# 顧客へのリアルな音声応答を実現するテキスト音声合成技術「Cralinet」

間野 一則 /水野 秀之
中嶋 秀治 /宮崎 昇
吉田 明弘

NTTサイバースペース研究所

**私たちは，リアルな音声応答を実現するテキスト音声合成技術Cralinetを開発し，お客さまに提供しています．本稿では，コンタクトセンタでの情報案内サービスへの技術導入効果，さらなる高品質化のための統計的な姓名のアクセント推定技術とイントネーションの改善技術，そして音声合成の将来像について紹介します．**

## テキスト音声合成とは

テキスト音声合成とは，テキストを音声に変換して情報を伝える技術です．NTTでは，大規模な肉声品質の音声辞書（コーパス）を用いるコーパスベースアプローチ(1)を取り入れたCralinet（開発コード名：クラリネット）テキスト音声合成技術を開発し，お客さまにリアルな合成音を提供できるようになりました．本合成技術を用いることにより，今回紹介するコンタクトセンタでの情報案内をはじめとして，さまざまなサービスへの利用が期待されています．

## コンタクトセンタへの音声合成の導入

現在，もっとも一般的に音声サービスが行われている場所といえば企業のコンタクトセンタでしょう．これまでのコンタクトセンタでは，顧客からの問合せに対してオペレータが直接電話を取って応対する場合が多く，機械による自動応答は，問合せ内容に応じて応対先を振り分けるための案内など，部分的な導入にとどまっていました．また自動応答が導入されている場合においても必ずしも合成音声ではなく，事前に録音された音声が用いられる場合も多くありました．その大きな理由の１つとして，従来の合成音声の品質はいわゆるロボット音声的であり，オペレータの音声の自然性からは大きく隔たった聞き取り難いものであり，顧客サービスの観点からコンタクトセンタの利用には適さなかったことが挙げられます．

しかし，我々が開発したCralinetはオペレータの話す音声に近いリアルな品質の合成音声を実現しているため，録音音声と同様に取り扱うことが可能となりコンタクトセンタへの導入も進んできています．

実際にチケットサービス業務を行っている，あるコンタクトセンタにCralinetが導入された結果，業務効率の向上に効果があることが実証されています．このコンタクトセンタではチケットの販売や説明，興行内容の問合せの対応を行っていますが，定型的な問合せの対応にオペレータの稼働が多く取られており問題となっていました．そこで，こうした問合せに対しては，**図１**のように，顧客の了解を得た後，合成音声

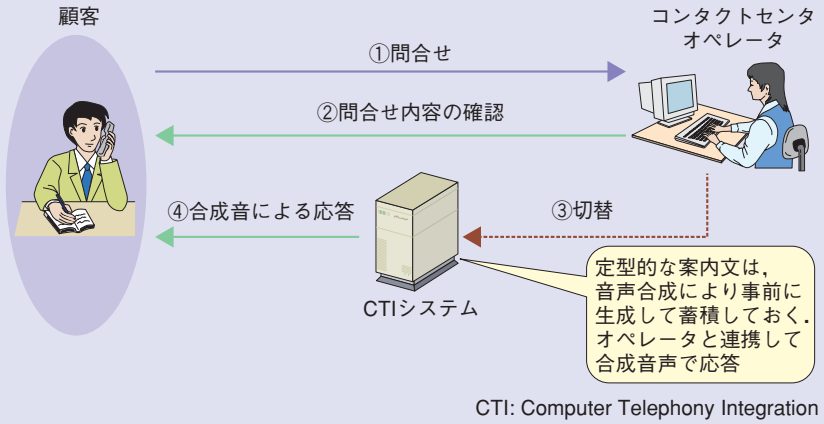


図１　オペレータと連携した合成音声案内システム

による案内に切り替えることにより，オペレータは次の電話を取ることができるシステムを導入して，オペレータの呼当りの通話時間を短縮し，受電率を向上しました．

音声合成と連携したコンタクトセンタシステムは，チケット業務以外にも銀行やクレジットカード，通販等の業務において，入出金や利用残高，ポイント数の案内等といったさまざまな問合せに対する応対に適用可能と考えられます．

またどのような業務のコンタクトセンタにおいても，なんらかの突発的な事情で通常以上に呼数が増加した場合に，接続待ちをしている状態の待呼や，接続をあきらめてしまった放棄呼の増大は，顧客サービスの観点から問題と考えられます．こうした状況においては，接続までの待ち時間や現在の受付状況の案内等を，逐次合成音声で案内するようなサービスを行うことで，顧客満足度の改善が可能ではないかと考えられます．

コンタクトセンタでの案内情報が日々更新される場合，それまで毎日録音していた作業も合成音声で作成することにより，発声にかかわる作業や声質の変化を気にすることなく，常に均一の音声をサービス提供できる点も音声合成を導入する利点の１つです．

このように，将来のコンタクトセンタでは音声合成技術が組み込まれたシステムを用いることで業務の効率化や顧客満足度の向上が図れるのではないかと期待されています．

## Cralinet音声合成概要

こうしたコンタクトセンタでの応用では，前述のとおり合成音声の品質が肉声に近いものであることのほかにも，お客さまの名前や住所，キーワードを正しく読むことが重要となります．テキスト解析部での正しい読み・アクセントの付与と音声合成処理部での正しいイントネーションを持った合成音声を生成することが重要となります．

Cralinetの構成は，**図２**のように，テキスト解析部と音声合成処理部からなっています．テキスト解析部においては，汎用性を高めるために，辞書にない未知語の入力に対する読み・アクセント推定機能を持ちます．音声合成処理部では，人間の発声パターンを網羅したコーパスを用いて，与えられた韻律目標に対して，音のつながりと声の高さをスムーズに，かつ，正しいイントネーションで音声を生成する機能を備えています．

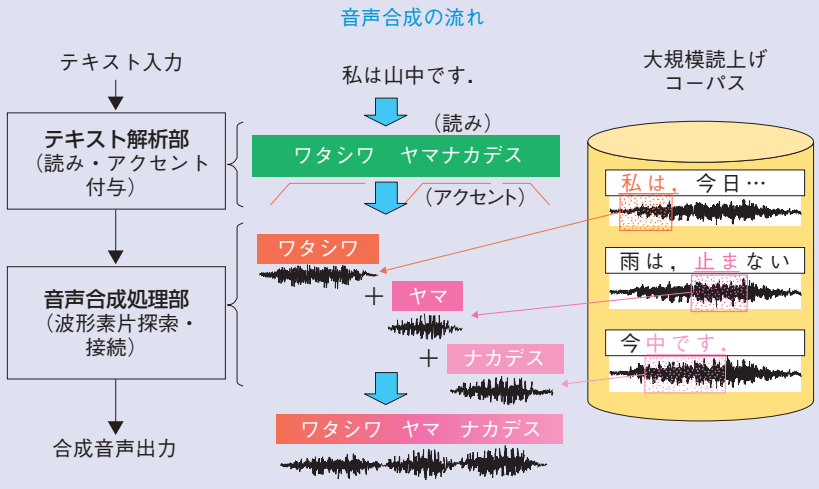


図２　Cralinet音声合成概要図

## テキスト解析部

顧客に音声合成を使って情報伝達を行うときに，「～さまの現在のご利用…は」と顧客の名前を声で呼ぶ場合がしばしばあります．通常顧客対応の窓口には顧客情報のデータベースがあり，それには顧客の漢字表記の名前とその読みが登録されています．しかし，アクセントの情報までは普通は登録されていませんので，正しく声に出して読むためには，名前の読みからアクセントを正しく推定する必要があります．

アクセントは音の高低で表現されますが，標準語の場合，例えば平仮名２文字で表される単語であれば，平仮名文字数に１を足した三通りのうち１つのアクセント型を取ります．例えば「ハシ（ヲ）」という読みに対して三通りのアクセント型があり得て，それぞれ「端（を）」「橋（を）」「箸（を）」に対応します．

従来のアクセント推定では，もっとも出現頻度の高いアクセント型を用いたり，人手で作成された規則による推定が行われていました．しかし，読みによっては，出現頻度の低いアクセント型もありますし，規則による方法では，新しい名前を処理するために規則の追加・修正を行う必要があります．その際，従来の推定との整合性を取るために，規則の複雑化と整備のコスト高となります．

アクセント型の推定という問題はその単語がどのアクセント型の単語群に

属するかという分類問題ですから，Cralinetでは統計的な機械学習によって事例から自動的に構成された分類器（SVM: Support Vector Machine）を使う方式を用いています．この方式では**図3**(a)に示すように，分類したい読み（アクセントが未知の単語の読み）と各アクセント型（分類クラス）を代表する複数の読みとの類似度が最大になるアクセント型に分類されます．

例えば，ニュースに出現した未知の姓と名に対して，従来法でのアクセント型推定率は，それぞれ82%，79%でしたが，Cralinetでは91%，86%と高い精度が得られることを確認しました．

図3(b)のように音声合成におけるテキスト解析処理では，入力された漢字仮名混じりの文を単語に分割し，辞書にある単語には辞書の読みとアクセントと品詞を付与します．辞書に登録しきれない人名のような語のアクセント型については，ここで紹介したアクセント型推定法を適用します．これらを用いて，文全体において適切なアクセントやイントネーション，ポーズなどの韻律情報を生成し，音声合成処理部に渡します．

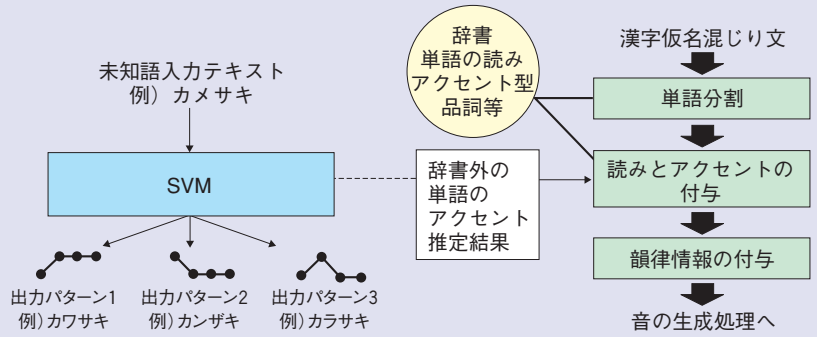


図3　テキスト解析部

## 音声合成処理部

音声合成処理部では，テキスト解析処理で得られた入力テキストの読みと合成したい音声の高さや長さを表す韻律情報（ターゲット情報）に合った波形素片（最小単位は音素）をコーパスから素片候補を探索し，ターゲット情報にもっとも近い素片の組合せを接続することで，合成音声を生成します．

コーパスのデータ量が大きいほど波形素片のバリエーションは増え，最適な波形素片を利用することができ，高品質な合成音声を生成できます．しかし，数十時間分の音声が登録されている大規模なコーパスを用いても，従来の波形素片選択処理アルゴリズムでは，品質劣化した合成音声を生成してしまうことがありました．これは，選ばれた波形素片それぞれがターゲット情報に近くても，生成される文全体の韻律特性を必ずしも反映していないことによります．実際調査したところでは，品質劣化の種類は，時間に伴う音の高さの上がり下がりであるイントネーションが不自然になる例が多くみられました（**図4**）．

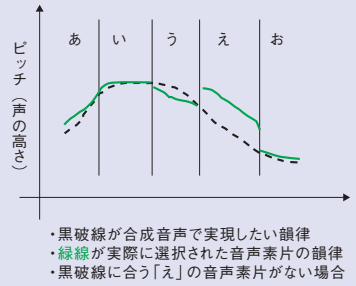


図4　ターゲットに対する合成音声候補の韻律情報の不自然な例

そこで，従来の素片選択アルゴリズムに加えて，合成音声のイントネーションの自然性を評価し，その評価結果に基づいて，文中の句全体にわたって総合的にイントネーションに不自然さのない候補を最終素片列として出力する方式を開発しました．このイントネーションの評価には，テキスト解析処理のアクセント推定でも利用されているSVMを用いています．

イントネーションの自然性を評価し合成音声を出力するまでの処理フロー例を**図5**に示します．まず従来の選択基準でもっとも良い波形素片の組み合わせに対してイントネーション評価をアクセント句ごとに行います．図5の例では候補A（赤線で処理の流れを示す）の“自然な”のイントネーションが不自然だと評価されます．そこで，候補Aを合成音声として出力せずに，新たな候補B（緑線で処理の流れを示す）を選択し，イントネーション評価を行います．すべてのアクセント句でイントネーションが自然であると判定された候補Bは合成音声として出力されます．

このイントネーションの評価を組み込んだ素片選択処理から生成される合成音声と従来の合成音声との対比較実験を行った結果，合成音声の約7割においてイントネーションが改善したという結果を得ました．

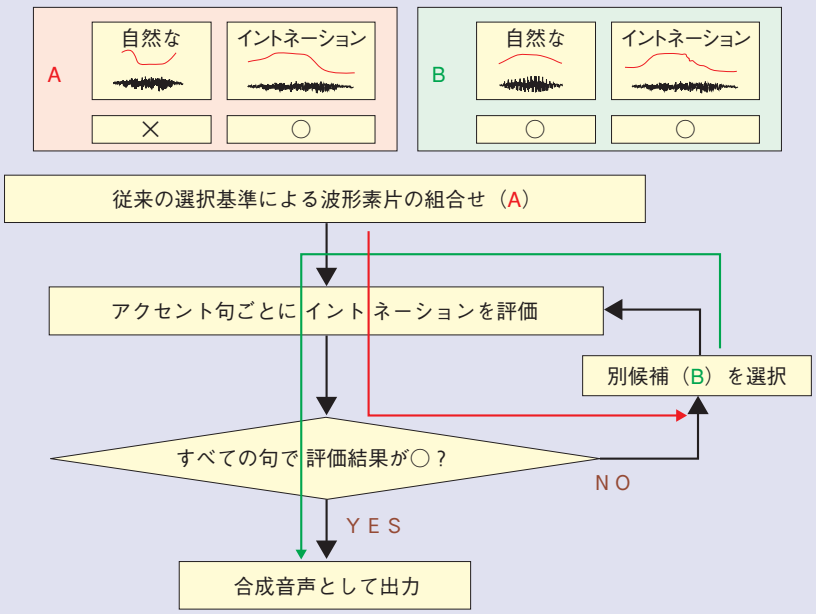


図5　イントネーションの評価に基づく合成音声出力フロー

## 音声合成技術の将来

今まで紹介してきた音声合成技術は，読み上げ音声と呼ばれる口調を主な適用先として考えられてきました．例えばコンタクトセンタの定型文案内代行のほか，株価情報，市況案内の代読など，淡々と情報を伝える場面を主としています．応用先もニュース，カーナビゲーションなどが主です．

しかし，このような適用先は，機械が人間の発声を代行するという音声合成本来の目的の，ほんの一歩目を踏み出した段階に過ぎません．私たちが普段耳にする音声には，人ごとに変わる声の変化（話者性）や，TPOにふさわしい口調の変化など，実にさまざまなバリエーションが存在します．現在でも，発声できる内容に制限を持たせた場合では，さまざまな声色，話者性を持つ点に特色を持つサービスも生まれています．代表的なものとして携帯電話の着ボイスやロボットの音声も受け答えをするためにさまざまな抑揚が表現されています．NTTサイバースペース研究所では，このように発話内容に制限がある場合だけでなく，どのようなテキストに対しても話者性や，口調に変化を持たせられるようになれば，音声合成技術の適用先を広げられるのではないかと考えています．

例えば，企業の電話自動応答メッセージ1つを取り上げても，さわやかな声や落ち着いたしっとりした声，元気のよい声など，企業が自社のイメージにふさわしいと思う合成音を選ぶことができます．また個人が用いるカーナビゲーションの案内音声などは，利用者によって声の質や話者の印象などに好みが分かれるところです．利用者の好みにきめ細かく応じて話者を入れ替えたり，流行の口調を導入したりといった，新しい楽しみ方が生まれてくるかもしれません．

残念ながら，現在の音声合成技術の枠組みでは，肉声と同等品質を保ったまま音声の話者性を変更することにはかなりのコストがかかります．また一口に「さわやかな」印象を与える音声といっても，どんなテキストに対してもそのような口調で音声を合成できるようにするためには，まだまだ技術的に未解決な部分が多く残されており，これらの課題に取り組んでいきます．

使う人，1人ひとりの好みに合った馴染みのある声が，TPOに沿った自然な口調で機械から流れ出すとき，今までの無機的なインタフェースに囲まれたIT社会とは一味違った世界が広がっていることでしょう．

**■参考文献**

(1) 水野・磯貝・長谷部・浅野・阿部：“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技術ジャーナル，Vol.16，No.1，pp.23-26，2004.

**（後列左から）宮崎　昇/ 間野　一則/ 中嶋　秀治**
**（前列左から）吉田　明弘/ 水野　秀之**

Cralinetをはじめ，多様な声を実現する音声合成技術を開発し，お客さまに感動と希望を与える音声合成サービスの実現に取り組みます．

**◆問い合わせ先**
NTTサイバースペース研究所
音声言語メディア処理プロジェクト
TEL　046-859-3938
FAX　046-855-1054
URL　https://www.ntt.co.jp/cclab/contact/index.html