論文の内容の要旨

論文題目　　新たな発話に対する柔軟性を備えた
コーパスに基づく音声合成システムに関する研究

氏　　名　　斉藤　隆

コンピュータ処理の飛躍的な進歩を背景に，主として大規模データベースを利用したコーパスベースの手法によって，音声合成の品質は近年大きく前進してきている．その一方で，音声合成の実際の応用に目を向けてみると，利用分野も徐々には広がりつつあるものの，市場展開の加速感はまだ感じられないと言っても過言ではない．音声合成の潜在的な市場を開花促進するために最も重要と考えられるのは，端的に言えば,「声の量産化」を目指す方向の技術である．つまり，ありとあらゆるアプリケーションの要求に応じられる豊富な声のバリエーションが，高品質であることは当然として，ごく手軽に得られること，言い換えれば，低コストで短期間に得られることである．ところが，発話の再現性をやや偏重した現在のコーパスベースの音声合成技術は，残念ながら「声の量産化」の方向を目指した技術とは言いがたい．

本研究の最終目標とするところは,「声の量産化」を可能とするような音声合成技術である．その量産化実現の先には，カーナビゲーション，e-ラーニング，ゲーム，携帯電話等のパイロット市場からも垣間見えるような音声合成の巨大市場が控えている．その大目標の下，本研究では，合成音声の高品質性を，新しい声や発話スタイルに対して，柔軟に，しかも，手軽に実現できるような音声合成システムを目指し，つぎの２つを技術課題とし

て設定する．

１．新しいスタイルへの柔軟性を高める

話者の音響的な特徴の再現性とともに，発話スタイル，特にイントネーションの多様性についても，柔軟に対応できるような仕組みを実現すること．

２．非専門家主導による音声データベース構築法の確立

声の量産化が行なわれるような市場においては，データベース構築の主役は，もはや音声の専門家ではなくなる．したがって，データベース構築プロセスにおける専門性の排除と，プロセスそのものの生産性向上が重要である．

第２章では，２つの主要課題に取り組むためにベースとなる音声合成システムとして，ボイスフォントに基づく音声合成システムを提案した．個人性を表わす２大特徴である音響特徴と F0 特徴をボイスフォントに蓄積し，それらの積極的な制御によって話者再現性を高めることに重点を置いた．多種多様な音声コーパスへの適応性を高めるためのシステム上の特長としては，（１）話者データベース更新負担の軽減： ボイスフォントは，音響・韻律特徴ともに生の「音声素片」データの集合として定義されており，音声コーパスから比較的容易に生成でき，しかも追加・削除が簡便に行なえること（２）話者データベース作成における専門性の排除と効率性向上：音声合成システムに合わせてデータベース作成処理を最適化し，両者を有機的に統合した作成ツールとして，非専門家にも扱えるシステムとして実現していること（３）適用分野での超高品質化を図る仕組み： イントネーションまで含めた録音音声との組み合わせ合成（シームレス音声合成）をベースライン・システムの自然な拡張として実現していること，以上の３点が挙げられる．

提案システムの主要な構成要素について下記のとおり検討を行なった．まず，音声生成の方式に関して，時間領域ピッチ同期重畳法の問題点とその改良法について議論した．ピッチマークの付与に関しては，ウェーブレットを応用した自動付与の手法を提案し，自動抽出の精度としては 96%を実現した．さらに，音声品質の観点からも評価を行ない，デバイス（ラリンゴグラフ）で声門閉鎖点を高精度に抽出する方法と比べて遜色ない性能を確認した．合成単位については，まずコーパスベース方式における合成単位をコーパス規模との関係から分類を行ない，その中での本研究の位置づけを明らかにした後，その考えに基づいて構成した合成単位定義とその選択特性等についての検討を行なった．

第３章では，コーパスベース手法のデータ再現性という最大の特長を保持しつつも，朗読音声だけでなく多様な発話特性をもつようなコーパスも柔軟に受け入れられるようなイ

ントネーション生成の方式を提案した．提案方式の特長は，①話者データが本来有している情報をできるだけ欠落させずにF0パターンを生成することによって再現性を高めること，②シームレス音声合成方式の導入によって多様な発話特性を任意文合成処理に自然な形で取り込んだこと，の２つに集約される．①と②の同時実現は従来のコーパスベース方式には見られなかったことであり，これによって実音声と合成音声のシームレスな融合が音響的な特徴だけでなく韻律的な特徴も含めた形で実現される．

提案方式について有効性を確認するための評価を実施した．まず，アニメーションの発話スタイルを含む複数の音声コーパスに適用して F0 予測能力についての評価を行ない，7 つのコーパスの平均でオープン評価の結果，文 F0 パターンの予測誤差として 0.21 [oct] を得た．主観評価としては従来システムとの比較評価を行ない，提案手法に対し 77.8 % のプリファレンス・スコアを得た．また，シームレス音声合成に関しても評価実験を通してその効果を確認した．

第４章では，ターゲットの音声合成システムに必要となる話者の音響的な特徴と韻律的な特徴の両方を，音声データから効率よく抽出する階層的な音声セグメンテーションの方法を提案した．本手法では，音声合成単位に必要となる音素区分の情報だけでなく，呼気段落，アクセント句，ピッチマークから構成される多階層の区分情報を，音声データから一括して抽出することで，波形素片の情報だけでなく，イントネーション生成のための F0 形状素片の情報も同時に得ることができる．音素セグメント境界抽出の頑健性を向上させることを目的として，得られた音素境界の信頼度を評価するための新たな尺度である，セグメント信頼度を導入した．

提案手法について複数の音声コーパスを用いて評価実験を行ない，音素境界誤差についての全話者平均として 10.9 ms を得た．この値としては，参考比較ではあるが HMM ベースの方法と比べても遜色ない結果である．また本手法は，異なる話者やイントネーションのスタイルに対しても，安定したセグメンテーションが行なえることを確認した．セグメント信頼度の導入によって，音素境界誤差が平均で 25.8 %，最大で 33.5 % 減少し，概して元の誤差が大きいほど改善効果が顕著に表われ，セグメンテーションの安定化における導入尺度の効果を確認することができた．

第５章では，データベース構築における負荷軽減に対して，特に，非専門家主導のデータベース構築という視点から議論した．非専門家によるデータベースの構築を前提におくと，自動処理の性能向上だけでなく、人手の介在する処理についても効率的なプロセスを

確立し，データベース構築過程全体の最適化が必須である．そういった考えに基づいたツールとして，ボイスフォント・ビルダーを提案した．さらに，ボイスフォント・ビルダーを用いて，データベース作成過程を定型化した作業として実施する手順を確立した．この手順はデータベース作成作業をあたかもコンピュータ・プログラムのコンパイル作業のように，定型化された処理作業の繰り返しとして実行可能なもので，音声の専門家でなくても短期間の学習で習得することができる．たとえば，音声アプリケーションの開発ベンダーがソフトウェア開発の一部として行なうことも想定できる．実際に，ソフトウェアベンダーによるフィールドテストを行ない，音声の専門家不在でも同等品質のデータベース作成が可能であることを確認し，開発期間については，従来の作業形態（音声専門家の指導の下で分析結果を網羅的にチェックする方法）の３分の１程度に短縮された．さらに，構築した話者データベース（ボイスフォント）を用いて話者性に関する主観評価実験を実施し，ベースライン・システムの話者性保持性能について確認した．

以上をまとめると，本研究では高品質であるとともに発話に対する柔軟性を備えた音声合成をテーマとして，新しいスタイルに対する柔軟性を高めることと，非専門家主導のデータベース構築法を確立することの２つを技術課題として設定し検討を行なった．その中で，ボイスフォントに基づく音声合成システム，スタイルに柔軟性をもつイントネーションの生成法，頑健性をもつ音声セグメンテーション法，非専門家主導のデータベース構築ツール，等を提案し，実験的評価やプロトタイプの実装を通じて，提案方式の有効性，システム実現の妥当性について検証した．その結果として，新しいスタイルに対する柔軟性に関しては,特徴的なイントネーションのスタイルをもつアニメーション音声に対しても，通常の読み上げ音声と同等レベルの話者性保持の効果が得られることを確認した．また，非専門家主導のデータベース構築に関しては，フィールドテストによって，非専門家であっても専門家と同等レベルのデータベース品質が得られること，また，提案ツールの利用によって従来の 3 分の１程度に開発期間が短縮できることを確認した．

これらの検討結果を総合すると，本研究の成果によって，新しい音声コーパスを用いた音声合成の開発容易性について従来に比べ少なからず進展がみられたものと考える．