コーパスベース音声合成とその応用

渡辺　聡　　岩木　健
兼安　勉　　三木　敬

沖電気では来るべき社会の姿を「e社会®*1)」として描き，全ての人に「欲しい情報を望む形で」提供できることが重要だと考えている。また，単なる情報の伝達を超えた，感情や共感を伝える豊かなコミュニケーション環境を提供したいと考えている。

一方，音声は最も基本的なコミュニケーション手段の一つである。より豊かな音声コミュニケーション環境を提供すべく，音声合成の高表現化に取り組んでいる。

音声合成技術

音声合成に関する技術開発は古くから行われてきた。音声合成技術を大別すると，録音編集方式と規則合成方式がある（**表1**）。録音編集方式は，人間による発声（自然音声）そのものを，単語やフレーズといった再利用可能な単位でデータベースに蓄積し，これらを適切に連結することで合成音を生成する技術である。自然性の高い合成音が得られる反面，その仕組み上，合成できる音声メッセージが限定されるという制約があった。また新たに単語や言い回しを追加する場合，同じ発話者が同じ発話スタイルで音声の収録を重ねなくてはならないという制約があった。録音編集方式は，あらかじめ合成する音声メッセージを特定できるシステムにおいて，広く実用化されている。

規則合成方式は，イントネーションなどを，あらかじめ規定した合成ルールに従って変化させながら，音声波形断片データを編集することで合成音を生成する技術である。任意のテキストを合成対象とするため，新たに単語や言い回しを拡充する場合にも音声収録は不要である。従来，この合成ルールや音声波形断片データの作成は，専門家が音声波形を分析し，その知見を用いて決定していた。ここではルールで記述できる表現力の限界，波形断片を作成・編集する際の信号処理による波形劣化が避けられない等の制約があり，結果として，口調のバリエーション（表現力）や人間らしさ（肉声感）に乏しい合成音となっている。規則合成が，当初の期待ほど実用化に至らないのは，上記音質的課題が原因の一つだと考えられる。

表1　音声合成技術の分類

	合成単位	合成対象	合成音質	適用例
録音編集方式	単語、フレーズ等	限定文	自然音声と同等	駅の構内放送 株価案内
規則合成方式（従来）	音素等	任意文	滑らかだが機械的	メール読上げ PC用スクリーンリーダ
規則合成方式（コーパスベース）	音素等	任意文	肉声感が高い	ボイスポータル 擬人化エージェント

*1)e社会は沖電気工業株式会社の登録商標です。

このような中，1990年代頃から，規則合成方式の一手法として，大規模な音声データベース（音声コーパス）を構築し，これに基づく統計的なアプローチでアルゴリズムやデータベースを決定するコーパスベース音声合成（Corpus-based Text-To-Speech：以下CTTSと略）の研究が行われるようになり[1)]，現在主流となってきている[2)]。CTTSは，任意のテキストを合成対象としつつも，録音編集方式に迫る表現力や肉声感を保てる技術として期待されている。当初問題であった計算量や記憶容量の問題も情報機器の進展等によって解決され，実用化の段階に入ってきている。

なおコーパスベース音声合成の確定的定義は明らかでない[3)]が，本稿では「あらかじめ蓄積した大量の音声波形を，音素単位で直接接続することで合成音を得る方式」と定義し稿を進める。

コーパスベース音声合成の特長

(1) CTTS技術の概要

CTTSに関する技術は，音声コーパス構築技術（**図1a**）と音声合成技術（**図1b**）に大別できる。

音声コーパス構築技術は，大量の音声データから音声コーパスを構築し，音声合成時に必要となる韻律モデルと音声データベースを作成する技術である。音声コーパスの構築においては，あらかじめ収集した数十分から数十時間の音声データに対し，音素・韻律・形態素といったさまざまなレベルで，テキストデータを時間的に対応さ

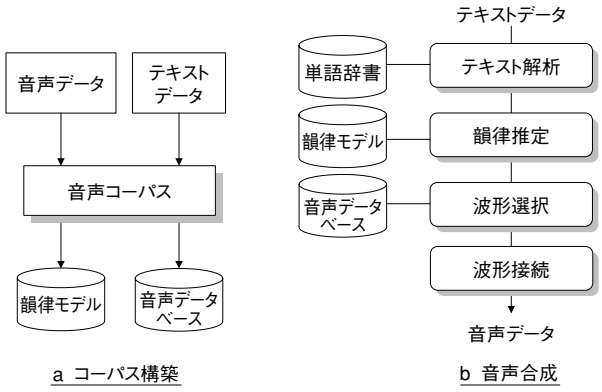

図1　CTTSの処理の概要

せた上で，各パラメータの抽出・統計データの算出を行う。続いて，構築された音声コーパスに対して，統計的なモデル学習等の客観的に定められた手続きを行うことで韻律モデルと音声データベースを作成する。韻律モデルとは，合成音のイントネーション・各音の継続時間長・ポーズの長さ等，合成音の口調的部分を決定するためのモデルであり，音声データベースとは，実際に合成時に直接接続する音声データを検索可能な形で蓄えたものである。

音声合成技術は，実際にテキストデータを音声データに変換する技術である。入力されたテキストの解析結果から，コーパス構築技術で作成した韻律モデルを用いて韻律情報を推定する。続いて推定された韻律情報に最もマッチする音素波形の組を音声データベースから選択し，これらを接続することで合成音を得る。

(2) CTTSの音質的特長

CTTSでは，自然音声の波形を音素単位で直接接続することで合成音を得るため，従来の規則合成方式と比較して，高い肉声感を持つ音声を得られる。従来の規則合成方式による合成音は，その了解性（何を言っているかわかるか？）において自然音声と大差がなかったにもかかわらず，広く受け入れられたとはいえない。これは合成音独特の機械的な音質が大きく影響していると考えられる。高い肉声感を確保できれば，より多くの人に，より多くの利用シーンで受け入れられると言える。

個人性や口調表現可能な音声合成を効率的に実現できることも，CTTSの音質的特長である。従来の規則合成方式では，特定個人や特定口調に関する合成ルールの決定や音声波形断片データの作成に対して，個別に音声技術の専門家による分析が必要であったため，さまざまな個人・口調を表現できる音声合成を開発するのは現実的ではなかった。CTTSでは，戦略的に収集した音声データを用いて音声コーパスを構築することで，特定の個人性・口調を表現可能な音声合成をシステマティックに開発できる。個人性や口調は音声メディアの持つ大きな特徴であり，たとえばサービス全体のイメージに合った音声を使ってサービスを提供することや，微妙なニュアンスの違いを使って感情や意図を伝えることが可能となる。

CTTSの技術開発動向

(1) 自然性の向上

十分な量の音声コーパスを扱えるようになったことで，合成音の自然性は格段に上昇している。読み上げ対象を特定のアプリケーション（たとえば天気予報やチケット予約）に限定することで，自然音声と同等の音質に達するものもある。また，ニュース音声に特化して高い自然性を実現するもある[4]。録音音声としてあらかじめ蓄積した固定の音声メッセージと張り合わせても違和感が少ないため，氏名や地名の合成のみにCTTSを適用する混合型のシステムも出てきている。

しかし，読み上げ対象を任意文に広げた場合の音質は，自然音声と比べてまだ十分ではない。また，波形接続で発生する不連続性を解消するために行うスムージング処理によって，自然性が低下してしまう。このような課題に向けて，任意文でも高い自然性が得られる大規模コーパスの効率的構築[5]，自然性劣化の少ない波形接続方式の開発，音声データベース内の素片に推定韻律を適合させる方式[6] 等の取り組みが進められ，徐々に成果が現れている。

(2) 個人性

個人性の再現については，特定タレントやアナウンサの音声でコーパスを構築し，本人の口調で合成するシステムが開発されている。音声収録に不慣れな一般人の発声に対しても，雑音等のない良好な収録状態の音声データを収集すれば，本人の声であることは十分表現できるレベルに達している。

その一方，高品質な合成音として個人性を再現するためには，各個人に対して，十分な量の高品質音声データを確保する必要となるが，これが必ずしも容易でないという問題がある。この問題に対して，特定個人の少量の音声データから得られる個人性のパラメータを用いて，既存の音声コーパスの合成音を声質変換する，話者変換技術[7] の研究開発が行われている。

(3) 感情表現

従来，大半のCTTSは職業アナウンサによる朗読調の音声でコーパスを構築しているため，いわゆる「読み上げ音声」である。しかし最近では，特定感情（“喜び”，“悲しみ” 等）を持たせて発声した音声から感情ごとにコーパスを構築し，これらを用いて音声合成するシステムも開発されており，コーパスを切り替えて使うことで，ある程度感情を表現することが可能となっている。

一方，さらにきめの細かい感情表現を考えると，必要な感情の種類や程度，感情同士の相互関係をはじめ詳細な分析に基づくコーパスの設計が必要となる。このような問題に対処すべく盛んに研究が行われている。

期待されるアプリケーション

上記課題の解決により表現力・自然性が高いCTTSが実現され，新たなアプリケーションが可能となる。

●音声ガイダンス

公共機器などの使いやすさを向上させるために音声ガイダンス機能の搭載が進んでいる。このようなシステムをより使いやすくするためには，利用者や利用状況にきめ細かく対応できる音声メッセージを低コストで提供することが求められる。現在の音声ガイダンスシステムは，録音編集方式もしくは録音音声そのものが主流である。したがって，音声メッセージの追加・変更を行うには，同一アナウンサによる再収録等のコスト問題が発生する。このような問題に対し，CTTSは音声データの制作コストを格段に低減し，大量な音声メッセージを追加・補充・変更を可能とすることで，使いやすい音声ガイダンスの普及に貢献する。ネットワーク化されたCTTSを利用することで，リアルタイムに音声メッセージの一斉更新を行うことも可能となる（**図2**）。

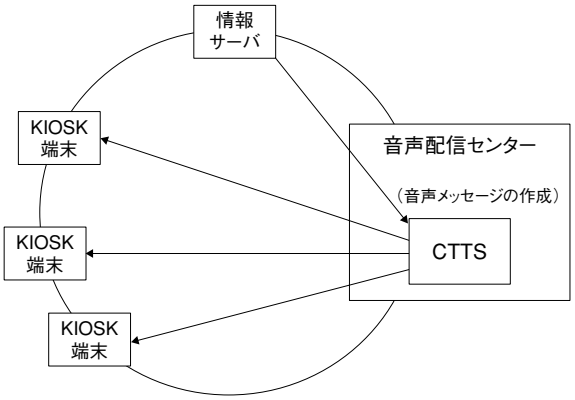

図2　音声メッセージの一斉更新の例

●携帯情報機器のインタフェース

携帯電話やモバイル機器の普及により，家庭やオフィスの外でも電子コンテンツを楽しみたいというニーズが出てきている。このような環境で使われる機器は，十分な表示機能を設けるスペースがないものも多く，音声出力は有力な情報提示手段となる。電子書籍や電子新聞の音声化といった，比較的長時間におよぶ音声コンテンツは，自然性に乏しく口調が単調な従来の規則合成に不向きであるとされていた。CTTSを用いれば，その自然性や高い表現力によって，これらを通勤電車や車の中で楽しむことが可能となる。

●福祉機器への適用

情報技術の進展によって，福祉機器の分野にも新たな製品が生まれつつある。CTTSによって多様な合成音を利用できるようになるため，視覚障害者に対してより親しみやすい音声を用いた電子メールやWebページ読み上げを提供できる。さらに，多くの音声書籍や教育教材を音声化し図書館等に備えることも可能となる。一方，発話障害者に対して，ユーザーの個人性を表現できる有用な発声器を提供することができる。すでに，咽頭摘出者に対する音声合成システムが開発されており[8]，今後数多くの用途に向けたシステムの開発が期待される。

実用化に向けた課題

沖電気では，長年にわたって規則合成方式の研究開発を行い，さまざまな製品を世に出すとともに，そのアプリケーションの普及に取り組んできた[9) 10)]。音声合成を実用化するには，「CTTSの技術開発動向」の節で挙げた技術項目以外にも解決すべき課題がある。CTTSに関わる実用化のための課題を示す。

なお，CTTSではアプリケーションによって求められる音質が大きく異なるため，技術開発はアプリケーションを明確にして進めることが重要である。当社では，サービスプロバイダやコンテンツプロバイダ等と協力してCTTSを使ったサービスの実証実験を推進し，その中で各サービスに要求される音質的特徴・水準等を明らかにしながらCTTSの技術開発を進めている。

(1) 合成音質の設計技術

現在のCTTSでは，「この位の品質の合成音を出すためには，この位の収録音声から構築した音声コーパスが必要」といった見積りが難しい。合成単位のカバー率[11)]，音声データ量[3)] に基づいた一般的なガイドラインは明らかにされつつあるが，ビジネスでの実用化を前提とした場

合にはまだ不十分である。所望の音声コーパスをビジネス上妥当なコストで提供可能かどうかを判断するためには，収録する音声データ量と内容，合成対象とするテキストのバリエーション，合成音に求められる音質，等の関係を定量的に論じられる枠組みをより具体的に作っていく必要がある。

特に，合成音質の評価に関しては，現在は明瞭性・了解性等について主観評価に基づく評価方法が開発されているが[12]，CTTSの高い表現力に対応できる新たな音質評価指標の開発が必要である。

(2) インタラクション設計技術

現行の音声合成を使ったシステムの中には，冗長性，揮発性，速報性といった音声メディアの特性を配慮していないものも少なくなく，結果として音声合成の普及が阻害されている側面がある。メディアの特性やそれを生かす機能を考慮しつつ，システム全体のインタラクションを注意深く設計することで，多くの音声合成アプリケーションにおいて実用性の向上が期待できる。特に，CTTSの高い表現力を十分に引き出すためには，このインタラクション設計が重要となる。

たとえば，CTTSを適用した発話装置では，テキスト入力機能，発声開始・中断機能，感情音声制御機能等を適切なGUI等として実装し，併せて提供することで初めて表現力の高い発話装置が実現できる。同様に，心地よい操作感を提供するマルチメディア情報端末を実現するには，各メディアに対する適切な提示情報の振り分けや時間同期を併せて設計する必要があるし，魅力的な電子小説読み上げサービスを実現するには，テキスト各部に対して最適な感情情報や発話ペースを与える機能等を併せて提供する必要がある。

(3) 権利問題

CTTSが普及すると，音声コーパスに対するさまざまな権利が問題となると予想できる。収録音声を含む音声コーパスの権利，音声コーパスから作成した韻律モデルや音声データベースの権利，それらを組み込んだ音声合成システムの権利，音声合成の結果生成された合成音データ（音声ファイル）の権利等である。積極的に実用化を進めるためには，これらの契約を結ぶためのガイドラインが求められる。

また，CTTSを悪用するケースも考えられる。たとえば，特定人物のCTTSを使ってモラルに反する内容の音声データを作成・配布する場合や，本人詐称（いわゆる振り込み詐欺等）に利用する場合などである。電子透かしや本人照合といった技術的解決策の検討と併せて，これらの使い方に関する社会的合意等が必要となる。

ま と め

CTTSの概要を紹介するとともに，期待されるアプリケーション・実用化に向けた課題を述べた。CTTSの特長である自然性・個人性再現・感情表現は一定レベルで実現されており，適用領域の探索とビジネスに向けての実用化フェーズに入ってきている。

沖電気では，さまざまなプレーヤとの共同実証実験を通して，新たなサービスの課題をいち早く抽出・解決し，CTTSのビジネスレベルでの実用化を推進していく。

◆◆

■参考文献

1）キャンベル 他：“CHATR:自然音声波形接続型任意音声合成システム”，信学技報 SP96-7，1996年5月
2）広瀬：“柔軟な音声合成”，パートナーロボット資料集成（エヌ・ティー・エス社），pp.58–85，2005年12月
3）河井 他：“[チュートリアル講演] 音声合成用大規模音声コーパスの構築”,信学技報 SP2005-9，2005年5月
4）世木 他：“可変長の音素環境依存音素列を単位とする波形接続型音声合成の検討”，音講論集，1-8-9，2003年9月
5）平井 他：“コーパス・ベース音声合成システムXIMERA”，信学技報SP2005-18，2005年5月
6）水野 他：“コーパスベースアプローチによるテキストからの音声合成”，NTTジャーナル，pp.23–26，2004年1月
7）磯貝 他：“多様な音声合成のためのモデル適応・適応学習アルゴリズムの検討”，信学技報SP2005-50，2005年8月
8）木村：“耳鼻咽喉科領域の医工融合：音声合成”，耳鼻科総説，2003年
9）矢頭 他：“テキスト音声変換技術と応用”，沖電気研究開発181号,Vol.66 No.2，pp.59–62，1999年10月
10）渡辺 他：“「ながら利用」向け音声合成Gatewayの設計と試作”，ヒューマンインタフェースシンポジウム2003論文集，pp.467–470，2003年9月
11）河井 他：“コーパスベース音声合成技術の動向[Ⅲ]ーコーパスの設計と評価尺度ー”，信学誌 Vol.87，No.3，2004年3月
12）“音声合成システムの性能評価方法ガイドライン”，JEITA-IT-4001，2004年3月

●筆者紹介

渡辺聡：Satoshi Watanabe. 研究開発本部 ヒューマンインタフェースラボラトリ チームリーダ
岩木健：Takeshi Iwaki. 研究開発本部 ヒューマンインタフェースラボラトリ
兼安勉：Tsutomu Kaneyasu. 研究開発本部 ヒューマンインタフェースラボラトリ
三木敬：Kei Miki. 研究開発本部 ヒューマンインタフェースラボラトリ ラボラトリ マネージャ