この問題をユーザが解決してくれる可能性を示している．人によるデータ創作と，そのデータを用いた統計的手法の発展および創作支援技術の改善を上手く連携させることで，正のスパイラルを生み出す枠組みを構築していきたい．

3.3 音声変換技術に対する社会的認知の確立

統計的手法による音声変換・合成技術において，音声データは極めて重要である．発声障害者が術前の音声を保存しておけば，失われた音声による発声を取り戻せる可能性がある．仮に自身の音声データが残っていない場合でも，大規模な音声データベースがあれば，類似した音声や好みの音声を見つけたり創り出したりすることができる．そのため，ボイスバンク[3]のように，多数の人の音声を登録するシステムは極めて重要であり，今後さらに発展させていく必要がある．

音声合成・変換技術により，我々の生活はより豊かになると信じている．その一方で，なりすまし等，悪用の可能性が増すことから目を背けてはいけない．音声合成・変換技術が社会的に認知され,その利点と欠点を直視した上で,正しく使用される社会を築いていく必要がある．

参考文献

1) 戸田，“統計的手法に基づくリアルタイム声質変換による音声生成機能拡張”，日本音響学会講演論文集，2-8-2, pp. 59—62, Sep. 2013.
2) 戸田，“統計的手法に基づく声質分析・変換・制御技術とその応用”，日本音響学会講演論文集，1-8-11, pp. 257—260, Sep. 2011.
3) J. Yamagishi *et al*., “Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction,” Acoustical Science & Technology, 33(1), pp. 1—5, Jan. 2012.

4. 日立の音声合成の現状と今後について（額賀）

4.1 はじめに

日立での音声合成研究は，約 40 年以上が経過した。90 年代後半から福祉分野をはじめとして情報機器の音声出力イ

ンターフェースとして採用され，近年では，カーナビや各種案内放送などでの利用が進んでいる。2000 年代中頃には，CPU リソースやメモリ量の制約から，組込み機器向けのコンパクトな実装が必要となり，搭載素片数や処理量削減手法の検討を行った[1]。近年では，機械学習技術によって最適なデータを自動学習することが可能となり，従来の人手による発見的な特徴量選択では困難であった日本語の細かな韻律特徴の考慮や話者の話し方の癖なども再現できるようになった[2]。また，外国語を合成するシステムの開発も行っている。

4.2 日立の音声合成システムの特徴

音声合成処理は，テキスト解析，韻律生成，波形合成から構成されている。波形合成では，各音節に対応する音声部品（素片）を素片 DB から取得し，音のつながりの良さ（接続コスト）を考慮して最適な組み合わせを選択，指定された韻律になるように変形・結合することで合成音声を生成する。波形合成では，波形重畳と波形接続の 2 つの手法を併用している。波形重畳では，ピッチに同期して波形を切り出し再配列する信号処理を用いて，素片の韻律を変形させる。素片データベースが小さく，目標韻律に近い素片がない場合にこの手法を用いる。一方，波形接続は，素片に韻律変形を施さず，スムージング処理で結合するだけで合成音声とする。

日立は，この 2 つの手法を動的に切り替えて使い分けるセレクティブ重畳型音声合成技術を開発した。本技術は，目標韻律との差や隣接する素片間の韻律の差をもとに，波形重畳で韻律を（どの程度）変形するか，あるいは変形なしの波形接続とするかを素片ごとに決定する。波形重畳は，目標韻律に一致した音声を合成できる半面，こもり感などの音質劣化が生じる。一方で，波形接続は，目標韻律に合致した素片が見つからない場合に不連続感などの音質劣化につながるという問題があった。本技術の開発で，合成音声の肉声感が大幅に向上した。

韻律生成では，様々な特徴量が付与された大量の数値を学習データとし，ある特徴量リストが与えられたときに最も精度良く数値を予測できる構造（決定木）を自動構築する手法を採用している[2]。韻律付与では，入力テキストに対応する特徴量リストを入力して，音素の継続長や基本周波数を予測する。この機械学習により，従来の人手による発見的な特徴量選択では困難であった日本語の細かな韻律特徴の考慮や話者の話し方の癖なども再現できるようになった。

機械学習の活用には，朗読調や会話調など，発話スタイルを変えて音声データを収録するだけで，新たな発話スタイルへの対応が可能になる利点もある。日立では現在，カーナビやスマートフォンでの音声対話技術をターゲットに，会話調の韻律モデル構築を進めている。

4.3 おわりに

素片選択型の音声合成システムを開発して 10 年以上が経過した。その間のバージョンアップも経て，お陰さまで幅広く利用頂いている。合成音声によるガイダンスも相当浸透したと感じるが，あるべき姿にはまだ程遠いのが実感である。以下は私見であるが，今後の課題として挙げたい。1) 会話調音声のアノテーション方式と制御方式の検討，2) 話者が希少である言語，書き言葉や正書法がない言語の保存及び合成，3) やはり埋まっていない人間の音声との差の解消，4) 自由に韻律・声質を制御することができる音声合成システム。これらは，社会的・経済的・技術的観点での合意形成が必要であると同時に，一層の研究開発が必要と考える。

参考文献

1) Nukaga et. al, “Scalable implementation of unit selection based text-to-speech system for embedded solutions,” Proceedings of ICASSP 2006, I-849, 2006
2) 孫他；統計的モデルを用いた波形接続方式音声合成における分割学習によるモデル構築法，音講論 3-Q-24,（2012.9）

5. エーアイの音声合成の現状と今後について（平井）

5.1 はじめに

近年、音声合成技術は統計的なモデルや機械学習の進歩とともに急速に向上し、音質や自然性が大きく改善された。それに伴い、弊社の音声合成エンジンも様々な分野での利用が広まっている。ここでは、弊社がこれまで開発を進めてきた音声合成エンジンの紹介と現状の課題、普及拡大のための今後の目標について簡単に述べる。

5.2 AITalk の特徴

AITalk とは弊社が開発しているコーパスベースの波形接続型音声合成エンジンである。弊社では、他社との差別化を図るため、音声の高品質化だけでなく、ユーザの希望する声(カスタム音声)を容易に提供できるエンジンの開発に力を入れてきた。音声合成エンジンの日本語への特化もその１つである。１つのエンジンで多言語に対応した場合、韻律や素片選択のモデルの自由度が増えることから、少ない収録音声では合成音声が不安定化すると言った問題が生じる。日本語に特化し、モデルのパラメータに様々な制限を加えることで、収録音声の量に関わらず、ロバストで高品質な音声の合成が可能となっている。また、専門のラベラーを育成し、日本語音声合成のラベリング用に最適化されたツールを日々改良することもカスタム音声の品質の向上に寄与している。その結果、弊社では 200 文程度の収録音声から個人性を再現できる音声合成の提供を行っている。

5.3 現状の課題

著名人の合成器を作成する場合、収録の拘束時間に制限があるため、およそ２～３００文程度しか収録することが

できない場合が多い。このような場合、話者によっては、個人性は再現できるが音質が劣化する場合がある。韻律生成には統計モデルを使用しているため、比較的少ない収録データからでも安定した結果を得ることは可能である。しかし、合成部には波形接続方式を用いているため、音素不足による接続歪が問題となる。この問題を解消する可能性のある方法としては話者適応を用いた HMM 音声合成[1]がある。しかし、声で知られた著名人やアニメのキャラクタなどは極端な特徴を持つ音声が多く、如何なる音声に対しても安定して個人性の再現が可能なパラメータ音声合成システムの構築が課題の 1 つであると考えている。

5.4 今後の目標

音声合成の利用される分野を広げることが重要である。現在、弊社の音声合成が主に利用されているのは、1) 駅の構内放送, J-ALERT など防災等の公共放送, 設備トラブル時の緊急放送等の自動アナウンス、2) 銀行のテレフォンバンキング, 図書館自動電話応答サービス等の IVR、3)カーナビゲーション、4) 「しゃべってコンシェル」[2]等の音声エージェント、5) 動画や資料の説明用のナレーションやアテレコ等のコンテンツ作成、6) WEB の読み上げやオーディオブック作成等の既存のコンテンツの音声化、7) その他、ゲーム, WEB キャンペーン, e-learning 教材等の Windows やスマートフォンのアプリケーションや専用機器への組込み等である。

これらのうち 1)-4)は能動的に聞く意思を持った聴取者に音声を提供するものであり、現在の TTS の性能であっても実用に耐え得るものである。合成文の多様さや緊急性などを考慮するとプロのアナウンサの収録音声と比較しても TTS を使うメリットはある。それに対し、5)6)では単純な情報だけでなく、コンテンツの意図を分り易く正確に伝える必要があり、音声の綺麗さだけでなく、イントネーションや速度、間の取り方などの話し方が重要となる。現状では、大量に処理を行う障害者向けのコンテンツの音声化など特殊な場合を除き、TTS がプロのアナウンサに優っているとは言い難い。よって、これからの合成音声には多様性の向上とともに、それらを簡便かつ適切にコントロールする技術が必要であろう。そのためには、魅力のある話し方の分析、自然性が高く多様な音声を簡単に生成できる音声合成技術、そして、それらを適切に制御するための言語処理技術など研究開発すべき事柄は多い。これらを解決し、音声合成の利用価値を高め、より広く普及させることが今後の目標である。

5.5 おわりに

音声合成技術の進歩は目覚ましく、文字を綺麗な音に変換し読み上げることに関しては、高品質なものに近付いてきている。将来的には、音の綺麗さだけでなく気持ちを伝えることのできる魅力的な言葉を生成できる音声合成を目指したい。

参考文献

1) 田村他, 信学論(D-2), vol.J85-D-II No.4 pp545-553, 2002.
2) https://www.nttdocomo.co.jp/service/information/shabette_concier/