Vol.2011-SLP-88 No.6
2011/10/28

電話応答サービスに適した音声合成の開発

加藤正徳　近藤玲史　三井康行†

電話応答サービスに適した音声合成の課題と、課題解決に向けた取り組みについて紹介する。電話応答用途では、電話回線伝送に伴う了解性の低下とサービス構築コストが主要な課題となる。電話回線伝送に伴う了解性の低下を改善する目的で、伝達経路の特性の影響を受けにくい韻律の生成に、モデルとパターンルールを利用する方法を導入する。また、了解性を重視した数字の読み変えを行う。サービス構築コストの削減に関しては、メモリ消費を効率化して最大同時応答数を向上した。更に合成音声エディタを導入し、合成音声の修正効率の改善に繋げた。

Development of speech synthesis for telephone answering service

Masanori Kato, Reishi Kondo and Yasuyuki Mitsui†

Problems of speech synthesis for telephone answering service and some approaches for solving the problems are presented. Intelligibility of synthesized speech conveyed through a telephone channel and construction cost of the service are major problems for telephone answering service application. In order to improve intelligibility of synthesized speech conveyed through a telephone channel, we introduce the method utilizing the models and pattern rules for generation of prosody which is less affected by a telephone transmission channel. Reading of numbers is also changed to enhance the intelligibility. For saving the service construction cost, the maximum number of simultaneous response is increased by the efficient memory usage of the speech synthesizer. Moreover, the efficiency of synthesized speech modification work is improved by introducing a speech synthesis editing tool.

1. はじめに

音声合成は、計算機を用いて音声を人工的に生成する技術であり、様々な音声インターフェースを実現する上で、音声認識などと共に必要不可欠な基盤技術の一つである[1]。1980 年代頃の合成音声の品質は、「ロボット声」に例えられるような人工的なものであり、アナウンサやナレータが発声した肉声を代替するには厳しい品質であった。ところが、1990 年頃に登場し、現在主流となっているコーパスベース音声合成の品質は、従来の音声合成に対する印象を覆すほど高く、電話応答サービスをはじめ、様々な商用サービスに利用できるほどのものとなった。

音声合成を電話応答サービスに利用する主な利点としては、短期間かつ低コストで電話応答サービスに用いる電話応答音声を作成できる点が挙げられる[2]。電話応答サービスに利用する音声を音声収録により用意する場合、アナウンサや収録スタジオの手配、収録作業、収録音声をシステムに組み込む作業などが必要となる。特に、電話応答音声に含まれることが想定される氏名や商品名等を予め全て収録する場合には、膨大な音声収録量が要求される。また、商品名などの変更・追加や、電話応答サービス自体の仕様変更に伴い、電話応答音声の内容は、電話応答サービスの運用中に変化することが多い。このため、音声収録を繰り返すことになるが、これが電話応答サービスの運用者にとって負担となっていた。

以上のような背景から、コーパスベース音声合成が電話応答サービス用途にも普及しつつある。実際、電話応答サービスへの音声合成の導入事例は近年増えており、電話応答サービスにおける音声合成の重要性は増大している。本稿では、電話応答サービスに適した音声合成における主要な課題である「電話回線伝送に伴う了解性の低下」と「サービス構築コスト」について説明し、これらの課題の解決に向けた取り組みを紹介する。

2. 電話応答サービスと音声合成

各種運営・運行状況などの情報提供サービスや、カスタマーサポート、商品予約・販売サービスなど、電話回線を利用した音声による電話応答サービスは、 様々な企業・団体から提供されている。音声合成の品質がサービスへの投入に耐えうる水準に到達してきたことから、電話応答サービスにも音声合成の利用が近年広がりつつある。

† NEC　情報・メディアプロセッシング研究所
Information and Media Processing Labs., NEC Corp.

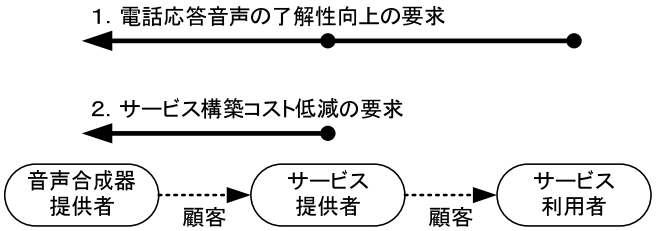

図 1　電話応答サービスの関係者

図 1 に示す通り、音声合成器の提供者には、電話応答サービスの利用者、提供者の二種類の顧客が存在する。サービス利用者は、サービス提供者の直接顧客に相当する。音声合成器の提供者には、サービス利用者から生じるニーズである了解性の向上が両者から要求され、サービス提供者からサービス構築コストの低減が要求される。それぞれのニーズを満足することが、電話応答サービスに適した音声合成器を開発する上で重要となる。次節では、音声合成器の提供者の顧客ニーズから浮上した課題である了解性、特に電話回線伝送に伴う了解性の低下と、サービス構築コストについて詳しく説明する。

3. 電話応答向け音声合成における課題

3.1 電話回線伝送に伴う了解性の低下

了解度とは、言語的に意味のある単語や短文が、どれだけ正確に伝わったかを測る尺度である[3]。電話応答サービスに当てはめると、電話をかけたサービス利用者が聞いたときに、応答音声の内容がどれだけ正確に伝わったかを測る尺度と言える。電話応答サービスでは、正確に伝ったか否かのみならず、合成音声を聞くサービス利用者に負担が少ない品質であることが求められる。そこで本稿では、本来の「正確に伝わったか否か」に加えて、聞き取り易さを含めて了解性と呼ぶこととする。つまり「了解性が高い」とは、サービス利用者が少ない負担で正確に聞き取れることを意味する。

電話応答音声は、電話音声帯域の周波数特性や音声符号化歪みなど、音声の伝送路特性の影響で、了解性が損なわれる。このことは、合成音声に限らず、人間が発話した自然音声にも当てはまる。図 2 は、ITU-T P.48 で規定されている電話音声の伝送路の周波数特性である[4]。電話音声は、図に示すような周波数特性を有する伝送路を経由して通話相手に届くため、PC などで標準的に再生される音声と比較して、こもった感じの音声となる。特に、高い周波数帯域の成分が支配的な /s/ や /sh/ などの摩擦音は、聞き取り易さが低下しやすい。

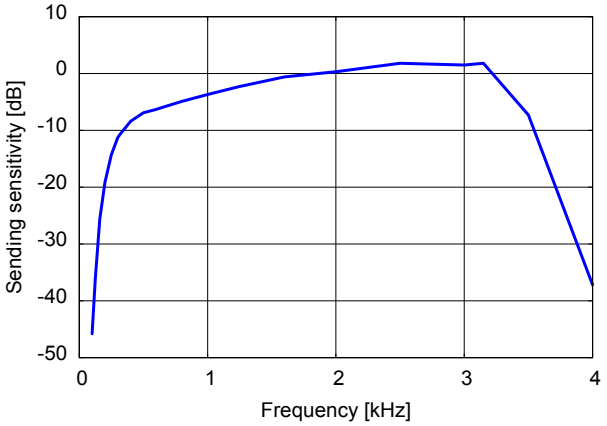

図 2 電話音声帯域特性

また、音声符号化歪みによる劣化も、了解性を低下させる要因となる。オーディオ符号化とは異なり、音声符号化はビットレートが低いため、通話音声の音質は音声符号化処理の影響を受けて低下しやすい。電話音声の伝送路特性を考慮して主観音質評価を遂行している例としては、例えば携帯電話向け音声通信に利用されているノイズサプレッサがある。ノイズサプレッサ(NS)は、通話音声に混入する背景騒音成分を抑圧し、聞き取り易くクリアな音声にする技術である。携帯電話の規格を策定する 3GPP は、携帯電話に応用する NS を評価する場合、NS 単体では無く、AMR コーデック処理や電話音声帯域フィルタ処理も含めた処理音の評価を要求している[5]。

音声の伝送路特性による了解性の低下は、元から了解性が低い合成音声のほうが自然音声よりも深刻である。通話音声の了解性を向上する観点からは、音声の伝送路特性の影響で劣化しやすい肉声感よりも、影響を受けにくい韻律の自然性が重要であると考えられる。

聞き誤りを低減するためには、場合によっては自然性が損なわれても、読み変えが有効である。電話応答サービスでは、日付や金額、番号などの数字の読上げが行われることが多い。特に数字には、通常の文よりも文脈からの推測が難しいという特徴がある。聞き誤り易い数字読みの組合せ例を表 1 に示す。通話音声の了解性を改善するためには、数字の読上げ方法にも工夫が必要である。

Vol.2011-SLP-88 No.6
2011/10/28

表 1. 聞き誤り易い数字読みの組合せ例

ヨーカ(８日)、トーカ(１０日)
ニガツ(２月)、シガツ(４月)
イチ(１)、シチ(７)
レー(０)、エー(A)

3.2 サービス構築コスト

電話応答サービスの構築コストに影響を与える要因として、最大同時応答数と合成音声の修正効率が挙げられる。最大同時応答数とは、１台の音声合成サーバで同時に応答可能な最大回線数のことである。電話応答サービスの構築者は、想定される同時着信数の最大値に基づいてシステムの仕様を策定する。もし想定される最大同時着信数が最大同時応答数を超える場合、複数の音声合成サーバを用意してシステムを構築する。複数のサーバで構成されるシステムは、サーバの追加や分散処理制御を必要とするため、サービス構築コストの増加を招く。つまり、１台の音声合成サーバが同時に生成できる合成音声の数を向上させることが、複数サーバ構成の回避、即ちサービス構築コスト増加の回避につながる。

合成音声の修正効率も、サービス構築コストに大きな影響を与える。現行の技術水準では、あらゆるテキストに対して常に正確な読みやアクセントを付与することは難しい。特に電話応答サービスでは、氏名や企業名、商品名など未知語となり易い固有名詞が多く含まれるため、言語解析に用いる辞書を強化しても、読みやアクセントの誤りは避けられない。読み・アクセントの誤りは了解性の低下に直結することから、誤りを人手で修正することが電話応答サービスの品質確保に有効である。

更に電話応答サービスでは、ポーズ制御の工夫も了解性の向上には有効である。例えば、伝達情報として重要な情報の前後には通常よりも長めポーズを挿入する、四桁数字であれば、やや長めのポーズで一桁ずつ区切る、などの方法が採用される。以上のような読み・アクセントの修正やポーズ挿入は、音声合成に関する技術的知識が不十分なサービス構築者によって行われる。こうした事情を考慮した対応が、サービス構築コストの低減には重要である。

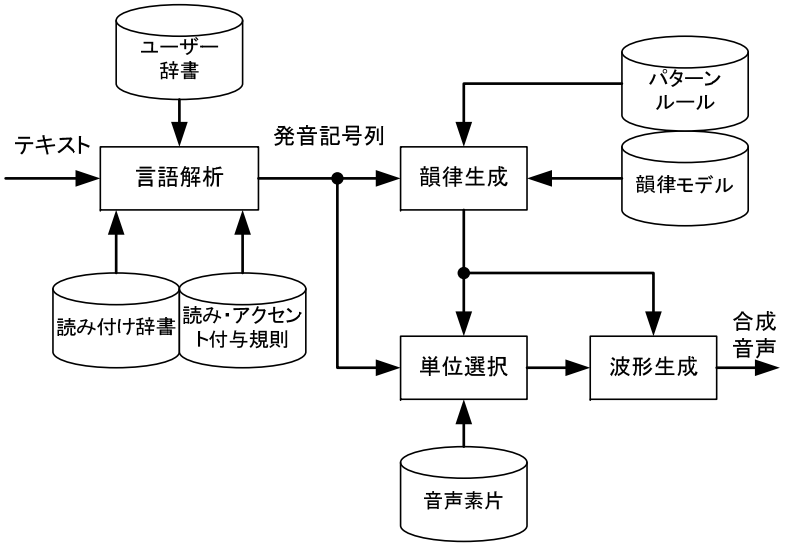

図 3 音声合成器のブロック図

4. 電話応答用途における課題の解決に向けた取り組み

前節で説明した課題の解決に向けた取り組みについて紹介する。 先ず開発した音声合成器の概要を説明し、続いて電話回線伝送に伴う了解性低下の改善策と、サービス構築コストの低減への貢献について述べる。

4.1 音声合成器の概要

音声合成器の構成を図 3 に示す。まず、入力された漢字仮名混じりテキストを言語解析処理によって形態素に分割し、辞書及び読み・アクセント付与規則に基づいて読みとアクセントを決定する。また、ポーズの位置や長さの推定も行われ、発音記号列に変換する。次に、音韻継続時間長(発話のリズム)や基本周波数時系列(声の抑揚)などの韻律を生成する。音韻継続時間長は、音韻の種類だけでなく、文頭や文末などの位置に応じて異なる。このことから、音韻の種類や文中の位置を参考に、音韻毎に計算する。声の抑揚を与える成分である基本周波数の時系列は、発音記号列に含まれるアクセント核やアクセント句区切り、ポーズなどの位置を参考に生成する。

単位選択では、発音記号列と生成された韻律に従って、合成の単位である音声素片を選択する。少ないデータ量でも任意の文を合成できるように、音声素片の最小基本単位として「CV 単位(C:子音、V:母音)」と呼ばれる合成単位を採用している。音声素片の基本周波数を大きく上下させたり、継続時間長を大きく伸縮させると素片の音質が著しく低下するため、音声素片を選択するときは、生成された韻律に近い韻律を有するものを選択する。また、隣接する素片同士のスペクトルが大きく異なると接続歪みが生じるため、素片の接続性が良いものを選択する。すなわち、生成韻律と素片韻律の差分や、隣接素片同士のスペクトルの差分が小さくなる素片を動的に選択する。最後に、選択された音声素片を接続して合成音声波形を生成する。その際、音声素片の継続時間長や基本周波数を、生成韻律に適合するように補正する。

4.2 電話回線伝送に伴う了解性低下の改善

電話音声帯域における了解性の向上を目的として、収録音声の話者に女性アナウンサを採用した。電話音声帯域における声の通りは男性よりも女性のほうが良いため、了解性の向上にも効果がみられた。電話応答サービスでは、数ギガバイトのディスク容量を低コストで利用できるサーバが使われるので、データ圧縮に伴う合成音声品質の低下を回避する目的で、音声素片のデータ圧縮は行わなかった。その結果、音声合成データサイズは 1 ギガバイトを超えたが、素片へのアクセス方法を工夫することにより、速度への影響無く利用できるようになった。

了解性を向上するためには、音声の伝送路特性の影響を受けにくい韻律の自然性が重要であることから、韻律生成の改良に取り組んだ。収録音声コーパスから、前後の単語や文の長さなどを基に、文章中の単語をどのような抑揚やリズムで読み上げるべきかを機械学習を利用して抽出した。音声合成時には、抽出した韻律モデルと、人手で作成した少数のパターンルールを組み合わせて韻律を生成する。韻律モデルとパターンルールの切り替えには、収録音声コーパスから抽出した統計情報を利用した。統計情報を参照して学習データ不足を判断するので、データ不足により韻律モデルを適切に学習できていない条件では、パターンルールに切り替えて韻律の乱れを回避することが可能となる。以上より、韻律モデルとパターンルールを組み合わせることで、自然でバリエーションの豊かな韻律を安定的に生成できる方式を実現した。

数字読みは、前節で説明したとおり電話応答サービスでは利用頻度が高く、文脈からの推測が効き難い。そこで、了解性向上を目的とした数字の読み替えを導入した。表 2 に例を示す。例に示されているように、別の数字に聞き間違えることが多いと予想される読みだけでなく、母音の無声化に伴い了解性が低下するものも読み変えの対象としている。

表 2. 了解性向上を目的とした数字読みの変更例
(%は無声化記号[6])

表記	変更前の読み	変更後の読み
7	シチ	ナナ
4	シ	ヨン
0	レー	ゼロ
8日	ヨーカ	ハチニチ
2日	フ%ツ%カ	ニニチ
２０日	ハツ%カ	ニジュウニチ

4.3 サービス構築コスト削減への貢献

4.3.1 最大同時応答数の向上

最大同時応答数を向上する目的で、計算量とメモリ消費量の削減に取り組んだ。計算量削減のため、単位選択における素片候補の絞り込みを導入した。計算量は単位選択処理が最も多く、選択候補である音声素片の数に大きく依存する。そこで、前後の音韻環境情報を考慮した候補の足切りを導入した。この足切りにより、生成韻律との差分の計算回数を削減できる。また、隣接素片との接続性を計算する回数を削減するため、生成韻律との差分が大きい素片を候補から除外した。

メモリ消費量の削減では、音声合成器を初期化するときに確保するメモリ領域(共有領域)と、合成音声生成時に随時確保するメモリ領域(ワーク領域)とを分割する手法を導入した。図 4 は、音声合成を搭載した計算機のリソース関係を模式的に示している。メモリ上の共有領域には、実体の保存先を示す合成辞書インデックスや、合成用データの一部など、アクセス頻度が高いデータが初期化時にロードされる。電話回線毎に用意されたワーク領域には、声の高さや話速などの初期設定パラメータや、発音記号列や生成韻律、最適単位系列など、音声合成処理中に生成される中間データが保存される。読み付け辞書や音声素片など、サイズが大きくアクセス頻度が低いデータは、ＨＤＤから直接ロードされる。

電話応答サービスに着信があると、予め定義されたフローに従って、テキスト入力制御部が音声応答用テキストを読み込み、合成音声の生成が開始される。合成時は、共有領域やワーク領域にロードされた初期設定パラメータ、中間データ、合成辞書インデックスなどを参照しつつ、ＨＤＤから合成に必要なデータを随時ロードして、合成処理を進める。生成された合成音声は、最終的に合成音声出力制御部を経由して各回線に出力される。計算量とメモリ消費量を削減した結果、計算機リソースやシステ

ム構成に応じて、数十以上の最大同時応答数を実現できるようになった。

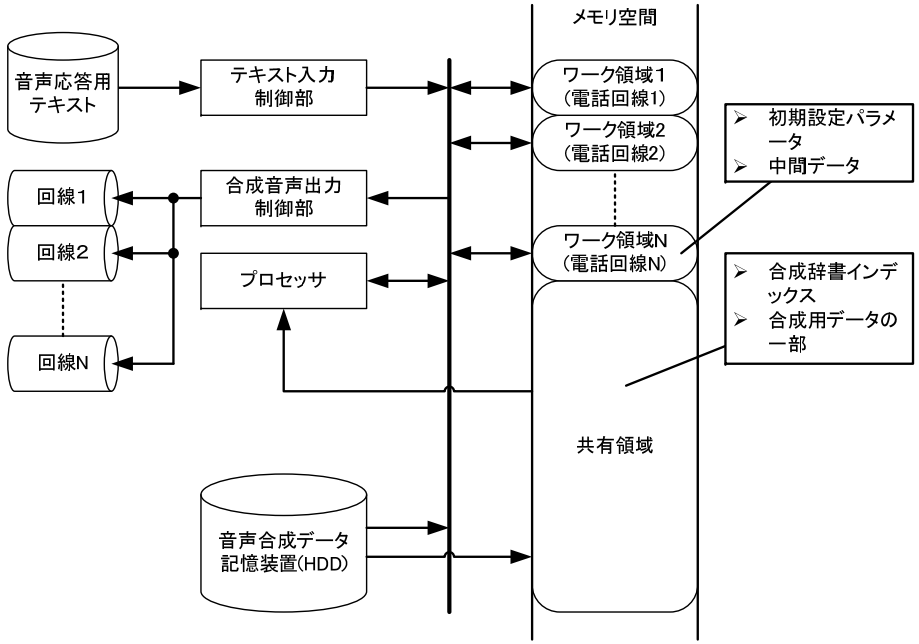

図 4 音声合成を搭載した計算機のリソース関係

4.3.2 合成音声の修正効率の改善

図 5 に示すようなインターフェースを有する合成音声エディタを用意して、合成音声の修正効率を改善した。このエディタでは、漢字仮名交じりテキストの他に、発音記号列、及びアクセント句単位に分解した発音記号列を表示する。また、アクセント位置の修正、ポーズの挿入・削除やポーズ長パラメータの変更、音節時間長とピッチパタンの修正を行うことが可能である。音節時間長は、修正値を変更倍率で指定する。ピッチパタンは、各アクセント句の始端、ピーク、終端位置における変更倍率を指定することで修正できる。

図 5 合成音声エディタ

図 6 に合成音声エディタによる作業フローを示す。先ず入力テキストを発音記号列に変換しつつ、合成音声を試聴する。合成音声の音質に不満があれば、次に発音記号列の修正に移行する。発音記号列の修正と合成音声の試聴を繰り返し、満足できる音質になれば作業を終了する。もし発音記号列の修正では満足できる音質に到達できないと判断した場合は、ピッチ・時間長の補正やポーズ長の変更など、発音記号列よりも細かい修正に移行する。そして、満足できる音質に到達するまで、修正と試聴を繰り返す。

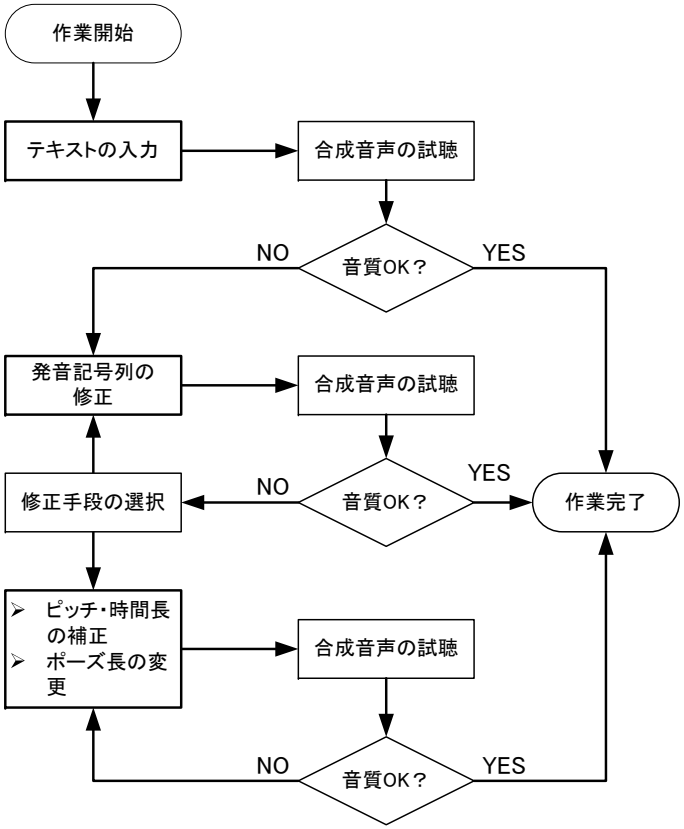

図 6 合成音声エディタによる作業フロー

5. おわりに

電話応答サービスに適した音声合成の課題と、課題解決に向けた取り組みについて紹介した。電話応答用途では、電話回線伝送に伴う了解性の低下とサービス構築コストが主要な課題となる。電話回線伝送に伴う了解性を向上する目的で、伝達経路の特性の影響を受けにくい韻律の生成に、モデルとパターンルールを利用する方法を導入した。また、了解性を重視した数字の読み変えを行った。サービス構築コストの低減に関しては、メモリ消費を効率化して、最大同時応答数を向上した。更に、合成音声エディタを導入し、合成音声の修正効率の改善に繋げた。

今後は、快適な電話応答サービスの実現に向けて、より表現力豊かな音声合成の開発に取り組んでいきたい。

参考文献

1) 小林隆夫: 小特集にあたって, 日本音響学会誌, Vol.67, No.1, pp.15-16 (2011).
2) 籠嶋岳彦: テキスト音声合成技術実用化の動向, 日本音響学会誌, Vol.67, No.1, pp.23-27 (2011).
3) 板橋秀一,他: 音声工学, 森北出版 (2005).
4) TELEPHONE TRANSMISSION QUALITY TRANSMISSION STANDARDS, ITU-T Recommendation P.48.
5) Minimum performance requirements for noise suppresser application to the AMR speech encoder, 3GPP TS 06.77 V8.1.1 (2001).
6) 日本語テキスト音声合成用記号, JAITA IT-4006 (2010)