# 朗読音声を用いた強調レベル推定により抑揚制御を学習する音声合成方式の検討

A Study of Speech Synthesis Schemes that Learn Inflection Control by Emphasis Level Estimation Using Reading Speech

和田　拓海
Takumi Wada

岡山大学 阿部研究室
Abe Laboratory, Okayama University

**概要** 従来の合成音声は，抑揚が小さく自然な音声であるとは言い難い．そこで，本研究では，抑揚が制御された合成音声の生成を実現させるために，合成音声と朗読音声を用いて強調場所の推定を行い，音声合成学習時に強調場所に強調レベルを付加する．このことで，音声合成時に，強調したい任意の場所に強調レベルを付加することで抑揚が制御された音声合成を検討する．

## 1　はじめに

TTS(Text to Speech) は，テキストを入力とし，音声生成を行う音声合成方式である．従来の TTS で生成される音声は，我々人間が話す音声と異なり抑揚の小さい音声になってしまっている．このような音声では，どこの部分を強調して話しているのかが分かりづらい．これにより，聞き間違いやミスリードが生じるおそれがある．

TTS は様々な場所で利用されている．例えば，コールセンターの自動応答，ショッピングモール，駅などでの構内放送である．このような状況での，聞き間違いやミスリードは大きな問題を生じうる可能性がある．このような問題を引き起こさないないためにも強調すべき箇所を強調できる自然な合成音声つまり，抑揚制御ができる TTS が必要であると考える．

そこで，本研究では，強調場所を見つけるために，アクセント句毎に合成音声と朗読音声の基本周波数の比較を行い，推定した強調場所に強調レベルを加え音声合成の学習を行う．これによって，合成音声生成時に強調したい場所を指定することができ，その部分に強調レベルを付加することで，抑揚が制御された合成音声の生成を提案する．

本報告では，提案方式と今後の課題についての検討を行う．

## 2　提案方式

提案方式は大きく分けて (a) と (b) の 2 つの段階に分けられる．これらを以下に示す．

### 2.1　提案方式 (a)　強調場所の推定と強調レベルの抽出

提案方式 (a) の概要図を図 1 に示す．

提案方式 (a) では基本周波数を使用し，朗読音声の強調場所を強調レベルとして抽出する．

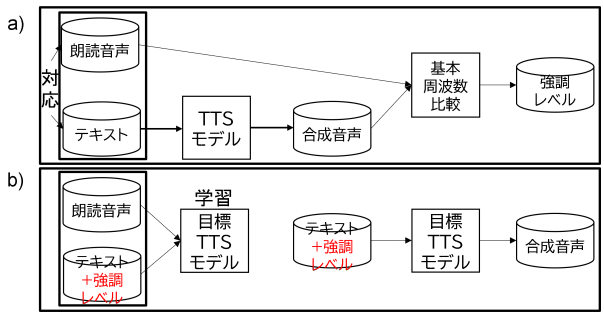


図 1:　提案方式 (a). 朗読音声と合成音声の基本周波数比較により強調レベル抽出．
提案方式 (b). 強調レベルを用いた音声合成モデルの学習と音声合成.

基本周波数は，話者が強調した部分を知覚するための主要な音響情報である [2]．また，強調された部分は基本周波数が高くなることが知られている．

J-KAC コーパスにある朗読音声とこの朗読音声に対応したテキストを用いて TTS で生成した合成音声の二つの音声を比較する．朗読音声と合成音声は長さが同等でないので動的時間伸縮を用いてフレーム数を同等にする．この二つの音声に対して WORLD 分析を行い，基本周波数を抽出する．朗読音声と合成音声は男声と女声なので，比較するために基本周波数のレベル合わせを行う．これは，男声と女声の基本周波数の値が大きく異なり，単純に比較できないからである．その後アクセント句毎に基本周波数を比較し，朗読音声の基本周波数が合成音声の基本周波数よりある閾値分大きい場合，そのアクセント句を強調場所と推定し，強調レベルとして抽出する．

### 2.2　提案方式 (b)　TTS モデルの学習と音声合成

提案方式 (b) の概要図を図 1 に示す．提案方式 (b) では，強調レベルを用いた目標 TTS モデルの学習と抑揚が制御された合成音声の生成を行う．

朗読音声と強調場所に強調レベルを付加したテキストを用いて目標 TTS の学習を行う．この目標 TTS を用いて合成音声を生成する．音声合成時，強調場所を

表 1: J-KAC

| | |
|---|---|
| サンプリング周波数 | 48 kHz |
| 話者 | 男声プロ話者 1 名 |
| 収録時間 | 9 時間 |
| 収録場所 | スタジオ収録 |

表 2: 使用するテキスト

| | |
|---|---|
| # | アクセント句境界 |
| ! | アクセント核 |
| ^ | 立ち上がり |
| ( | 文末 |
| 2 | 句読点などの記号 |
| 大文字 | 無声化 |

手動で選択し，強調場所に強調レベルを付加する．これは，同じ音韻情報でも強調する場所が異なる場合があるからである．例,「私は岡山で公務員をしています」の文の場合，どこを強調するときは「岡山」，何をしているのかを強調するときは「公務員」を強調するのが一般的である．

### 2.3 使用したコーパス

使用したコーパスを表 1 に示す．J-KAC(japanese kamishibai and audiobook corpus)[3][4] はオーディオブックと紙芝居の朗読音声からなる．収録されている作品は著作権の消滅した作品やライセンスフリーの作品である．

## 3 テキスト処理

### 3.1 テキスト形式

本研究で TTS(Text to Speech)[1] で学習，合成時に使用するテキストはかな漢字ではない．学習時，テキストと音声を対応づける．この時，かな漢字より以下のようなアクセント文を使用することでアクセント指定ができ，より自然な合成音声を生成することが可能になる．

本研究では以下のテキストを TTS 学習，合成時に使用する．

```
# k I ^ t a n o # u ! m i n i m o #
s u ! N d e # i ^ t a # n o ^ d e a r i
m a ! s U (
```

上記のテキストは,「北の海にも棲んでいたのであります」をアクセント文に書き換えたものである．

### 3.2 強調レベル

学習時に強調レベルを付加したテキストを用いる．テキストは 3.1 節に示したものであり，2.1 節に示した通りアクセント句毎に基本周波数を比較し，強調場所を推定し，強調レベルを抽出する．テキストでは#がアクセント句境界なので,「北の」,「海にも」,「棲んで」,「いた」,「のであります」の 5 つのアクセント句に分解することができる．朗読音声と合成音声の基本周波数を比較し，強調場所が「北の」だと推定されたとする．この時「北の」に強調レベルを付与する．例えば以下の通りである．

```
# k' I' ^ t' a' n' o' # u ! m i n i m o #
s u ! N d e # i ^ t a # n o ^ d e a r i
m a ! s U (
```

このように強調場所に強調レベルを付与することで，TTS 学習時に強調場所を認識し，合成時に強調場所を指定することで抑揚が制御された音声が合成可能になると考える．

## 4 まとめと今後の課題

本報告では，合成音声と朗読音声の比較をもとにテキストに強調レベルを付加し，声合成時に，強調したい任意の場所に強調レベルを付加することで抑揚が制御された音声合成を可能にする音声合成方式についての提案を行った．これにより同じ音韻情報でもシステム使用者が強調したい場所を指定し，音声合成が行えるようになると考える．今後の課題として，強調レベルの具体的な方法や朗読音声と合成音声の基本周波数の比較方法を検討する．

## 参考文献


[1] Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, Xu Tan, “ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit,” arXiv:1910.10909, Feb. 2020.

[2] 長尾 恭子，天野 成昭，“発話における強調表現の知覚：基本周波数による発話意図の識別，” 電子情報通信学会研究報告，vol.100，no.254, pp47–54, 2000.

[3] 高道 慎之介, 中田 亘, 郡山 知樹, 丹治 尚子, 井島 勇祐, 増村 亮, 猿渡 洋, “J-KAC：日本語オーディオブック・紙芝居朗読音声コーパス，” 情報処理学会研究報告, vol.2021-MUS-131，no.14，pp.1–4, 2021.

[4] 小口 純矢 , 金井 郁也 , 小田 恭央 , 齊藤 剛史, 森勢 将雅，“ITA コーパス：パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価，” 情報処理学会研究報告, vol.2021-MUS-131, no. 31, pp. 1–6, 2021.