E-026

# 録音した音声から違和感のない合成を行う提案手法
# Suggestion technique realization of synthesis from recorded speech

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、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。

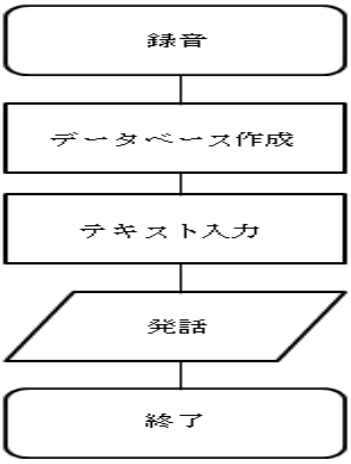


図１　波形接続型音声合成の処理過程

## 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。

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」などに分ける。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。

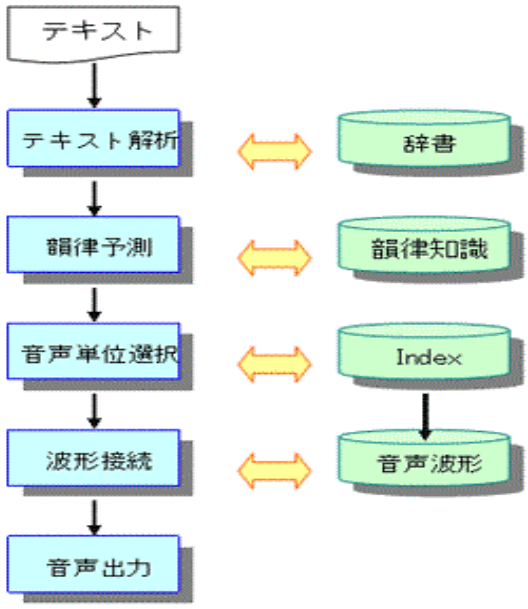


図２　コーパスベース方式の音声合成処理過程

### 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。

† 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



Copyright © 2017 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

## 3. 研究内容

### 3.1 提案手法

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。

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。また、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。そこでデータを補うために、音声の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のではないかと考えた。違和感のない音声を実現するために有効な手法であるかを、音声の周波数スペクトル解析ソフトを利用して実験を行う。

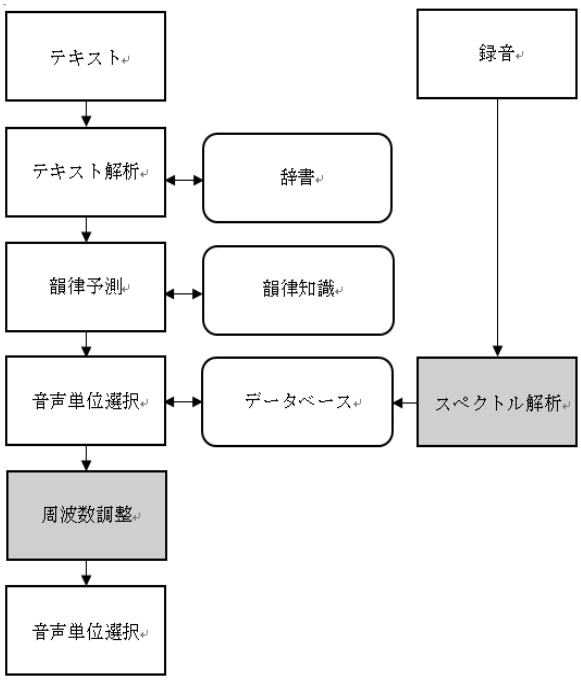


図３．提案手法

### 3.2 実験方法

まず人間が話した自然音声を使用するため、あるゲームの１人のキャラクターのセリフを録音する。自分の音声ではなく、すでに録音された人の音声を利用するため言葉に制限ができるが、その中から音声合成に利用できる音声を選択する。理由として、実際に音声の障碍者から音声サンプルを得る際に、こちらが指定した言葉を正確に発することが出来ないためである。音声の内容は、１１３個の文章と、その文章を単語で区切った音声と、音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。また、録音した文章の音声データと、単語や音節で区切った音声データの時間に制限はなく、音節で区切った音声データの場合、0.5 秒以下のものがほとんどであった。

文章：わたしはかえる（私は帰る）

単語で区切る：わたし／は／かえる

音節で区切る：わ／た／し／は／か／え／る

図４．パターン例

次に、データベース作成のため、３パターンの音声波形を保存し、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、音節で区切った音声は、音節で区切った音声のみで合成を行う。保存の際は、明瞭に発音しているものとされていないものに分ける。

合成した音声データは、波形接続部分のつなぎ目に違和感ができるため、それを修正する必要がある。合成した音声データの波形をそれぞれスペクトル分析し、人間の発話に必要なアクセントやイントネーションを加えるため、波形接続部分に適切な周波数を設定する。また、音節で区切る際に発音が明瞭でない音声データにも、適切な周波数を設定し明瞭に聞こえるようにする。この処理を行うことで、明瞭に発音している音声データとして利用でき、データベースの情報量を増やすことができるのではないかと考えた。それらのデータから任意の文章を合成する。

それらの合成結果を、オピニオン評価によって音声の比較を行う。「録音した元の音声」を５段階評価の５として、「単語のみで合成した音声」と「音節のみで合成した音声」を比較した場合を５段階評価し、合成した音声の違和感の程度を調査する。

## 4. おわりに

今回は、違和感のない音声を実現するため、録音した音声を単語や音節で区切ってそれぞれ合成し、その波形をスペクトル分析し、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加えるために適切な周波数を設定し、それらのデータから任意の文章を合成した。

合成した音声の違和感の程度を調査するためのオピニオン評価はまだ行っていない。結果は出ていないため、本研究の提案手法が有効であるかはまだ検証できてない。

また、制限された言葉の中から音声合成に利用できる単語や音節を抽出する際は、0.5 秒以下の短いサンプルでも適切な周波数を設定することで、明瞭に発音している音声データを得ることができた。しかし、明瞭の判断は主観的に行っているため、音声サンプルの結果にもオピニオン評価を導入することにより、明瞭の正確性を客観的に測ることができるのではないかと考えた。

これらの結果は、データベースから適切な音声波形を探索／選択をする基準となるため、今後、この結果をデータベースの作成時に利用し、音声合成システムの実装を行う。

## 参考文献

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



Copyright © 2017 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.