E-023

少ない音声サンプルからの違和感のない自然な波形の生成
A study of generating a natural waveform with no discomfort from a few voice samples

高野　加奈絵† 　前田　涼佑† 　藤村　真生‡
Kanae Takano　Ryosuke Maeda　Masao Fujimura

1. はじめに

音声合成は、文章や文字などのテキストを人間の発話に近い音に変換する技術である。現在、日常の様々なところで実現されている。身近な例として電車の構内アナウンスがあげられる。従来、駅の係員が放送室などから電車の行先駅または停車駅のアナウンスを行っていたが、業務の効率を上げるために自動化を導入した。近年は電車の遅延や電車が運転休止した場合にアナウンスで放送するなど、乗客に対するサービスが向上している。

音声合成技術は、大きく分けて波形接続型音声合成とフォルマント合成がある。フォルマント合成は、録音された人間の音声は使わず、周波数、音色、雑音レベルなどのパラメータを調整して波形を作り、人工的な音声を作る方法である。合成された音声はロボットに近い音声になるが、波形接続型音声合成のような音声データベースは不必要なためデータのサイズは小さくて済む。また、イントネーションや音色を自由に変化させることができる。

一方で波形接続型音声合成は、図１に示すような処理手順をとる。あらかじめ人間の発話した音声を録音し、これを基にデータベースを作成しておく。実際の合成処理では入力されたテキストを解析して発話に適した単位に分解し、分解されたテキストの単位に適した人の音声の断片を選択し連結して合成する方法である。

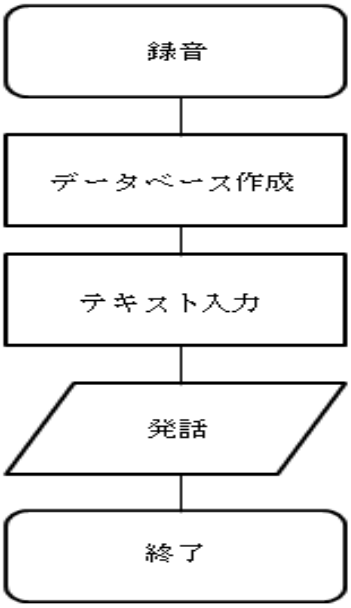

図１　波形接続型音声合成の処理過程

2.　研究背景

2.1　コーパスベース方式

本研究では、より自然な発話を実現するために、人間の声に近い音声合成が可能である波形接続型音声合成を使用する。波形接続型音声合成のうち、コーパスベース音声方式と呼ばれる方法は現在主流となってきている。コーパスベース音声方式の処理過程[1]を図２に示し説明する。

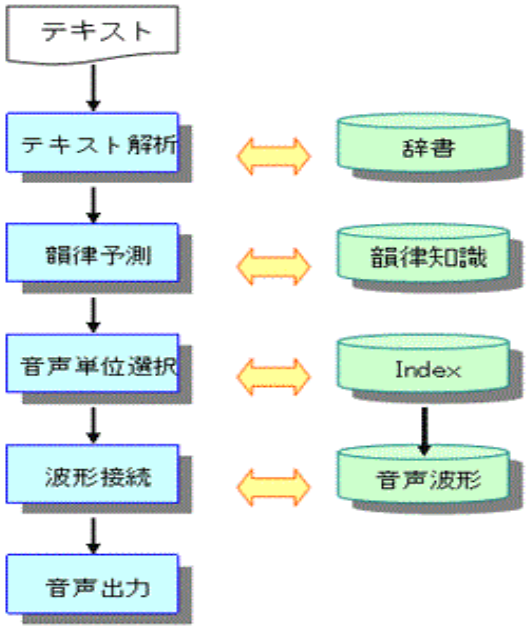

図２　コーパスベース方式の音声合成処理過程

コーパスベース方式の音声合成では、数十分から数時間の録音された音声からなるデータベースを使用する。

データベースを作成するためには、録音した音声を「音」、「音節」、「形態素」、「単語」、「成句」、「文節」に分ける。それぞれ分けた音声を波形として表し、各パラメータの抽出や統計データを算出する。

実際に音声を合成する際には、作成したデータベースから最も適した音声波形を探索／選択し、合成する。

基のデータベースが人間の声であることにより、人間に近い声に合成することが可能になる。しかし、接合部分で不自然になる場合がある。より自然に聞こえる音声を合成するにはデータベースの情報量を増やす必要がある。また別の人の声で音声合成を行う際は、新たに音声を録音しデータベースを作成しなおさなければいけないため、多くの時間とコストがかかる。

したがって任意の人の声を合成することには膨大な手間がかかり、現実的には不向きな手法であるといえる。

† 大阪工業大学大学院, Graduate School of Engineering, Osaka Institute of Technology
‡ 大阪工業大学, Osaka Institute of Technology

2.2　本研究の目的

本研究はコーパスベースの音声合成手法を基本的な手法として採用した。最終的には任意の人の音声を合成することが目標ではあるが、現段階では限定的な人の音声を合成することをまず考える。ここで音声合成分野は医療の分野でも応用されていることに鑑み、研究の対象を音声の障碍者とした。あらかじめ音声に障碍のある人に音声を録音してもらい、コーパスベース音声方式に基づいてデータベースを作成し、任意のテキストを入力すると録音した人の声で読み上げるシステムを実装する。

2.3　実現方法

本研究が対象としている音声の障碍者は、数十分から数時間話すことが困難なので、波形接続型音声合成で処理するとデータが少なく不自然な発話になる。そこでデータを補うための音声波形部分をフォルマント合成で処理することによって「より自然な発話」に近づけることができる。音声の障碍者に録音してもらう際は短時間で済むよう、あらかじめ膨大なデータベースを作成する必要がある。違和感のない音声を実現するため、音声波形の接続部分のアクセントやイントネーションを変化させスペクトル分析した。それらのデータから適切な波形を調査した結果を説明する。

2.4　開発環境

本研究のシステム実装で必要な開発環境は、Windows、音声の周波数スペクトル解析ソフトである。

3.　適切な波形の選択

3.1　実験目的と方法

提案する合成手法について図 3 を用いて説明する。同図では通常の合成手法とは異なる部分について網掛けで示している。今回は解析ソフトを利用して音声波形の接続部分のアクセントやイントネーションをどのように変化させスペクトル分析をすれば違和感のない音声を実現することができるか調査する。

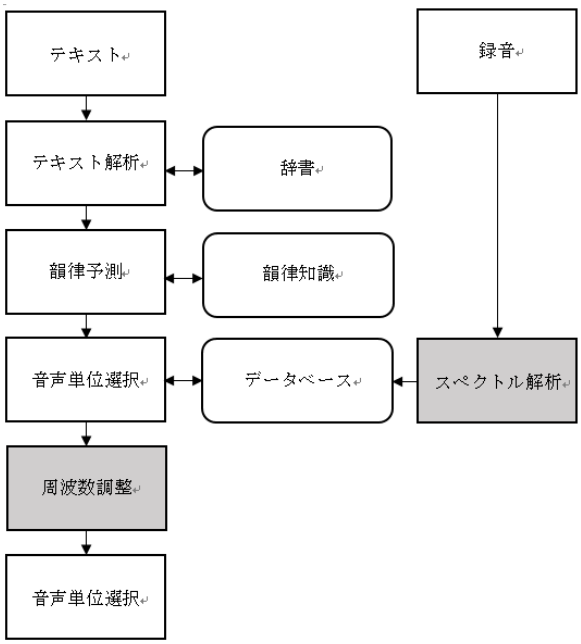

図３．提案手法

まず人間が話した自然音声を使用するため音声を録音する。音声の内容は、文章とその文章を単語で区切った音声と文節で区切った音声の３パターン用意する。パターンの例を図４に示す。

文章：私は大阪に住んでいる。

単語で区切る：私／は／大阪／に／住んで／いる。

文節で区切る：私は／大阪に／住んで／いる。

図４．パターン例

次に３パターンの音声波形をそれぞれスペクトル分析する。文章の音声はそのままスペクトル分析を行い、単語や文節で区切った音声は、区切った箇所を１つずつスペクトル分析する。分析した波形の結果を１つずつ保存し、それらの波形を組み合わせて音声合成を行う。単語で区切った音声は単語で区切った音声のみで合成を行い、文節で区切った音声は文節で区切った音声のみで行う。また単語や文節で区切った音声波形を合成する際にアクセントやイントネーションを考慮し、適切な周波数を設定する。そして適切な周波数を設定した音声波形を単語ごとまたは文節ごと合成を行い、録音した自然音声の文章に近づける。

最後に合成した音声と自然音声をオピニオン評価で比較し、違和感の程度を調査する。録音した自然音声の文章、解析ソフトで合成した単語のみで構成された文章、文節のみで構成された文章の３つを比較する。録音した音声を５段階評価の５として、単語のみで構成された文章と文節のみで構成された文章を比較した場合を５段階評価してもらう方法をとる。

4.　おわりに

今回は、違和感のない音声を実現するため、音声波形の接続部分のアクセントやイントネーションを変化させスペクトル分析し、またそれらのデータから適切な波形を調査した。自然音声にはまだ劣るものの、適切な周波数を設定した波形を合成することで、単語のみで構成された音声合成や文節のみで構成された音声合成でも違和感のない音声を実現することができることがわかった。

参考文献

[1] 音声合成システム「Wizard Voice ™ SDK」
http://www.atr-p.com/products/wv.html

[2] 富士通研究所「音声合成方式の紹介」
http://www.fujitsu.com/jp/group/labs/resources/tech/techguide/list/voice-processing/p04.html

[3]株式会社 AI 「AITalk」
http://www.ai-j.jp/about/