テキスト音声合成技術

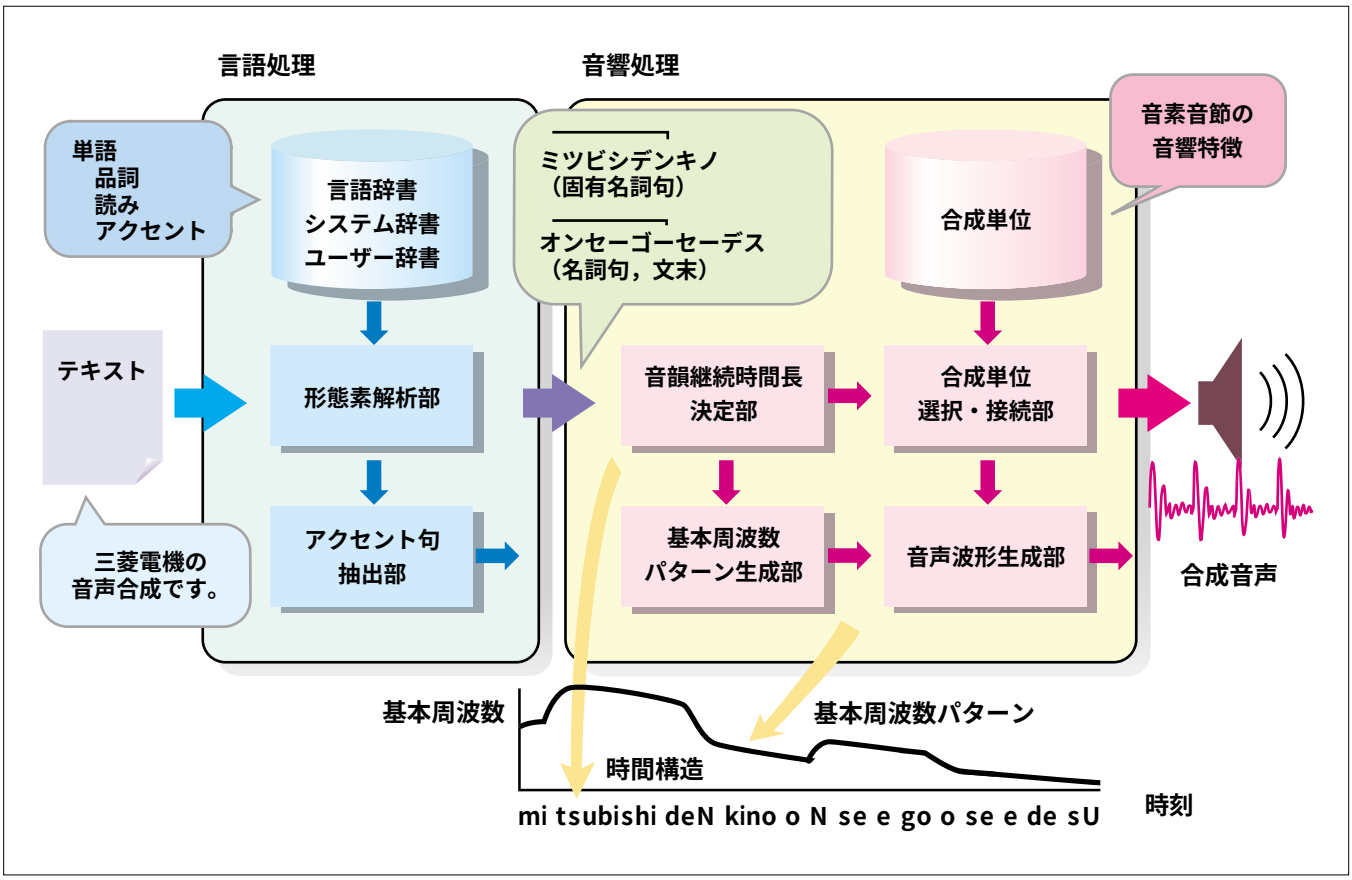
藤井洋一*
石川　泰*

要　旨

任意のテキストを音声に変換するテキスト音声合成技術は，運転中で目が走行方向から離せないカーナビや，ディスプレイが小さい携帯端末向けのインタフェースとして注目されている。また，障害者や老人をIT社会における情報弱者としないためにも，ますますこの技術への要求が高まっている。しかし，テキスト音声合成の品質は，最近の研究の蓄積によって向上しているものの，今後の本格的実用化と利用分野拡大には，更に品質向上を果たす必要がある。

このテキスト音声合成の処理の概要を図に示す。処理は，言語処理と音響処理に大別される。言語処理では，入力文から，読みの単位となるアクセント句と，その読み，アクセント位置，文の構造を抽出する。音響処理では，まず，基本周波数(声の高さ)のパターンを規則によって生成し，個々の音素の音韻継続時間を決める。次に，音素や音節の音響的な特徴を記憶している合成単位を接続し，合成音声を生成する。したがって，言語処理の性能は，合成音声の読みやアクセントの正確さを決め，音響処理は，自然性，肉声感，明瞭性を決める。このうち，音響処理の品質向上では，生成された音声の特徴がある特定の音声データに近いだけではなく，利用者の要求に応じて発話速度や抑揚のパターンなどの制御ができることが必要となる。

本稿では，聴覚特性と言語構造に着目し，自然性が高くかつ制御が容易な音韻継続時間長制御方式を中心に開発方式を説明し，今後の課題と，応用分野についての展望を述べる。



テキスト音声合成方式の構成

テキスト音声合成では，入力されたテキストを言語処理で解析し，アクセント句(発声上のまとまり)，その読みとアクセント位置，句の言語的属性を抽出する。これら情報から，音響処理は，個々の音素の時間長を決定し，文全体のリズムを生成するとともに，基本周波数(声の高さ)のパターンを生成し，文の抑揚を作成する。さらに，読みの情報から音声の基本的な単位の特徴を接続し，作成した声の高さで合成音声を生成する。

*情報技術総合研究所