局所的な句構造の情報を用いたニューラル音声合成

海木　延佳† ・ サクティ　サクリアニ†‡ ・ 中村　哲†‡

2021/6/19

†奈良先端科学技術大学院大学
‡理化学研究所 革新知能統合研究センター

局所的な句構造の情報を用いたニューラル音声合成

[背景] ニューラル音声合成により合成音の音質の自然性は向上した
しかし、まだ韻律の自然性は不充分

[目的] 構文情報を導入し、発話者の意図を反映した韻律を生成

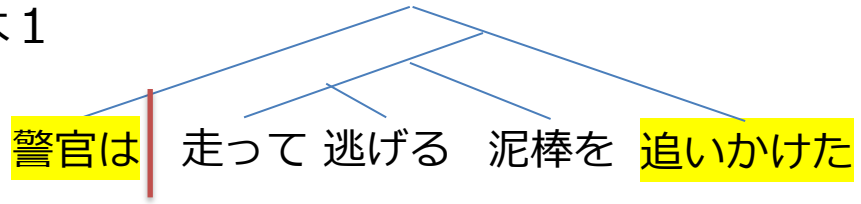
[提案] 句境界に『**局所的な句構造**』の情報を挿入

[結果] 発話者の意図（構文）に従った韻律の生成が可能に
従来法に比べ**自然性が向上**

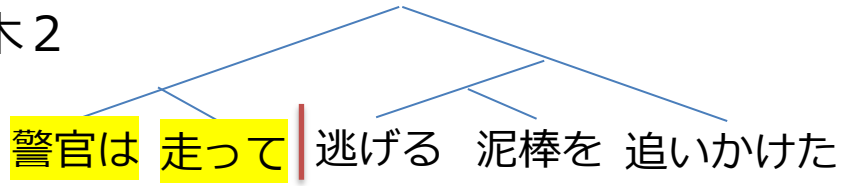
従来法（アクセント情報の導入）：

提案法（局所的な句構造の情報を導入）：

構文木１



構文木２



局所的な句構造の情報を用いたニューラル音声合成

海木　延佳† ・ サクティ　サクリアニ†‡ ・ 中村　哲†‡

2021/6/19

†奈良先端科学技術大学院大学
‡理化学研究所 革新知能統合研究センター

背景

[背景]ニューラル音声合成により合成音の音質の自然性は
大きく向上した
まだ日本語の**韻律の自然性は不充分**と思われる。
近年、アクセント情報の利用により韻律の自然が向上

[目的]ニューラル音声合成に更に、**局所的な句構造の情報**
（構文情報）**を取り入れ**、
発話者の意図を反映した、より自然な韻律を生成する

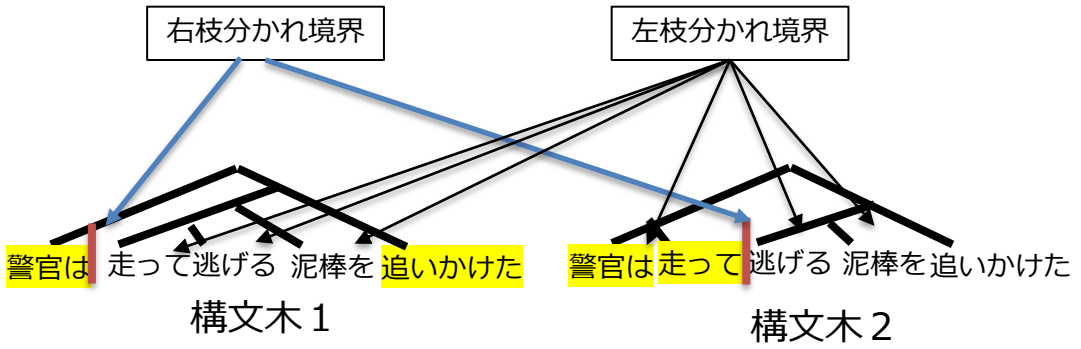
入力:　　警官は走って逃げる泥棒を追いかけた。
音素列:
k e　e k a N w a　h a　sh i　cl t e　n i　g e　r u
d o　r o b o o o　o　i k a k e　t a

アクセント情報の追加:
k e ^ e k a N w a # h a ^ sh i ! cl t e # n i ^ g e ! r u #
d o ^ r o b o o o # o ^ i k a k e ! t a (

表1 アクセント情報を取り入れた韻律記号[1]

	韻律記号
アクセント上昇記号	^
アクセント下降記号	!
アクセント句の区切り	#
文末（通常）	(
文末（疑問）	?
ポーズ	_

（２つの構文木の可能性　⇒　２つの韻律）



[1] K. Kurihara, N. Seiyama, T. Kumano, “Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS,” IEICE Trans. Inf. & Syst., Vol.E104–D, no2, Feb. 2021

２つの提案モデル

２つの提案モデル(局所的な句構造の情報を導入)

提案１：　従来モデルに、アクセント境界に係受けの深さを挿入

提案２：　F0生成過程モデル（藤崎モデル）に基づき、
フレーズ指令とアクセント指令を導入

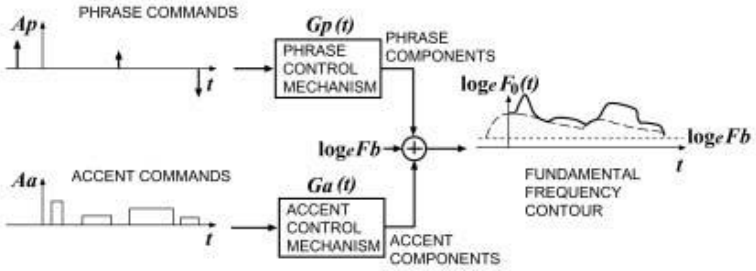


表１　従来モデル[栗原1]：（韻律記号）

	韻律記号
アクセントの立上り	^
アクセント核	!
アクセント句境界	#
文末（通常）	(
文末（上昇）	?
ポーズ	_

表２　提案モデル１：（句構造の情報を追加）

	韻律記号
アクセントの立上り	^
アクセント核(立下り)	!
アクセント句境界の係受けの深さ	**#1, #2, #3, #4, #5, #6**
読点	,

表３　提案モデル２：（F0生成過程モデルに基づく）

	韻律記号
アクセント指令の立上り	/
アクセント指令の立下り	¥
フレーズ指令（係受けの深さ）	**#2, #3, #4, #5, #6**
読点	,

入力:　　**警官は　走って　逃げる　泥棒を　追いかけた**
音素列:　　k e　e k a N w a　h a　sh i　cl t e　n i　g e　r u　d o　r o b o o o　o　i k a k e　t a
従来法:　　k e ^ e k a N w a # h a ^ sh i ! cl t e # n i ^ g e ! r u # d o ^ r o b o o o # o ^ i k a k e ! t a (
提案法１(アクセント情報に句構造の情報を追加):
　構文木1:　k e ^ e k a N w a #4 h a ^ sh i ! cl t e #1 n i ^ g e ! r u #1 d o ^ r o b o o o #1 o ^ i k a k e ! t a
　構文木2:　k e ^ e k a N w a #1 h a ^ sh i ! cl t e #3 n i ^ g e ! r u #1 d o ^ r o b o o o #1 o ^ i k a k e ! t a
提案法２(F0生成過程モデルに基づく):
　構文木1:　k e / e k a N w a ¥ #4 h a / sh i ¥ cl t e　n i / g e ¥ r u　d o / r o b o o o ¥ o / i k a k e ¥ t a
　構文木2:　k e / e k a N w a ¥　h a / sh i ¥ cl t e #3 n i / g e ¥ r u　d o / r o b o o o ¥ o / i k a k e ¥ t a

音声データベース

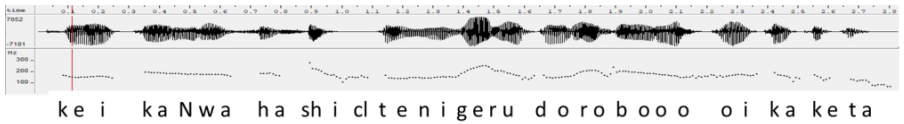
・全11,615文　(Webで公開されている音声・テキスト)
　１話者
　朗読（アラビアンナイト[千夜一夜物語]）
　全26時間 26分

テキスト処理：
　テキスト　⇒　**Open Jtalk**　⇒　音素, アクセント句、形態素
　形態素　⇒　**Chabocha**　⇒　構文木

音声処理：
　文章　⇒　**CTC Segmentation**　⇒　文単位に分割
　文　⇒　**Montreal-Forced-Aligner**　⇒　音素単位に分割

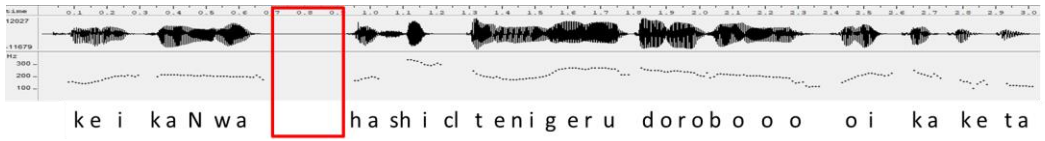
提案法：句境界におけるポーズの生成

従来法:

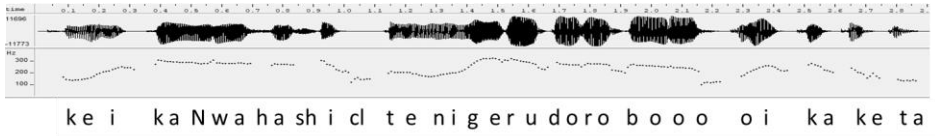


提案法１:

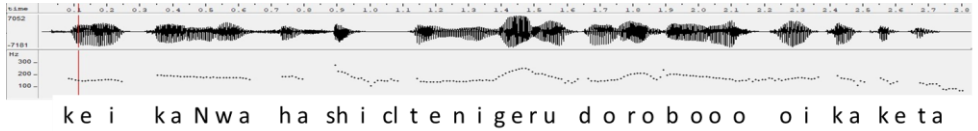
構文木１



構文木２

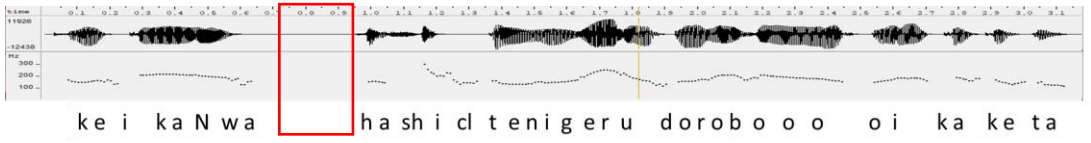


従来法:

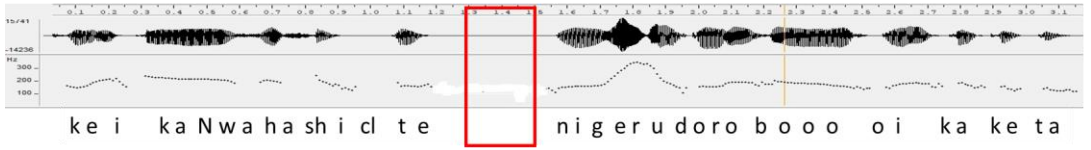


提案法２:

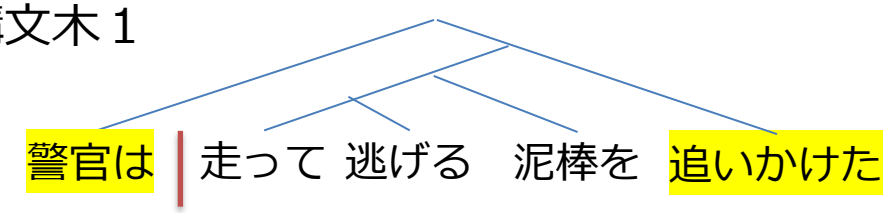
構文木１



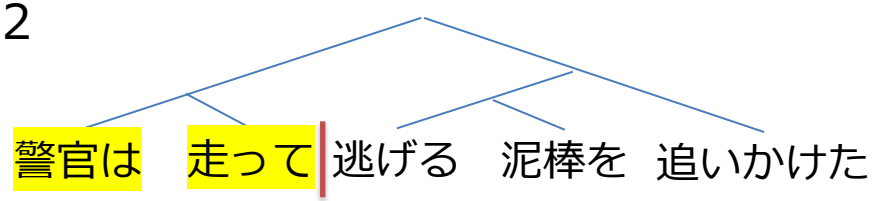
構文木２



構文木１

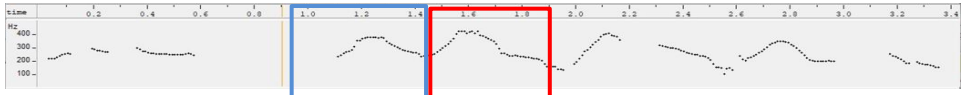


構文木２



提案法：句境界におけるフレーズの立直し

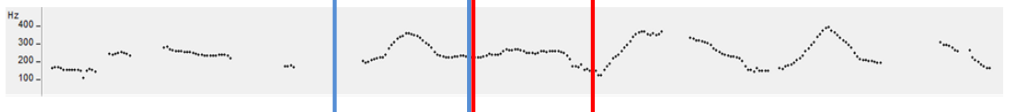
従来法:



提案法１：

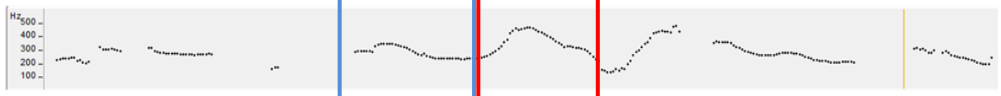
構文木１

隣の句に係る：　前の句より**低い**F0



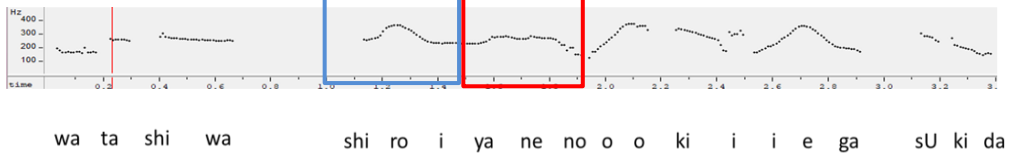
構文木２

隣の句に係らない：　前の句より**高い**F0



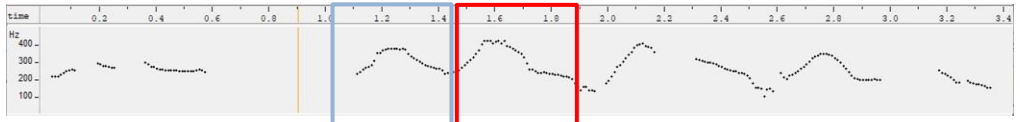
構文木３

隣の句に係る：　前の句より**低い**F0



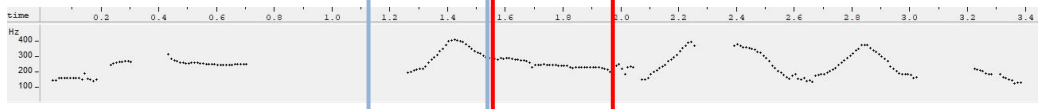
wa ta shi wa shi ro i ya ne no o o ki i i e ga sU ki da

従来法:

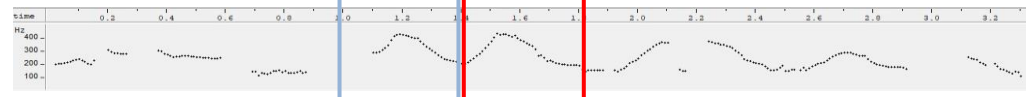


提案法２：

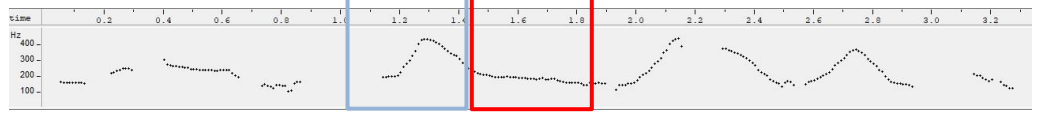
構文木１



構文木２

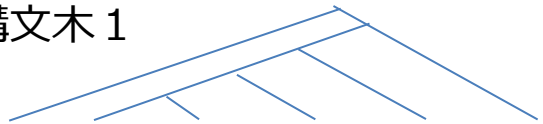


構文木３



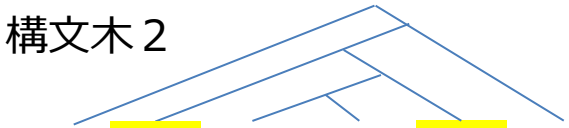
wa ta shi wa shi ro i ya ne no o o ki i i e ga sU ki da

構文木１



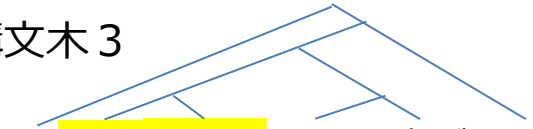
私は、白い 屋根の大きい 家が 好きだ

構文木２



私は、白い 屋根の大きい 家が 好きだ

構文木３



私は、白い 屋根の大きい 家が 好きだ

主観評価

主観評価方法（韻律の自然性）

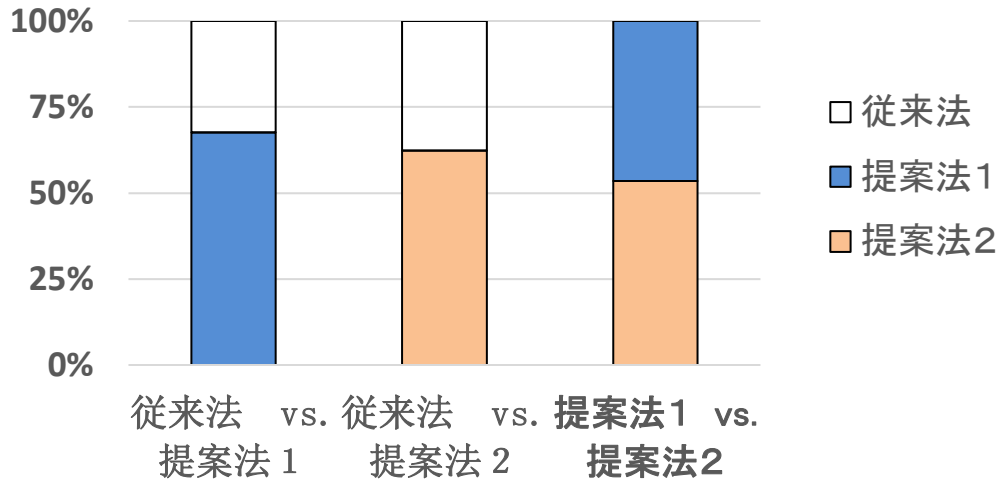
- 13人のネイティブ日本語話者
  １対の合成音声を聞き、どちらの韻律が自然かを判定
- 評価文
  作成した音声DBの評価用250文から20文を選択
- 60対の文音声の比較評価

従来：	提案１	20文＊2
従来：	提案２	20文＊2
提案１：	提案２	20文＊2

主観評価結果

1) 提案した２つの方法で合成した音声は、
従来法の合成音声に比べ、有意に韻律の自然性が高い。
・提案法1 ＞ 従来法　　68% ＞ 32%
・提案法2 ＞ 従来法　　62% ＞ 38%

2) 提案法１と提案法２の間に、合成音声の韻律の自然性に
有意な差はなかった　(有意差 5%)



韻律の自然性の対比較評価結果 (ABテスト)

まとめ

[目的]　ニューラル音声合成に構文情報を取り入れ、
発話者の意図を反映した自然な韻律の生成を目指す

[提案]　『局所的な句構造』 の情報を用いた２つのモデルを提案
１）句境界に『局所的な句構造』の情報を追加挿入する
２）F0生成過程モデルに基づき、フレーズ指令と
アクセント指令を導入

[結果]　発話者の意図（構文）に従った韻律
（ポーズの挿入、F0の立直し）の生成が可能に。
従来法に比べ韻律の自然性が向上。