音声合成システム

２２１４０００３　佐倉仙汰郎

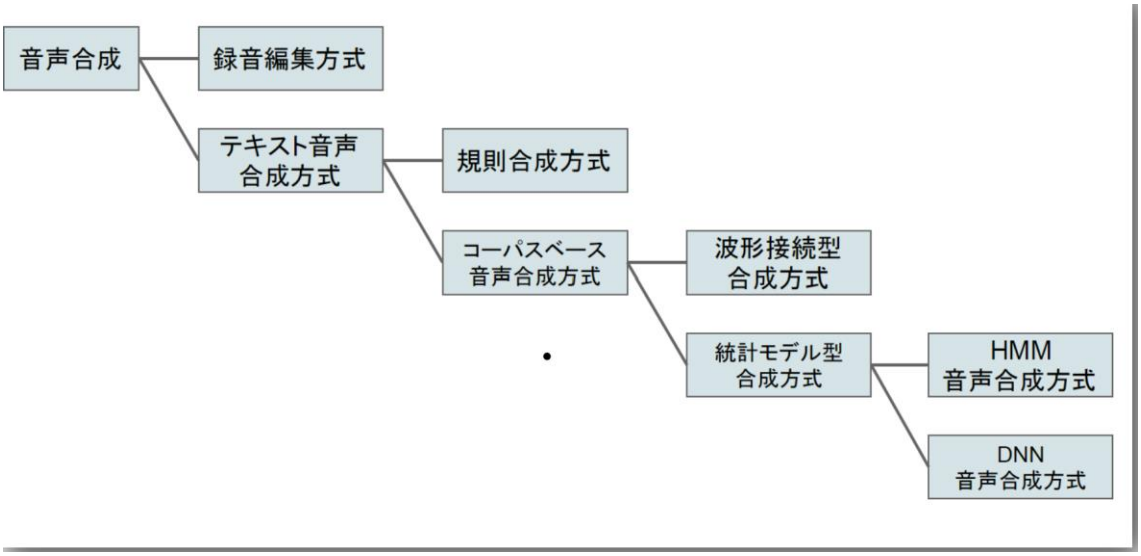
- 音声合成システムの概要

音声合成システムとはコンピューターを用いて人工的に人間の声を作り出す技術である。これによってテキストから実際の発声を聞ける。

１７９２年にフォン・ケンベレンが機械式音声合成機を発明し音声合成の研究が発展した。１９４０年代に入りコンピューターが発明されさらに研究が加速していき１９５０年代に日本で初めて現代のようなコンピューターを使って音声合成システムが完成した。２０００年代からは初音ミクを筆頭として、音声合成システムが多くの一般の人にひゅきゅうされるようになった。また 2013 年には初の深層学習方式の音声合成が Google により発表しさらに音声合成システムの研究は発展を続けている。

- 音声合成システムの仕組み

現代使われている音声合成システムにはいくつか種類が存在しており、ここではそれらを紹介する。



（think it より引用）

１．録音編集式

これは、事前に単語をレコーディングして音声を入力する。そしてそれを組み合わわれている。事前に、１００m先、２００m先、などと音声を入れて置き適切なタイミングで１００m先交差点を左です。などと組み合わせて伝えるのである。これは読み上げる内容が限定的なときのみに使え、そうでないときにはさらに大量なレコーディングが必要である。

２．テキスト音声合成方式

これは現代主に使われている技術である。音を非常に細かい単位で分け、それらから音声を合成する方式である。これによって非常に柔軟に様々なテキストに対応することができる。そしてこの方式は「規則合成方式」と「コーパスベース合成方式」に分けられる。

（1） 規則合成方式

これは事前に音のルールを決定してそのルールに基づき音声を合成していく方式である。現代はあまり使われていない技術で、理由としてはこのシステムの構築には研究者が多くの時間をかけてルールを設定しないといけないからである。

（2） コーパスベース合成方式

これは現代の主流である。大量の音声録音データとそのテキストをデータベース化した「音声コーパス」を作り、それに基づき統計的な手法で合成音声を生成する手法。コンピューターの性能が上がると同時にこの分野は発展してきた。またこのコーパベース合成方式はさらに 2 種類のタイプがある。

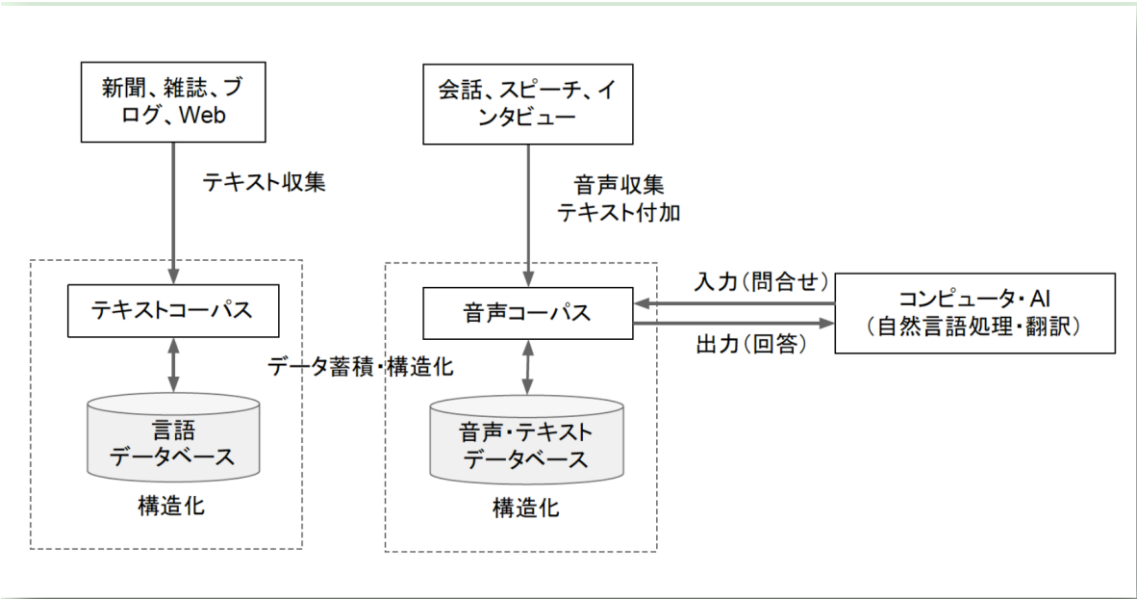
- 波形接続型合成方式

  あらかじめ録音された声を適切な単位に分割し、それらを組み合わせていく方式。これは録音編集方式に若干似ているがこちらのほうが細かく区切られている。欠点としては非常にたくさんの区切られた音を合成しなければいけないので、その切れ目を違和感なくつなげるのが難しいという点である。

- 統計モデル型合成方式

  事前処理として、音声コーパスから抽出した音響特徴量を機械学習を使って分析・モデル化し、それに基づきテキストの言語解析結果からその音声を予測し合成します。波形接続型合成方式に比べて少ない音声データでも安定した合成音声を生成できるため、近年この手法の利用が急速に進んでいます。また最近では、波形接続型合成方式と統計モデル型合成方式を組み合わせることによって双方のデメリットを最小限に抑えようとする「ハイブリッド方式」の開発も行われています。（coestation より引用）

統計モデル方式には主な手法が二つあり、HMM 音声合成と DNN 音声合成がある。DNN 音声合成は近年の主流であり、大量なデータを用いて機械学習を行いより鮮度の高い音声合成を目指している。



（think it　より引用）

- 音声合成システムの応用

  コールセンター、ニュースのアナウンサー、医療福祉現場での対話ロボット、音楽生成などがあげられる。

  特に音楽生成の精度は近年目を見張るものがある。AI が曲と歌詞を作り有名なアーティストに似せた声を使って音楽を作るのである。様々な問題が懸念視される中、応用分野の発展はとどまることを知らない。

まとめ

音声合成システムは様々な手法があり、現代応用されているものもたくさんの種類がある。しかし近年の研究の発展は素晴らしく AI 技術の向上とともに発展を続けている。これから音声合成システムによりさらに豊かな生活が享受できることを期待している。