日本語テキスト音声合成のための句境界予測モデルの検討

二又航介　朴炳宣　山本龍一　橘健太郎
LINE 株式会社

kosuke.futamata@linecorp.com

1　はじめに

句境界はテキスト音声合成システムを構築する上で，音声の自然性に寄与する重要な要因の一つである．句境界とは，連続する句の間に挿入される音声的なポーズとして定義され，息継ぎやアクセントの変化などにより発生する．表 1 に句境界を含まない発話文例と句境界を含む発話文例を示す．

表 1 における’w/o PB’ は句境界を含まない発話文例，’w/ PB’ は句境界を含む発話文例を表す．また，’w/ PB’ における’/’ は句境界を表す．表 1 の’w/o PB’ では，句境界が一切挿入されておらず，聞き手に対して単調で話速の早い発話として印象を与える．一方で表 1 の’w/ PB’ は，複数の句境界が適切に挿入されており一定のリズムで発話されるため，聞き取りやすい印象を与える．このように句境界の有無によって音声の自然性が大きく異なる．

現在までに，LSTM といった深層学習を用いた手法や品詞タグなどの特徴量の効果が，英語を対象とした句境界予測において検証されている [1, 2, 3]．しかし，深層学習と従来の機械学習アルゴリズムを組み合わせた手法 (LSTM+CRF) や BERT[4] など大規模分散表現を用いた手法ついては，句境界予測に対する効果が検証されていない．また，日本語を対象とした句境界予測では，Random forest や CRF などによる手法が研究されているものの [5, 6, 7]，深層学習の有用性については十分に検証されていない．そこで本稿では，日本語テキスト音声合成システムの品質向上のために，様々な特徴量やモデルの句境界予測に対する効果を検証する．

句境界予測にあたり，入力として与えられる発話文のトークン系列から各トークンの後に句境界が挿入されるべきか否か出力する系列ラベリングのタスクとしてモデルを構築した．また，句境界予測に用いたコーパスを分析した結果，非句境界と句境界数のラベル分布が大きく偏っていることが判明した．このようなラベルの分布が偏ったデータに対して，有効である Focal-loss[8] を用いることで句境界予測の精度上昇を図った．自動評価実験および主観評価実験の結果，BERT および Focal-loss により，句境界予測の精度が大きく向上し，かつ音声の自然性が向上することが明らかになった．

2　コーパス及び評価尺度

2.1　句境界コーパスの分析

句境界予測モデルを構築するにあたって，複数話者から構成される日本語話し言葉コーパス (CSJ) [9] の書き起こし文および単一話者から構成される独自に収集したコーパスを使用した．CSJ では，200 ミリ秒以上の無音区間を句境界として認定し，独自コーパスには予め定義された句境界を利用した．また，各コーパスは Sudachi[10] を用いて分かち書きし，トークン数が 128 以下の発話のみ使用した．各コーパスの統計情報を表 2 および表 3 に示す．

表 3 に示す独自コーパスは単一話者から構成され，句境界の挿入位置が予め統制されている．一方で表 2 に示す CSJ は複数話者による少数の発話から構成される．そのため，息継ぎタイミング，思考によるポーズなど句境界が挿入される箇所が話者によって大きく異なる．したがって，CSJ では句境界の分布が話者によって大きく異なる．このようなコーパスから，どのような句構造が自然な音声になりうるか判別することは難しい．しかし，多くの話者に共通する句境界の分布を得ることができれば，万人に共通する自然な音声になると想定される．

2.2　評価尺度

前述の問題を解決するために，自動評価尺度として F-β を使用した．式 1 に F-β の式を示す．

式 1 に示す F-β は F-1 を一般化した評価尺度であり，$\beta \leq 1.0$ として評価することにより，*Recall* よ

表 1　句境界を含まない発話文例と句境界を含む発話文例

w/o PB	先日球場で行われたコンサート後に同球場の天然芝が傷んでしまったと一部で報道された
w/ PB	先日 / 球場で行われた / コンサート後に / 同球場の天然芝が傷んでしまったと / 一部で報道された

表 2　CSJ のコーパス統計情報

	Train	Val	Test
発話数	157,976	1,799	1,729
一文あたりの句境界数	3.54	3.46	3.22
話者数	2463	31	31

表 3　独自コーパスの統計情報

	Train	Val	Test
発話数	99,807	500	500
一文あたりの句境界数	1.59	1.58	1.53
話者数	1		

り *Precision* が重視されて評価されるため，多くの話者に共通する句境界が重点的に学習されることが期待される．また，一般的に誤った箇所に句境界を挿入するより，句境界を挿入しない方が自然な音声になることが知られている [1]．したがって，F-β を用いることで，誤った箇所に句境界が挿入されることが少なくなり，かつ多くの人に共通する句境界が学習されることが期待される．本研究では，経験上 $\beta = 0.25$ として *Precision* を 4 倍偏重して評価を行う．また，単一話者から構成される独自コーパスについては，F-1 を用いて評価を行う．

$$F_\beta = (1+\beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \quad (1)$$

3　自動評価による実験

句境界予測モデルを構築するにあたって，3 つの実験を行った．まず初めに，品詞タグ，構文情報など計 5 種類の特徴量の句境界予測に対する効果を検証した．次に，CRF や BERT などのモデルの効果について検証した．最後に，出力ラベルの分布が大きく偏ったデータに対して有効である Focal-loss の効果について検証した．

3.1　特徴量の比較実験

まず初めに，様々な特徴量の句境界予測における効果を検証した．表 4 に実験に使用した特徴量を示す．表 4 における'DEP' には現トークン，親トークンの構文タグおよび相対位置を用いた．'W2V' は wikipedia による事前学習済みの embedding を表す．実験には，2 層の BiLSTM を使用し，次元数は 512 とした．

実験の結果，CSJ では全ての特徴量において F-0.25 が上昇した．また，'UD', 'POS', 'W2V' の特徴量を用いたモデルでは，'Token' のモデルより 2 ポイント以上 F-0.25 が上昇し，'ALL(POS)' では，F-0.25 の値が 63.5 と最も高い結果となった．一方で，独自コーパスでは単一の特徴量を追加するだけでは精度が上昇しなかった．'ALL(POS)' のみわずかに精度が上昇し，F-1 は値は 90.5 であった．これは，独自コーパスが単一話者による発話，かつ句境界の挿入が箇所が統制されているため，CSJ と比較して句境界の分布が整然としているからであると推察される．特徴量追加による実験結果の詳細は付録の表 12 に示す．

表 4　実験に使用した特徴量

特徴量	次元数	詳細
Token	512	単語トークン
+ UD	16	Universal dependency tags
+ POS	48	品詞タグ (細分類含む)
+ DEP	64	依存構造情報
+ CHAR	64	文字単位 embedding
+ W2V	512	学習済み embedding
ALL (UD)	-	UD 含む全ての特徴量
ALL (POS)	-	POS 含む全ての特徴量

3.2　モデル構造の比較実験

前項の実験では，独自コーパスにおいて特徴量を追加しても句境界予測精度の改善がみられなかった．そこで，前項の実験で用いた BiLSTM に加え，CRF および BERT を用いることにより，句境界予測の精度改善を図った．実験には'BiLSTM', 'BiLSTM+CRF', '$\text{BERT}_{\text{last}}$', '$\text{BERT}_{\text{last}}$ + *CRF*', 'BERT_{mix}', 'BERT_{mix} + CRF' の計 6 つのモデルを対象とした．BiLSTM ベースのモデルには，前項の実験の結果最も性能が高かった'ALL (POS)' を用いた．BERT の学習済モデルには日本語 Wikipedia で事前学習を行った'cl-tohoku/bert-base'[1)] を使用した．$\text{BERT}_{\text{last}}$ は通常の BERT の使い方にならい最終層の出力を利用する．一方 BERT_{mix} では，学習可能なパラメータを用いて各層の出力に対する加重平均

1)　https://github.com/cl-tohoku/bert-japanese

を利用する．BERT は各層において，構文情報や意味情報など異なる特徴量を暗黙的に学習していることが知られているため [11]，それらの特徴量が活用されることを期待する．表 5 および表 6 に CSJ および独自コーパスによる実験結果を示す．

表 5 に示す CSJ の実験では，BERT_{mix} の F-0.25 が一番高く，全ての特徴量を含んだ BiLSTM ベースのモデルより F-0.25 が高かった．また，全ての層における出力を利用した BERT_{mix} において，最終層の出力のみ用いた $\text{BERT}_{\text{last}}$ より F-0.25 が高かった．したがって，BERT の最終層の出力だけではなく各層における出力を用いることで，様々な特徴量を暗黙的に利用し精度が上昇したと推察される．

表 6 に示す独自コーパスによる実験では，BERT_{mix} + CRF の F-1 が一番高く，その他のモデルより優れていた．また，前述の表 5 における CSJ の実験と同様に $\text{BERT}_{\text{last}}$ より BERT_{mix} の F-1 の方が高い結果となった．

以上の結果より，句境界予測に対して，明示的な特徴量を用いた BiLSTM ベースのモデルより，BERT ベースのモデルの性能の方が高いことが明らかになった．BERT ベースのモデルでは，BERT の最終層の出力のみを用いた $\text{BERT}_{\text{last}}$ より，BERT 全層の出力の加重平均を用いた BERT_{mix} の結果の方が高かった．BERT は各層異なる特徴量を暗示的に含むため，全層の出力を用いることによって句境界予測の性能が向上したと推察される．また，$\text{BERT}_{\text{last}}$ の性能は BiLSTM ベースのモデルとほとんど性能に差がなかったため，句境界予測において BERT を用いる際，異なる複数の層における特徴量を利用することが重要であると考えられる．

表 5 CSJ によるモデル構造比較の F-0.25 評価結果

	F-0.25	Precision	Recall
BiLSTM	63.5	68.1	30.7
BiLSTM + CRF	65.2	70.9	28.7
$\text{BERT}_{\text{last}}$	63.9	67.3	35.3
$\text{BERT}_{\text{last}}$ + CRF	64.2	67.2	**37.2**
BERT_{mix}	**67.4**	**72.9**	30.8
BERT_{mix} + CRF	64.0	68.0	33.1

3.3 Focal-loss の効果検証

表 2 および表 3 に示した CSJ および独自コーパスにおける非句境界 (<NB>) と句境界 (
) の分布は大きく偏っている．表 7 に CSJ 及び独自コーパスに含まれる非句境界と句境界の数および比率を示す．

表 6 独自コーパスによるモデル構造比較の F-1 評価結果

	F-1	Precision	Recall
BiLSTM	90.5	91.9	89.0
BiLSTM + CRF	90.1	91.6	88.5
$\text{BERT}_{\text{last}}$	90.8	92.2	89.5
$\text{BERT}_{\text{last}}$ + CRF	91.7	92.9	90.6
BERT_{mix}	92.0	94.1	90.0
BERT_{mix} + CRF	**92.8**	**94.3**	**91.4**

表 7 から，CSJ および独自コーパスにおける非句境界と句境界の比率は大きく偏っており，非句境界の数が圧倒的に多いことがわかる．前節までの実験では，Cross-entropy loss により句境界予測モデルの性能を測ったが，Cross-entropy loss は，非句境界と句境界の損失を同等に扱うため，非句境界に対する loss が多く伝播される傾向にある．そこで，物体検出の分野で一般的に用いられる Focal-loss を導入する．Focal-loss の式を式 2 に示す．

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{2}$$

式 2 における γ および α は調整可能なパラメータを表しており，損失の減衰加減を調整する．これにより，出力確率 p_t の大きいサンプルに対する損失が減衰される．Focal-loss (FL) における γ および α は $\gamma = 2.0$，$\alpha = 0.4$ とし，BERT_{mix} を用いた．また，独自データに関しては，CSJ による学習済みモデルに対して Fine-tuning (FT) した結果も示す．表 8 および表 9 に実験結果を示す．

表 8 および表 9 に示す結果から，Focal-loss を用いることにより，F-0.25 および F-1 が上昇することが明らかになった．また，独自データにおいて CSJ による学習済みモデルを Fine-tuning した結果，わずかに F-1 が上昇した．以上の結果から，Focal-loss を用いることにより，出力確率の高い非句境界に対する loss が減衰され句境界予測における精度が上昇することが明らかになった．

表 7 各コーパスに含まれる句境界の比率

	<NB>	 	<NB> /
CSJ	5,072,106	571,458	**8.875**
独自コーパス	1,185,077	161,072	**7.357**

表 8 CSJ による Focal-loss の効果

	F-0.25	Precision	Recall
BERT_{mix}	67.4	72.9	**30.8**
+ FL	**68.9**	**78.3**	23.5

表 9　独自コーパスによる Focal-loss の効果

	F-1	Precision	Recall
BERT_{mix}	92.0	94.1	90.0
+ FL	93.7	94.6	92.5
+ FL + FT	**94.0**	**94.7**	**93.7**

4　主観評価による実験

句境界予測モデルの導入によって実際に音声の自然性が向上するか調査するために，テキスト音声合成において自然性に関する主観評価実験を行った．

実験に使用したテキスト音声合成システムは，音素列および句境界情報から音響特徴量を予測する音響モデル，音響特徴量から音声波形を生成するボコーダの二つによって構成される．音響モデルには FastSpeech 2 [12]，ボコーダには Parallel WaveGAN [13] を使用した．学習には 5.5 時間からなる単一女性話者の音声コーパスを用いた．詳細なモデル構造及び学習条件は [14] に従った． また，句境界予測モデルには独自コーパスにより訓練されたものを用いた．

主観評価実験では，聴取実験による平均オピニオン評点 (MOS: Mean Opinion Score) テストおよび，ABX テストにより音声の自然性を評価した．それぞれの実験において被験者は健常な聴覚である 25 人の成人日本語母語話者が評価した．MOS テストでは、評価者はランダムに掲示される音声サンプルを聴取し，自然性に関して 5 段階（1: 非常に悪い，2: 悪い，3: 普通，4: 良い，5: 非常に良い）で評価した．ABX テストでは，評価者はランダムに提示される音声サンプルのペアを聴取し，A と B のどちらがより自然であるか，または同じであるか (Neutral) 評価した．評価対象として，読点の後のみを句境界とした’Rule-based’ に加え，’BiLSTM (Token)’，’BiLSTM (All)’，’BERT_{mix}’，’BERT_{mix}+ FL + FT’，正解データである’Reference’ の計 6 条件を用いた．MOS テストでは，テストセット 30 発話に対して 6 条件の計 180 発話を，ABX テストでは，テストセット 30 発話のペアに対して 5 条件の計 300 発話をそれぞれヘッドホン聴取により評価した．表 10 に MOS テストの実験結果，表 11 に ABX テストの実験結果をそれぞれ示す．

表 10 に示す MOS テストの結果では，’Reference’ を除く BERT_{mix}+ FL + FT の評価値が一番高かった．また，’Rule-based’ および BiLSTM ベースのモデル，BiLSTM ベースのモデルと BERT_{mix} + FL + FT の間に 5%水準で有意な差が見られた．しかし，BERT および BERT_{mix}+ FL + FT の間には有意な差が見られなかった．

表 11 に示す ABX テストの結果では，’Rube-based’ より’BiLSTM (Token)’ の評価値が高く，’BiLSTM (All)’ より’BERT_{mix}’ の評価値が高かった．しかし，’BERT_{mix} + FL + FT’ より’BERT_{mix}’ の評価値の方が高かった．これは，両方の BERT ベースモデルの F-1 の値の差がわずかなものであり，ABX テストに用いられたサンプルによっては予測結果にほとんど差がない，または’BERT_{mix}’ の予測結果の方が’Reference’ に近いサンプルが多くみられたからであると推察される．実際に，両モデルを比較した ABX テストの回答では’Neutral’ を選択する割合が他の ABX テストの回答より極端に高かった．

以上の結果から，テキスト音声合成システムに句境界予測モデルを導入することにより，音声の自然性が向上することが明らかになった．また，句境界予測モデルの性能が音声の自然性に大きく関係していることも明らかになった．

表 10　MOS テストの結果 (95%信頼区間)

Model	MOS
Rule-based	3.72 ± 0.07
BiLSTM (Token)	3.89 ± 0.07
BiLSTM (All)	3.86 ± 0.07
BERT_{mix}	3.91 ± 0.07
BERT_{mix}+ FL + FT	**3.95** ± 0.07
Reference	4.06 ± 0.07

表 11　ABX テストの結果

Target A	Target B	A	B	Neutral
Rule-based	BiLSTM (Token)	0.173	**0.455**	0.372
BiLSTM (Token)	BiLSTM (All)	0.200	**0.211**	0.589
BiLSTM (All)	BERT_{mix}	0.215	**0.221**	0.564
BERT_{mix}	BERT_{mix}+ FL + FT	**0.136**	0.112	0.739
BERT_{mix}+ FL + FT	Reference	0.157	**0.260**	0.583

5　おわりに

本稿では，日本語テキスト音声合成のための句境界予測に対して，様々な特徴量，モデル構造の効果について検証した．自動評価実験の結果，BERT を用いた句境界予測モデルを導入することにより，その他の句境界予測モデルより大きく予測精度が上昇した．また，主観評価実験の結果，句境界予測モデルの精度上昇により，音声の自然性も向上することが明らかになった．

参考文献

[1] Viacheslav Klimkov, Adam Nadolski, Alexis Moinet, Bartosz Putrycz, Roberto Barra-Chicote, Thomas Merritt, and Thomas Drugman. Phrase break prediction for long-form reading tts: Exploiting text structure information. In *Proc. Interspeech 2017*, pp. 1064–1068, 2017.

[2] Anandaswarup Vadapalli and Suryakanth V. Gangashetty. An investigation of recurrent neural network architectures using word embeddings for phrase break prediction. In *Interspeech 2016*, pp. 2308–2312, 2016.

[3] Paul Taylor and Alan W. Black. Assigning phrase breaks from part-of-speech sequences. *Computer Speech & Language*, Vol. 12, No. 2, pp. 99 – 117, 1998.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] 博子武藤, 勇祐井島, 昇宮崎, 秀之水野, 澄宇阪内. Analysis and evaluation of factors relating pause location for natural text-to-speech synthesis. *IPSJ Journal*, Vol. 56, No. 3, pp. 993–1002, mar 2015.

[6] Deok-Su NA and Myung-Jin BAE. A variable break prediction method using cart in a japanese text-to-speech system. *IEICE Transactions on Information and Systems*, Vol. E92.D, No. 2, pp. 349–352, 2009.

[7] Shigeru FUJIO, Yoshinori SAGISAKA, and Norio HIGUCHI. Prediction of major phrase boundary location and pause insertion using a stochastic context-free grammar. *The Transactions of the Institute of Electronics,Information and Communication Engineers.*, Vol. 00080, No. 00001, pp. 18–25, jan 1997.

[8] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, Vol. abs/1708.02002, , 2017.

[9] Hanae KOISO and Kikuo MAEKAWA. The corpus of spontaneous japanese : Its design and transcription criteria. *IPSJ SIG Notes*, Vol. 143, pp. 41–48, may 2001.

[10] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a Japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[11] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works, 2020.

[12] Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and high-quality end-to-end text-to-speech. In *Proc. ICLR (in press)*, 2021.

[13] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. ICASSP*, pp. 6199–6203, 2020.

[14] Ryuichi Yamamoto, Eunwoo Song, Min-Jae Hwang, and Jae-Min Kim. Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators, 2020.

A 付録

特徴量追加による実験結果の詳細を表 12 に示す．

表 12 CSJ および独自コーパスによる特徴量追加の実験結果

特徴量	CSJ			独自コーパス		
	F-0.25	Precision	Recall	F-1	Precision	Recall
Token	58.5	63.9	24.8	88.9	93.0	85.1
+ UD	61.6	65.8	30.4	89.0	90.8	87.2
+ POS	61.6	65.8	30.5	88.6	90.5	86.8
+ DEP	59.3	64.0	27.1	88.7	91.4	86.0
+ CHAR	60.1	65.0	27.1	89.4	92.1	86.8
+ W2V	60.9	66.9	24.9	87.6	91.4	84.1
ALL (UD)	62.9	67.5	30.3	89.3	91.4	87.4
ALL (POS)	**63.5**	**68.1**	**30.7**	**90.5**	**91.6**	**89.0**

主観評価に利用した発話文例を表 13 に示す．各発話文における’
’ は句境界を表す．各発話文は Github 上のページ[2)]から再生できる．

表 13 主観評価に利用した発話文例

Example(1)	
Rule-based	メキシコでは麻薬密売組織に殺害された人の切断遺体が 道路際に置き去りにされることが多い
BiLSTM (Token)	メキシコでは ** ** 麻薬密売組織に殺害された人の切断遺体が 道路際に ** ** 置き去りにされることが多い
BiLSTM (All)	メキシコでは ** ** 麻薬密売組織に殺害された人の切断遺体が 道路際に置き去りにされることが多い
BERT_{mix}	メキシコでは ** ** 麻薬密売組織に殺害された ** ** 人の切断遺体が ** ** 道路際に置き去りにされることが多い
BERT_{mix}+ FL + FT	メキシコでは ** ** 麻薬密売組織に殺害された人の切断遺体が ** ** 道路際に置き去りにされることが多い
Reference	メキシコでは ** ** 麻薬密売組織に殺害された人の ** ** 切断遺体が ** ** 道路際に置き去りにされることが多い

Example(2)	
Rule-based	オリジナルのモデルは 2007 年 1 月に、** **最新モデルの iPhone8 iPhone8 Plus iPhone10 は 2017 年 9 月 12 日に発表されました
BiLSTM (Token)	オリジナルのモデルは ** ** 2007 年 1 月に、 ** ** 最新モデルの ** ** iPhone8 ** ** iPhone8 Plus ** ** iPhone10 は ** ** 2017 年 ** ** 9 月 12 日に発表されました
BiLSTM (All)	オリジナルのモデルは ** ** 2007 年 1 月に、 ** ** 最新モデルの ** ** iPhone8 ** ** iPhone8 Plus ** ** iPhone10 は ** ** 2017 年 ** ** 9 月 12 日に発表されました
BERT_{mix}	オリジナルのモデルは ** ** 2007 年 1 月に、 ** ** 最新モデルの ** ** iPhone8 ** ** iPhone8 Plus ** ** iPhone10 は ** ** 2017 年 ** ** 9 月 12 日に発表されました
BERT_{mix}+ FL + FT	オリジナルのモデルは ** ** 2007 年 1 月に、 ** ** 最新モデルの ** ** iPhone8 ** ** iPhone8 Plus ** ** iPhone10 は ** ** 2017 年 ** ** 9 月 12 日に発表されました
Reference	オリジナルのモデルは ** ** 2007 年 1 月に、 ** ** 最新モデルの ** ** iPhone8 ** ** iPhone8 Plus ** ** iPhone10 は ** ** 2017 年 ** ** 9 月 12 日に発表されました

2) https://github.com/matasuke/nlp2021-pbp-audio-examples