本審査

Rakugo Speech Synthesis: Toward Speech Synthesis That Entertains Audiences

加藤集平

2021-02-01

序論

研究の背景・動機

機械は人間と同じくらい自然に文章を読み上げられるか？ Yes!

- 最新の音声合成システムでは，自然音声に極めて近い音声を合成可能.
  - 少なくとも，英語の読み上げ音声の場合.
- もちろん，様々な発話スタイル（感情など）を合成する研究も盛ん.

研究の背景・動機

機械の音声合成能力は十分か？ No!

- 例えば，**人を楽しませる能力は不十分**.
  - 落語を含む話芸を音声合成に代替させるのは，現状では想像し難い.
  - 人間による相当な調整を経た作品でさえ，出来映えは不十分.
- 機械と人間との間の，このような隔たりを埋めたい.
- **人を楽しませる音声合成**としての，落語音声合成を行う.

解決すべき課題 1

音声合成に適した落語音声データベースが存在しない.

- 落語のCDやDVDは数千種類が市販されている.
- 多くはライブ録音で，雑音や残響が多く音声合成には不適.
- **独自に音声データベースを構築する必要がある.**

解決すべき課題 2

落語音声のスタイルは非常に多様である.

- 落語の本編は基本的に登場人物の会話から成り立っている.
  → **発話スタイルが多様.**

- 落語は演者がアドリブで，あるいは記憶を頼りに発話している.
  → **はっきり発音しているとは限らない.**

- 登場人物は，その性別・年齢・身分などによって異なる日本語を話す．→ **通常のテキスト処理が困難.**

- このような特徴により，通常の音声に比べて，**モデルの設計・学習がより困難**である.

解決すべき課題 3

音声を聞いて，容易に登場人物の区別が付き，かつ内容が理解されるべきである.

- 落語の本編は基本的に登場人物の会話から成り立っている.
- 当然，**登場人物の区別が付くべき**である.
- （楽しんでもらうためには）**噺の内容が容易に理解されるべき**である.

解決すべき課題 4

合成された落語音声は，聞き手を楽しませるべきである．

- 人を楽しませる音声合成として落語音声合成を開発している以上は，聞き手が楽しめるような音声合成にするべきであるし，どの程度楽しんだかを測定すべきである．

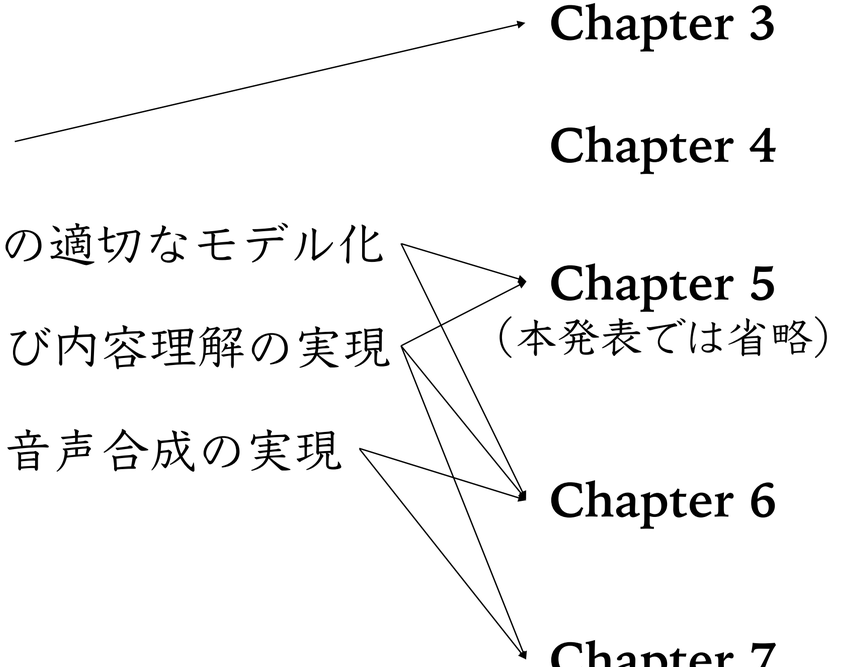
論文の構成

課題1: データベースの構築

課題2: 多様な発話スタイルの適切なモデル化

課題3: 容易な役の区別および内容理解の実現

課題4: 聞き手を楽しませる音声合成の実現



Chapter 1	Introduction
Chapter 2	Rakugo
Chapter 3	Database
Chapter 4	Speech Synthesis and Its Relationship to Entertainment
Chapter 5 （本発表では省略）	Initial Modeling of Rakugo Speech Using Sequence-to-Sequence Speech Synthesis
Chapter 6	Modeling of Rakugo Speech Using Tacotron 2 with Self-Attention, Global Style Tokens, and Manually Labeled Context Features
Chapter 7	Comparison with Human Professionals
Chapter 8	Conclusion

Chapter 2
落語について

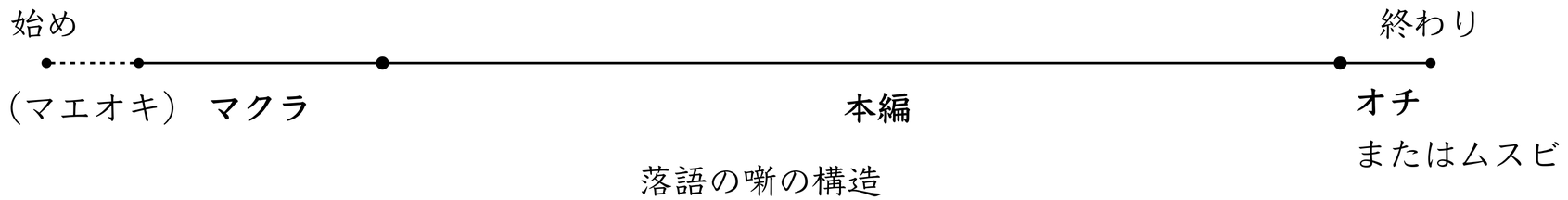
落語とは

- 300年以上の歴史を持つ伝統芸能の一つ.
- 江戸落語（東京落語）と上方落語が存在.
  - 本研究では，**江戸落語**を扱う.
    - 上方落語と異なり，見台・拍子木を使わない，つまりほとんど声だけで表現されることが理由.
- 古典落語（1920年代までに成立）と新作落語（1930年代以降に創作）に分けられる.
  - 本研究では，**古典落語**を扱う.
    - 正統性，共通の演目での比較が容易，著作権の関係などが理由.

東京を代表する4つの寄席（定席）．左上: 鈴本演芸場，右上: 末広亭，左下: 浅草演芸ホール，右下: 池袋演芸場.

噺の構造

- 噺は大きく分けて5つのパートからなる: マエオキ，**マクラ**，**本編**，**オチ**，ムスビ (野村, 1994)．
- マクラ（枕）は本編の導入部分で，即興の話である．
- オチ（落ち）は噺を締めくくる結末の部分．落ちのある話だから，落語という．
  - ムスビはオチのない噺か，噺を途中で打ち切る際に用いられる．
- 噺の長さは通常10分から45分程度（様々）．
  - 寄席（主に落語が上演される常設劇場）での持ち時間は，通常一人15分（主任（トリ）のみ30分）



落語の噺の構造

落語の演じ方

- 演者はステージ（高座）に置かれた座布団の上に座り，一人で演じる．
- 小道具は扇子と手拭いのみ（江戸落語の場合）．
- 演者は**複数の登場人物を一人で演じ分け，主に登場人物の会話だけでストーリーが進んでいく**．

落語を演じる春風亭正太郎．

落語の言葉遣い

- 古典落語の本編で用いられる日本語は，**やや古い**．
- 登場人物の**性別・年齢・身分などによって言葉遣いが異なる**．
  - 例（「青菜」より）

隠居（商人）	エー植木屋さんは、菜のお浸しはお好きですか。
植木屋（職人）	ああお浸しえええええ、もうあっしはね、でえ好きなんですよ。

  - 逆に言えば，言葉遣いによって性別・年齢・身分などを表現している．

江戸落語の身分制度

- 江戸落語の噺家（落語家）には，下から順に**前座**・**ニツ目**・**真打**という身分制度がある．
  - **前座**（期間は3–5年）：寄席の楽屋仕事（雑務一般）や師匠の世話をしながら稽古に励む．寄席では文字通り前座として落語を披露することがある．プロとはみなされない．
  - **ニツ目**（期間は10年程度）：ここからがプロだが，寄席で主任を務めることはなく，弟子も取れない．
  - **真打**（終身）：寄席で主任を務めることができるほか，弟子も取れる．

Chapter 3
データベース

NII落語音声データベース

- 適当な落語音声データベースが存在しなかったため，データベースを独自に構築した.
  - Database I: モデル構築・評価用
  - Database II: 評価専用
- 音声収録用の防音スタジオで収録した.
- 録音時にスタジオ内にいたのは演者一人で，観客および反応は一切なかった.
- 言い間違いや言い直しなどは，演者本人が希望した場合を除いて再録音を行わずそのまま収録した.

収録の様子

Database I（モデル構築・評価用）

演者

- 柳家三三（真打）

録音した演目

- 江戸落語の古典落語25演目

アノテーション

- 書き起こし（音素）を行うとともに，文ごとにコンテキストラベルを付与した.

Database II（評価専用）

演者

- 柳家三三（真打），柳亭市童（二ツ目），柳家小ごと（前座）

録音した演目

- 演目「味噌豆」

アノテーション

- 行っていない.

Database I の書き起こし

- 私が書き起こしを行った.
- **フィラーや笑い声などに対しても特別な表記は定義せず，聞こえたとおりに音素を書き起こした.**
- 文は以下のような基準で区切った.
  - 文法的に文の切れ目であり，かつ後にポーズが続く箇所.
  - 話者交代が起きている箇所.
  - 上昇調のイントネーションの直後.

書き起こしに使用した音素の一覧

音素	母音	a, e, i, o, u
	子音	b, by, ch, d, dy, f, fy, g, gw, gy, h, hy, j, k, kw, ky, m, my, n, N, ng, ny, p, py, r, ry, s, sh, t, ts, ty, v, w, y, z
	その他	cl（促音）
ポーズ		pau（読点）, sil（文の始めおよび句点）, qsil（疑問符）

Database I の文ごとに付与した コンテキストラベルの一覧

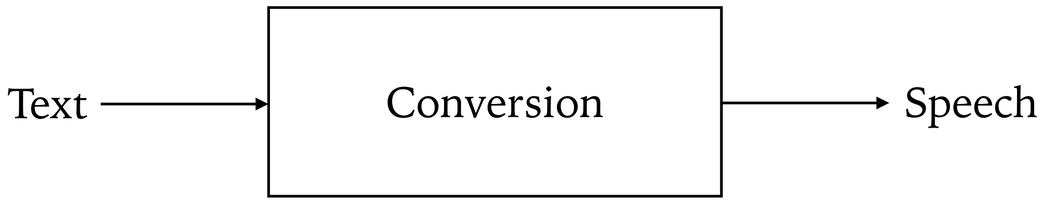
グループ	名前	説明	詳細
ATTRibution 発話者の属性	role	登場人物の役	［性別］噺家，男，女 ［年齢］噺家，子供，若者，壮年，老人 ［身分］噺家，武士，職人，商人，その他町人，田舎者，その他方言，現代人，その他
	individuality	登場人物の個性	噺家，間抜け
CONDition 発話者の状態	condition	登場人物の状態	平常，感心している，諫めている，気取っている，怒っている，懇願している，ゴマをすっている，陽気である，不満である，自信がある，困惑している，納得している，泣いている，憂鬱である，飲んでいる，酔っ払っている，食べている，励ましている，興奮している，怖がっている，不審がっている，体調不良である，眠たい，心苦しく思っている，疑念を持っている，拍子抜けしている，寒がっている，苛立っている，幽霊のようである，喜んでいる，もじもじしている，興味を持っている，正当化している，掛け声をかけている，大声で話している，笑っている，甘えている，説教している，軽蔑している，焦っている，ペットに向かって話している，とぼけている，我慢している，反抗している，拒否している，悲しんでいる，誘惑している，呆れている，叫んでいる，小声で話している，なだめている，力んでいる，驚いている，得意になっている，からかっている，叱っている，息が切れている，思い出そうとしている，高を括っている，不快である（60種類）
SITuation 発話者の置かれた状況	relationship	話し相手との関係性	噺家，地の文，独り言，目上，目下
	n_companion	話し相手の数	噺家，地の文，独り言，1人，2人以上
	distance	話し相手との距離	噺家，近，中，遠
STRucture 噺の構造	part	パート	マクラ，本編，オチ

注:「噺家」はマクラにおいてどの登場人物にも該当しない文に相当する.

Chapter 4
音声合成について

音声合成

- 音声合成とは: 機械が音声を合成する過程のこと.
- 多くの場合，文字列を音声に変換する (text-to-speech; TTS).
- 自動音声応答装置 (IVR)，カーナビゲーションシステム，音声アシスタント，本の読み上げ，公共交通機関のアナウンスなどで幅広く実用化.



パイプライン／frame-by-frame方式

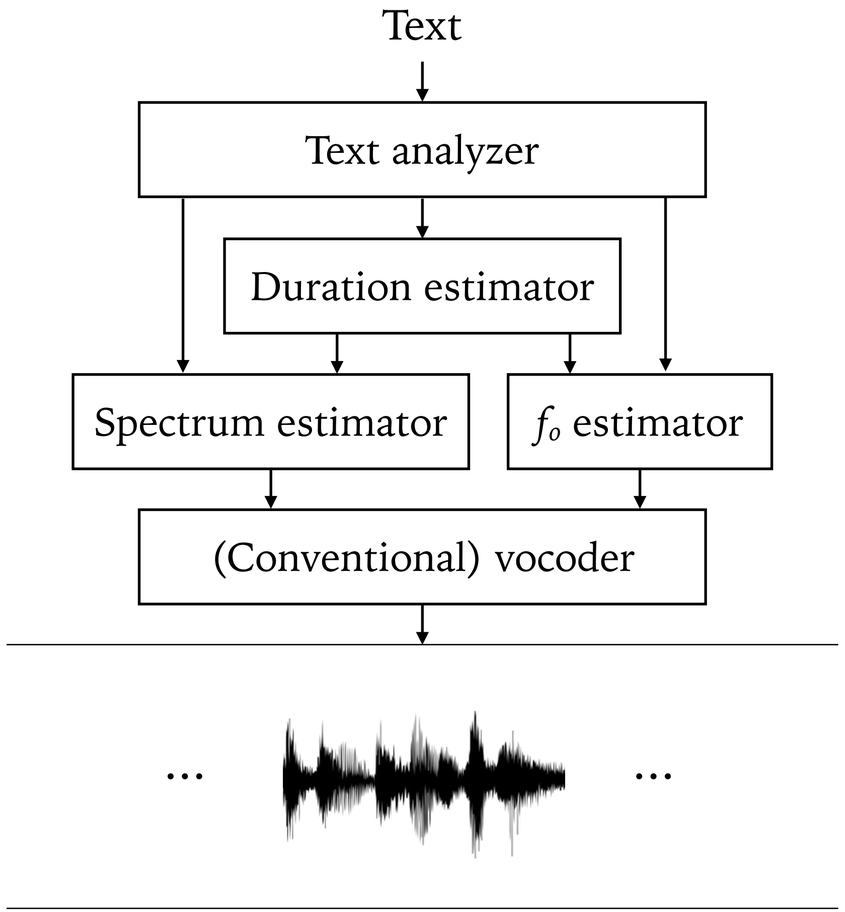
- 入力された文字列を直接出力波形に変換するのは困難.
- 順番に，少しずつ変換していく方式.
- フレームごとに音響特徴量を予測していく.

利点

- モジュール化されているので，デバッグがしやすい.

欠点

- 合成された音声の自然性に改善の余地がある.
- **標準的でないテキストには事実上対応できない.**
  - 方言，古語など.



End-to-end/sequence-to-sequence方式

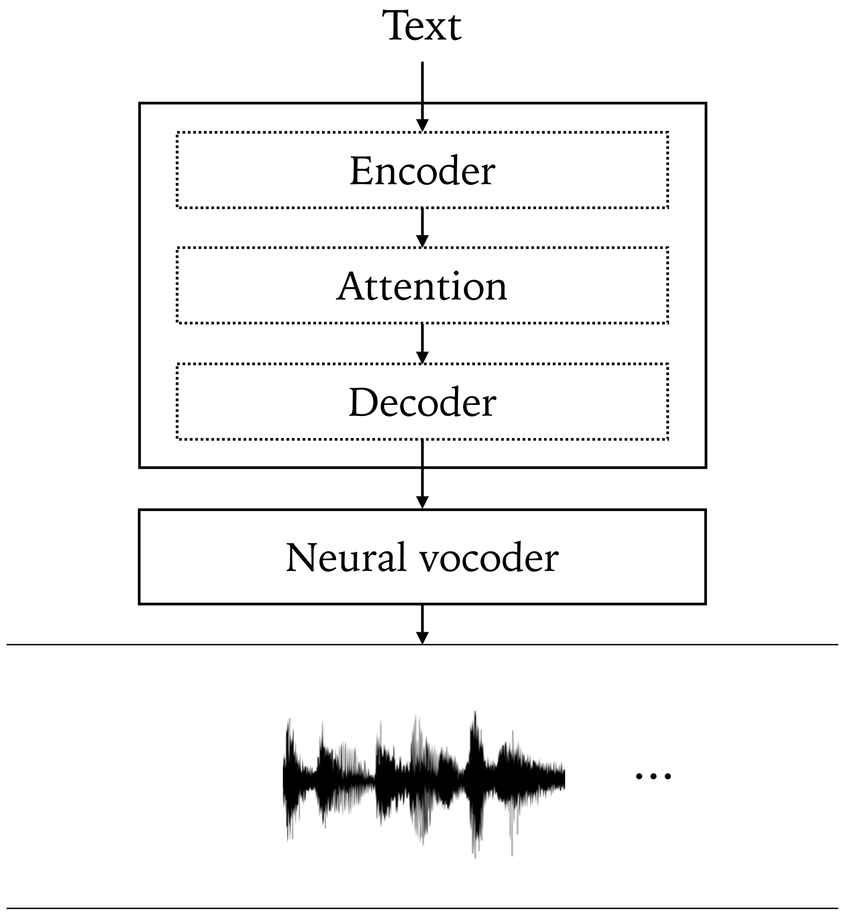
- 入力文字列から出力波形への直接変換を目指して，モジュールの数をなるべく減らした方式.

利点

- 人間と同程度の自然性を持つ音声を合成可能.
  - 少なくとも英語の読み上げ音声の場合.
- **標準的でないテキストにも対応可能.**
  - 日本語の場合は現状では結局音素や仮名の入力が必要だが，それらの情報だけでも比較的高品質な音声が合成可能（本研究では，アクセント情報も入力していない）.
    - 一般的に，パイプライン／frame-by-frame方式では，テキスト解析器は読み（音素）のほかに，アクセント，句境界情報，品詞などの情報を出力する.

欠点

- デバッグがしづらい.



End-to-end/sequence-to-sequence方式が落語音声のモデリングに向いている理由

テキスト解析が困難

- 落語は相当程度くだけた話し言葉を使う上に，古典落語では言い回しもやや古めかしい.

- そのため，読みの推定・アクセントの推定ともに困難.

基本周波数の抽出が困難

- 表現がかなり豊かであるために，基本周波数（音の高さ）の抽出が困難.

Chapter 6

Tacotron 2，self-attention，GSTs，およびコンテキスト特徴量を用いた落語音声のモデル化

本章は論文誌 *IEEE Access* に掲載された論文の内容に基づいている．

Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki, and Junichi Yamagishi, “Modeling of rakugo speech and its limitations: Toward speech synthesis that entertains audiences,” *IEEE Access*, vol. 8, pp. 138149–138161, Jul 27 2020.

本章の動機

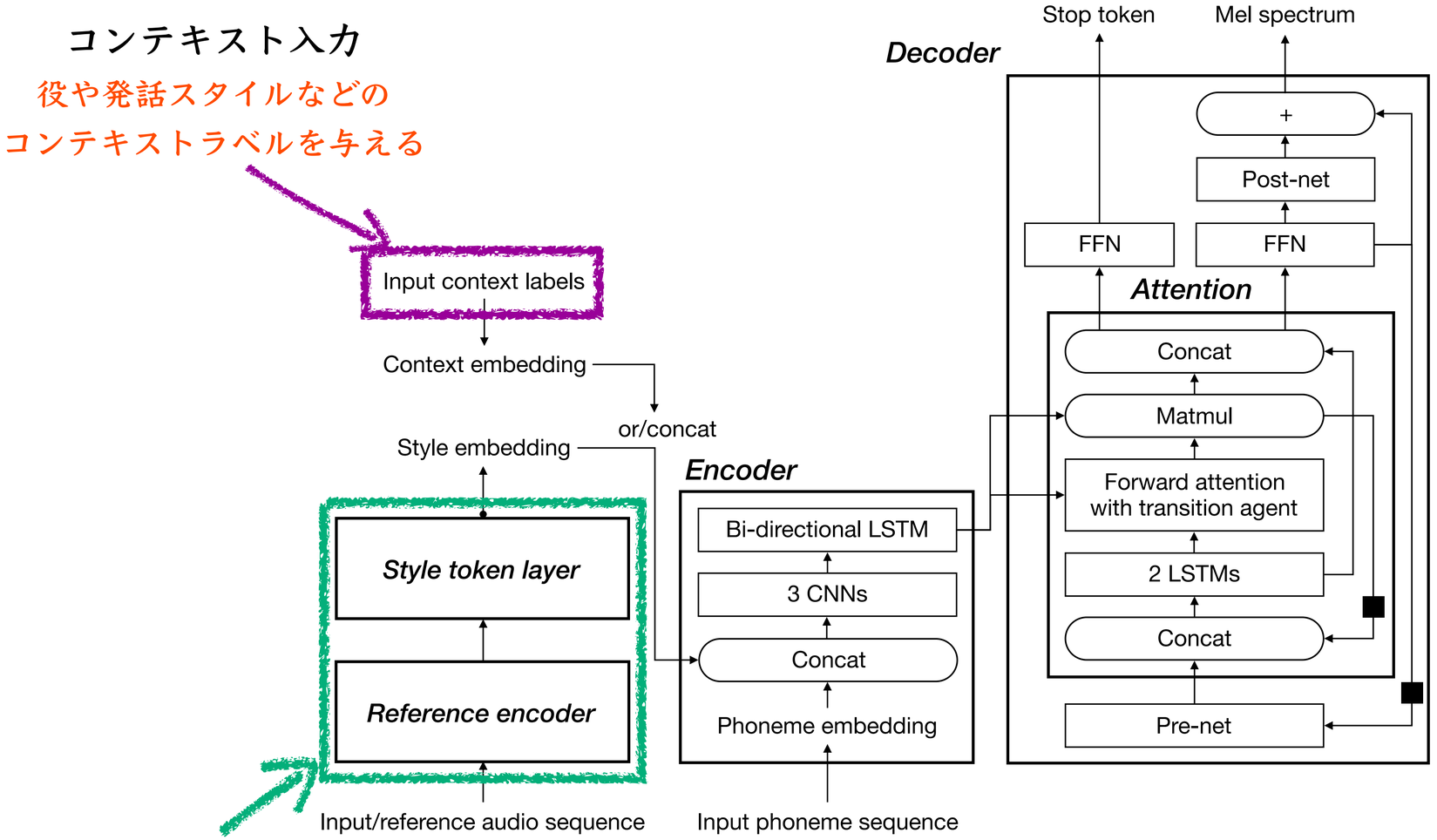
- 読み上げ音声合成で最高性能を誇るTacotron 2と，それをself-attentionで拡張したモデルを使用することで，どこまで落語音声をモデル化できるかを測定したい.

- 「聞き手をどれだけ楽しませたか」を測定したい.

Tacotron 2に基づく落語音声合成モデル

コンテキスト入力

役や発話スタイルなどの
コンテキストラベルを与える



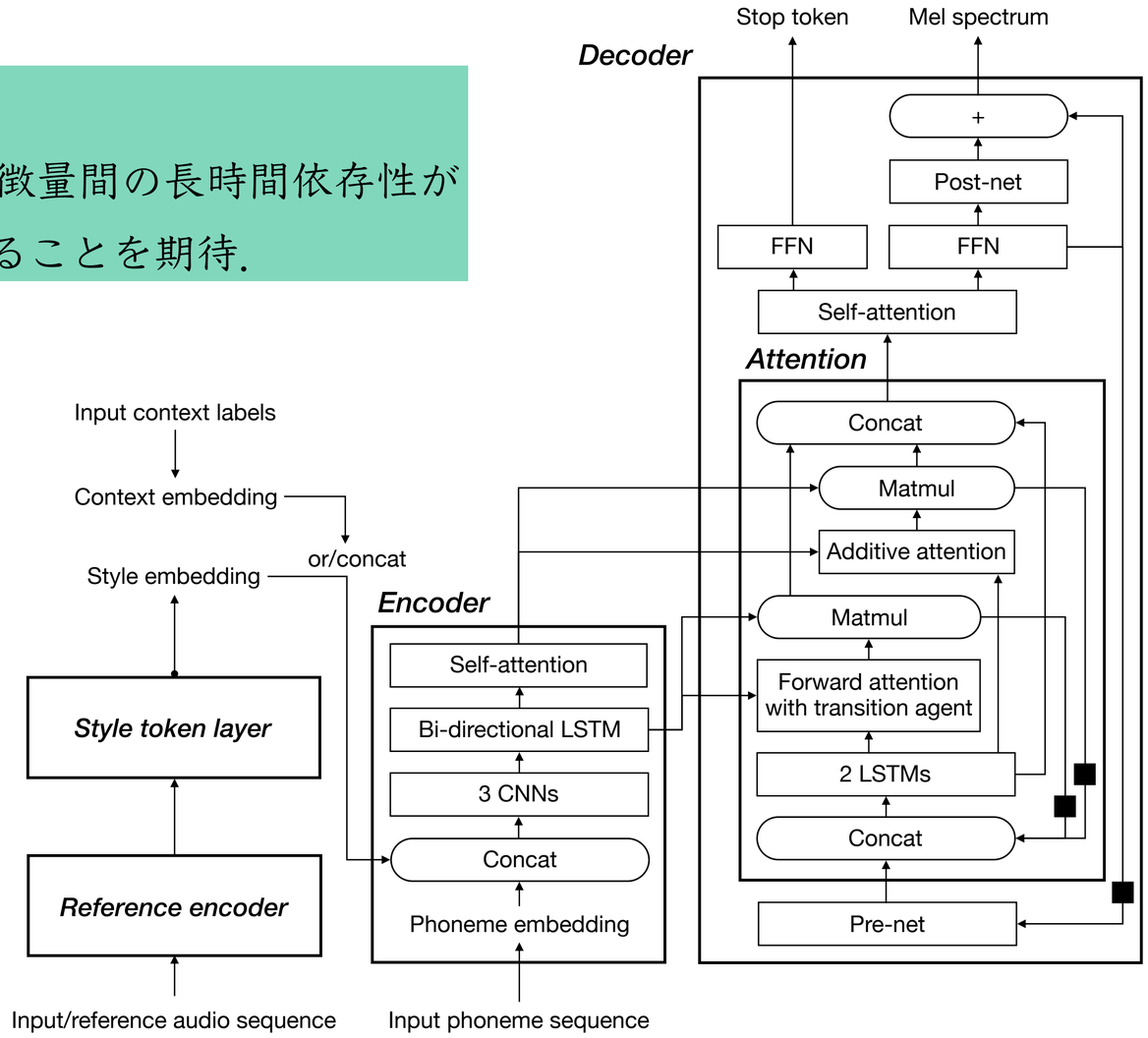
Global style tokens (Wang et al., 2018)

役や発話スタイルなどに係る音声表現について，
参照音声からモデルに推定させる

Self-attentionを用いた拡張 (SA-Tacotron)

Yasuda *et al.*, 2019に着想.

Self-attentionにより，特徴量間の長時間依存性がより適切にモデル化されることを期待.



聴取実験の目的

- 合成された落語音声は，どの程度自然で，役の区別が付いて，内容が理解できて，**聞き手を楽しませる**のか？
- **分析合成音と比べて，どうか？**

聴取実験の条件

データ	Database Iより16演目（7,341文，4.31時間）．ただし，0.5秒未満および20秒以上の音声を除いている．sv56により，全体の音量を-26dBovに正規化．
サンプリング周波数 / ビット / チャネル	48kHz / 16bit / mono
学習セット	6,437文（3.74時間）
検証セット	715文（0.40時間）
テストセット	189文（0.17時間）
音響特徴量	80次元のメルスペクトログラム．ただし，学習・検証・テストセットをあわせた全体で次元ごとに平均が0，分散が1になるように標準化した．
ボコーダ	WaveNet vocoder（学習・検証・テストセットの全てを用いて学習） 入力：メルスペクトログラム 出力：16kHz / 16bit mono waveform

Tacotron 2に基づく落語音声合成モデル

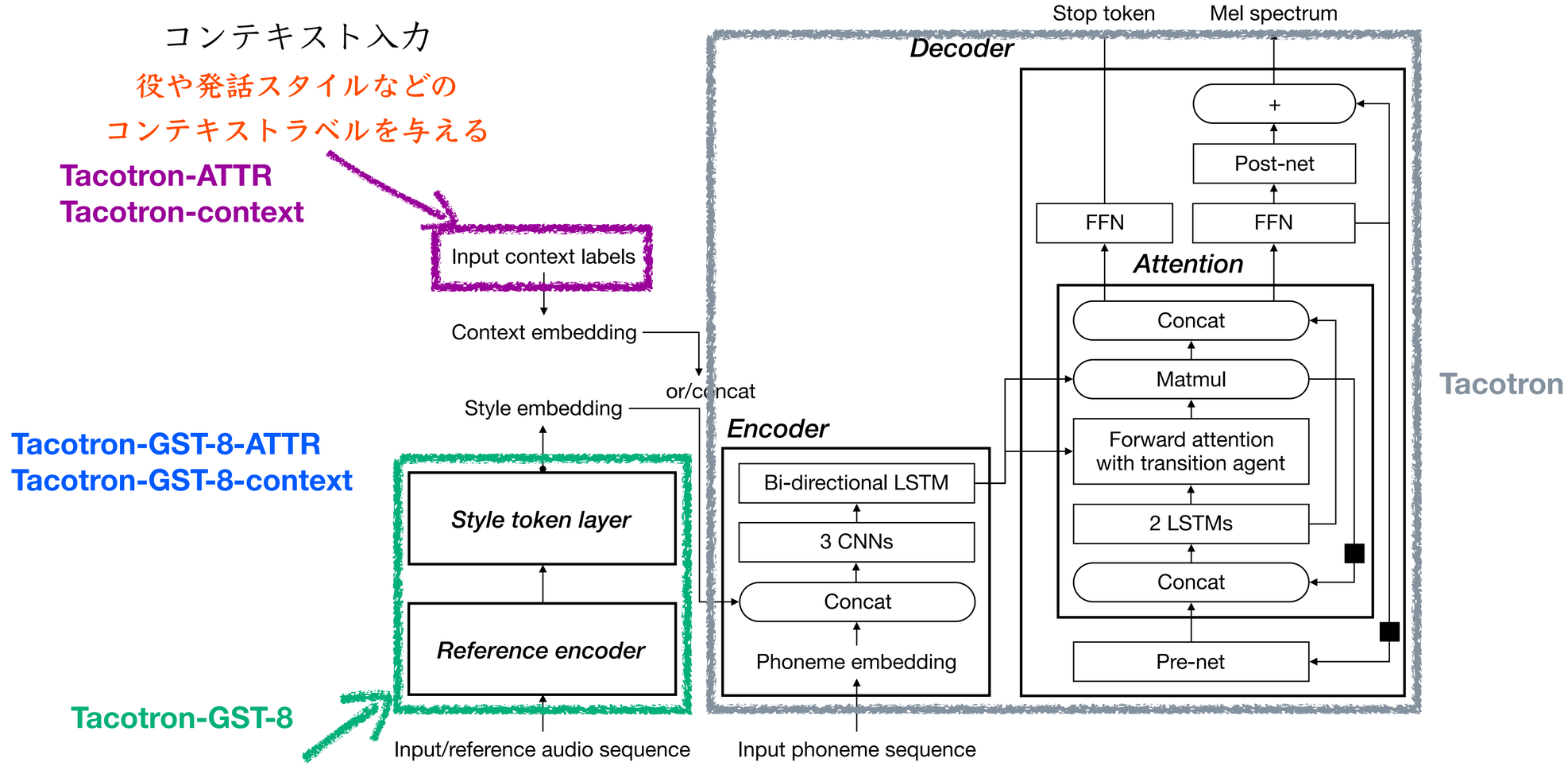
コンテキスト入力

役や発話スタイルなどの
コンテキストラベルを与える

Tacotron-ATTR
Tacotron-context

Tacotron-GST-8-ATTR
Tacotron-GST-8-context

Tacotron-GST-8



Global style tokens (Wang et al., 2018)

役や発話スタイルなどに係る音声表現について，
参照音声からモデルに推定させる

聴取実験の条件

- テストセット189文からなる13の小噺の音声を用意した.
  - 文単位で合成した音声を連結して作成し，文と文の間のポーズは，録音時と同じだけの長さの無音を挿入した.
- 聴取者は文単位ではなく，小噺単位で評価した.
- 分析合成音 (analysis-by-synthesis; AbS) も評価対象とした.
- 5段階のMOS試験を実施.
  - 質問項目は，1) 自然性，2) どの程度役が区別できたか，3) どの程度内容が理解できたか，4) どの程度楽しめたか.
  - 1つの評価ラウンドでは，13の小噺（異なるいずれか1つのシステムで合成）をそれぞれ評価した.
  - 音声（および合成システム）の提示順はランダムとした.
  - 合計189人が189ラウンドを評価した.

得点の標準化

- より公平に比較を行うため，以下の2段階の手順でMOS試験の得点の標準化を行った.

1. **聴取者内の得点の平均が 0，標準偏差が 1 になるように標準化した.**

   - 聴取者の違いによる評価のばらつきを吸収するため.

2. **さらに，システムごとに，分析合成音に対する評価の平均が 0，標準偏差が 1 になるように標準化した.**

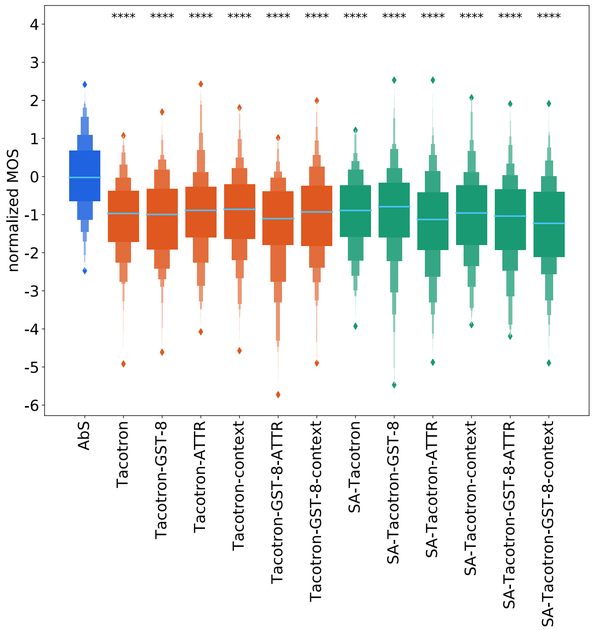
   - 内容が面白い話は得点が高くなり，面白くない話は得点が低くなるので，システムの評価でなく，話の内容の評価となってしまうことを防ぐため.

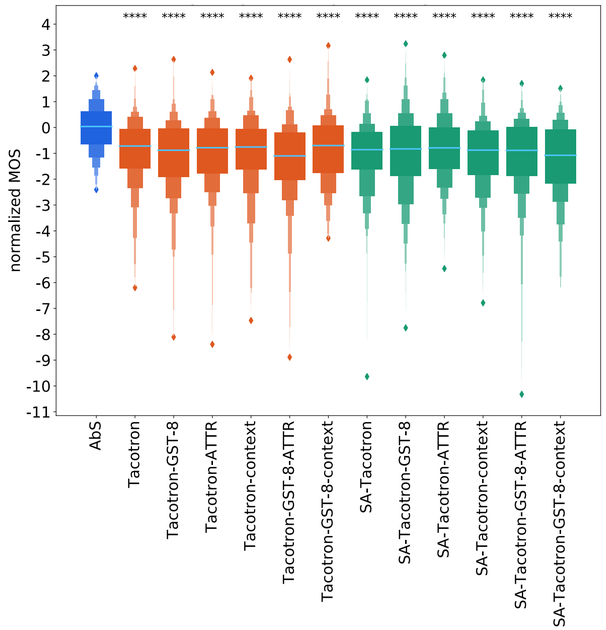
聴取実験に用いた音声の例

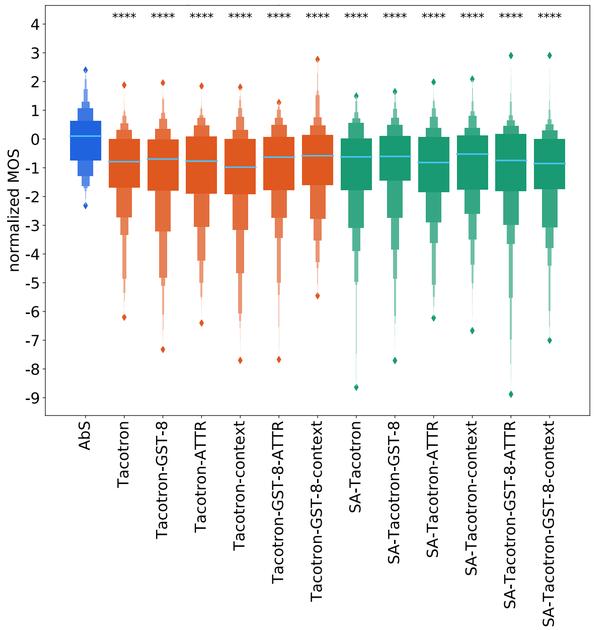
若い男	オー見なよ見なよ見なよエエ？
―――	この蟹。
―――	この蟹おかしいよ？
―――	蟹ってのは横に這うだろ？
―――	縦に歩ってんだいどうしたんだろうね.
噺家	たら蟹が顔上げて。
蟹	アすいません，エちょっと酔ってるもんですから.

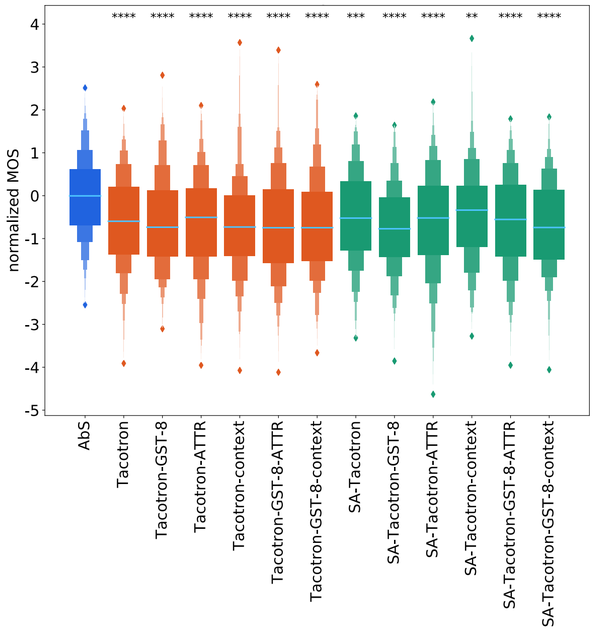
合成音声　分析合成音

聴取実験の結果









凡例

青: 分析合成音

赤: Tacotron 2系

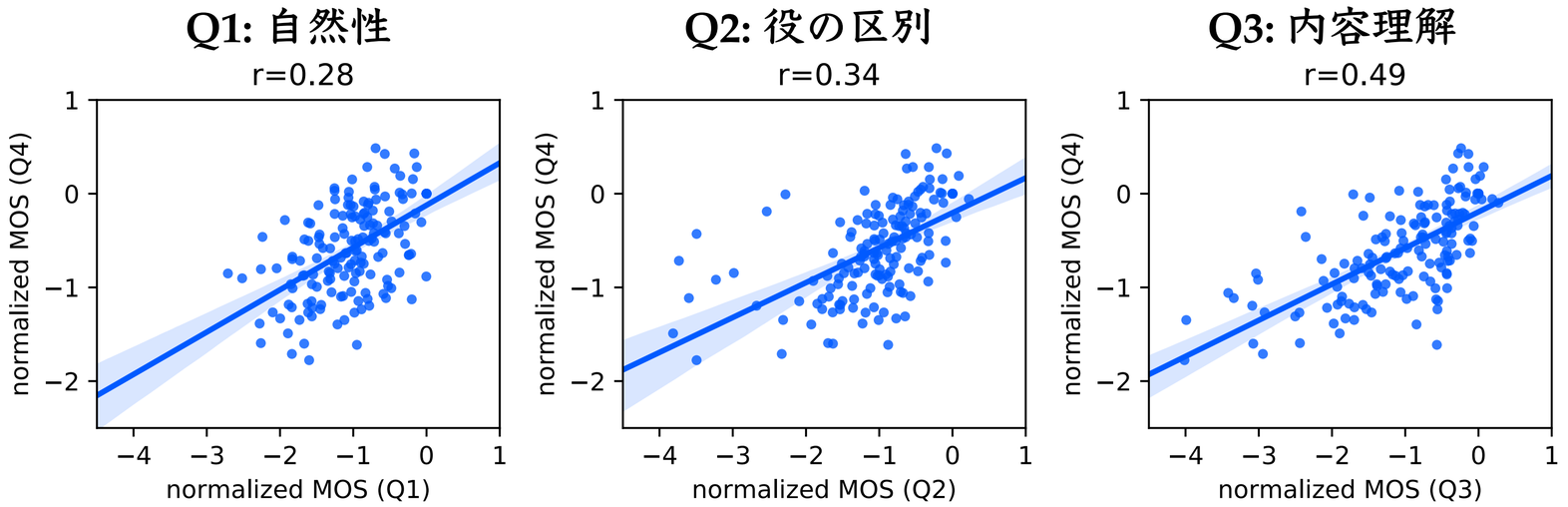
緑: SA-Tacotron系

**: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$

合成音声をどのように改善すべきか？

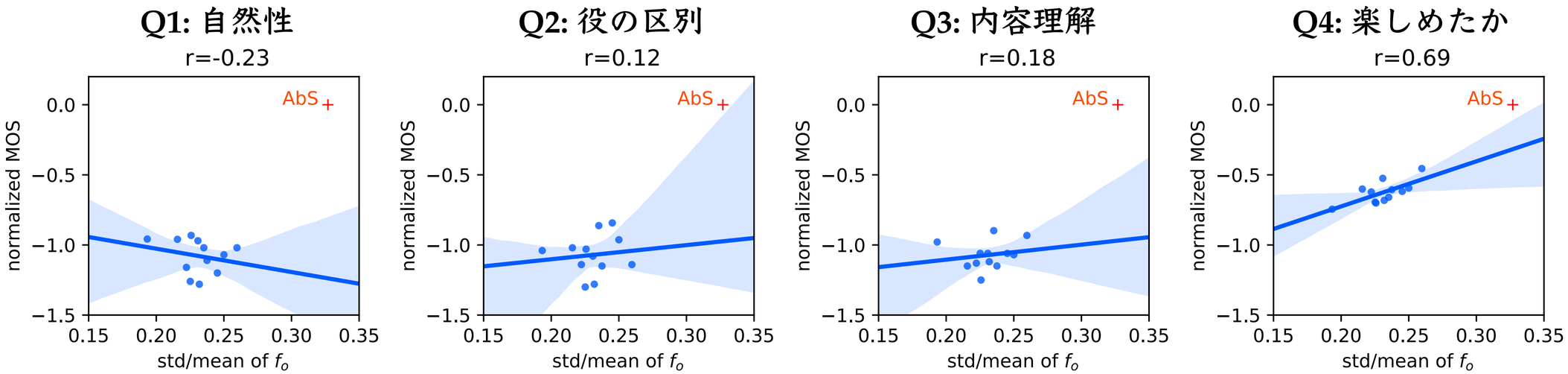
- 合成音声は，いずれの質問・モデルにおいても，分析合成音との間には評価に有意差があった.
- 合成音声をどのように改善すれば，聞き手がより楽しめるだろうか？

Q4（楽しめたか）とその他の質問との間の得点の相関関係



「どの程度楽しめたか」との相関は，
「自然性」よりも，「役の区別」「内容理解」のほうが強い．

基本周波数のばらつきと得点の相関関係

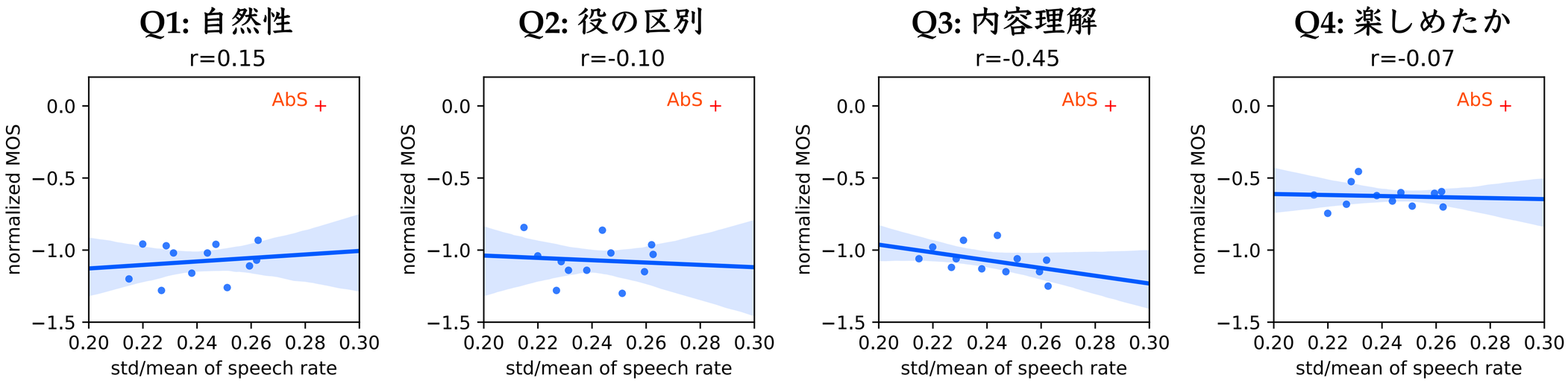


横軸: 小噺内の基本周波数の標準偏差を同平均で割ったもの

縦軸: 標準化された得点

「どの程度楽しめたか」と基本周波数のばらつきには，
ある程度強い相関がある．

話速のばらつきと得点の相関関係



横軸: 小噺内の話速の標準偏差を同平均で割ったもの

縦軸: 標準化された得点

各質問に対する評価と話速のばらつきには，
明らかな相関関係はない．

本章のまとめ

- 読み上げ音声合成で最高性能を誇るTacotron 2と，それをself-attentionで拡張したモデル (SA-Tacotron) を使用して，落語音声をモデル化した．

- 聴取実験の結果，**いずれの音声合成モデルも，分析合成音との間には評価に有意差があった．**

- しかしながら，**今後の改善に向けた，いくつかの興味深い示唆を得ることができた．**

1. 聴取者をより楽しませるためには，**従来音声合成の品質評価で重要視されてきた自然性を向上させるだけでなく，役の区別や内容理解の程度にも着目して，それらを向上させる必要がある．**

2. 聴取者をより楽しませるためには，現状の落語音声合成よりも**基本周波数のばらつきを大きくする必要がある．**

Chapter 7
人間の落語家との比較

本章は国際会議 ICASSP に投稿した論文の内容に基づいている.

Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, and Junichi Yamagishi, “How similar or different is rakugo speech synthesizer to professional performers?,” arXiv:2010.11549 [eess.AS], Oct 22 2020, submitted to ICASSP 2021.

本章の動機

- 江戸落語には3つの身分（前座・二ツ目・真打）があるにもかかわらず，前章では，音声合成と（モデル構築に用いた）真打との比較しか行わなかった．

- 現状の落語音声合成が真打の水準でないことは前章で明らかになっているが，それではどの程度の水準なのかを，各身分の人間の落語家と比較することで明らかにしたい．

- 前章では聴取実験に小噺を用いたが，より適切な評価のために，（独立して演じられる）演目を用いて評価したい．

聴取実験に用いた音声（人間）

- Database IIとして収録した音声を使用した.
  - 演者は真打（柳家三三，音声合成モデル構築に用いた話者と同じ）のほか，**ニツ目**（柳亭市童）と**前座**（柳家小ごと）の**3名**.
  - 小噺ではなく「味噌豆」という演目を収録.
    - 短いながらも，独立して演じられることのある演目（真打・柳家三三の助言により選定した）.

聴取実験に用いた音声（合成）

- 前章で最も高い評価を得たモデルである，SA-Tacotron-context（入力: テキスト＋全てのコンテキストラベル）を使用.

聴取実験の条件

データ	**Database Iより16演目**（7,341文，4.31時間）．ただし，0.5秒未満および20秒以上の音声を除いている．sv56により，全体の音量を-26dBovに正規化．
サンプリング周波数 / ビット / チャネル	48kHz / 16bit / mono
学習セット	6,362文（3.67時間）
検証セット	706文（0.42時間）
テストセット	273文（0.22時間）
音響特徴量	80次元のメルスペクトログラム．ただし，学習・検証・テストセットをあわせた全体で次元ごとに平均が0，分散が1になるように標準化した．
ボコーダ	WaveNet vocoder（学習・検証・テストセットの全てを用いて学習） 入力：メルスペクトログラム 出力：24kHz / 16bit mono waveform

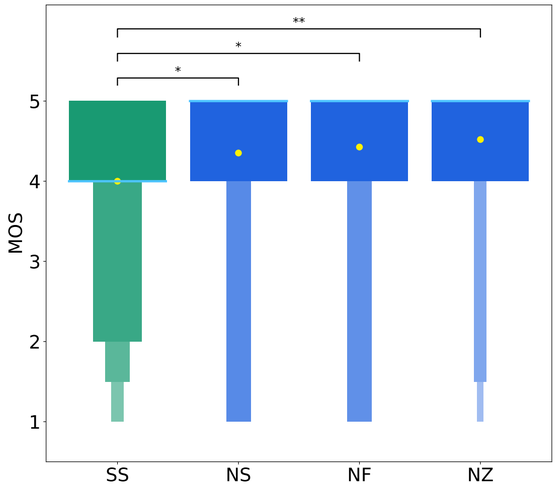
聴取実験の条件

- 演目「味噌豆」の音声を聴取実験用に用意した.
  - 合成音声は文単位で合成した音声を連結して作成し，文と文の間のポーズは，基本的に録音時と同じだけの長さの無音を挿入した．ただし，咀嚼音などの音声以外の音はモデル化しておらず，録音時のものをそのまま使用した.
- 聴取者は文単位ではなく，演目全体について評価した.
- 人間の落語家の音声は分析合成音ではなく，自然音声を用いた.
- 5段階のMOS試験を実施した.
  - 質問項目は，1) 自然性，2) どの程度役が区別できたか，3) どの程度内容が理解できたか，4) どの程度楽しめたか，5) 落語家としての技量の程度はどの程度か.
  - 聴取者292人が，それぞれいずれか1つのシステムで合成した音声を評価した.

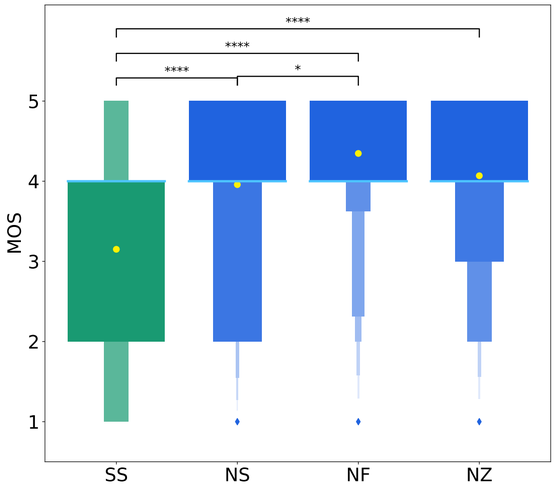
合成音声　真打　二ツ目　前座

聴取実験の結果

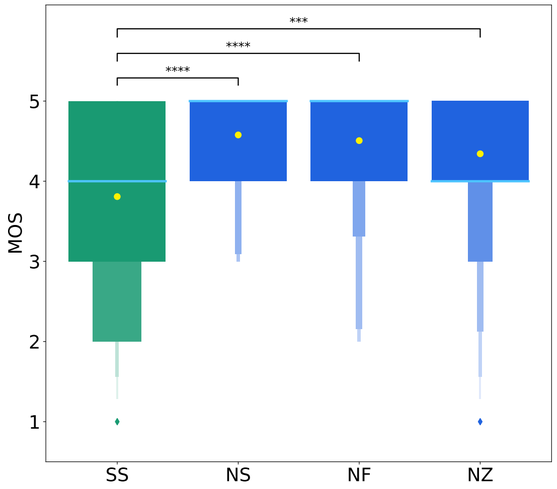
Q1: 自然性



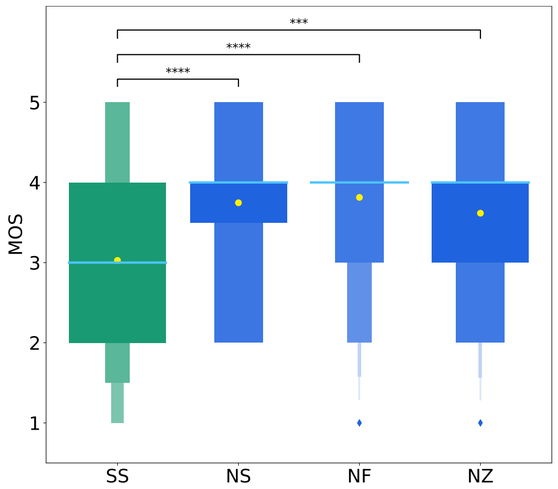
Q2: 役の区別



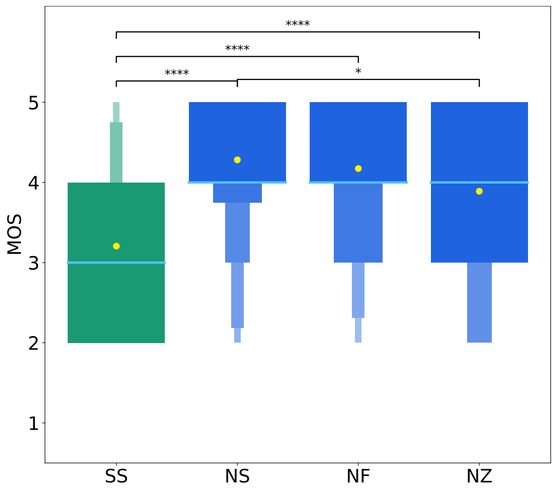
Q3: 内容理解



Q4: 楽しめたか



Q5: 技量



凡例

SS: 合成音声

NS: 真打

NF: ニツ目

NZ: 前座

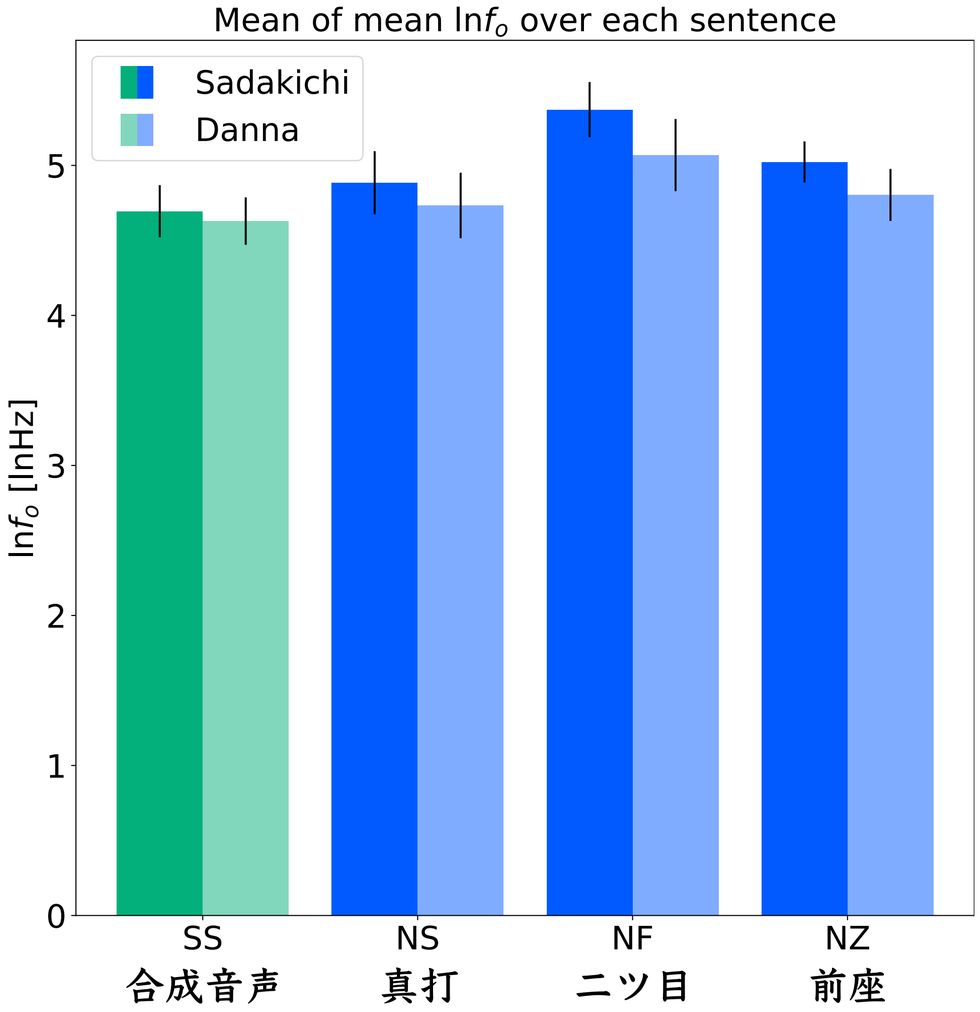
*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$

質問項目間の得点の相関関係

	Q2（役の区別）	Q3（内容理解）	Q4（楽しめたか）	Q5（技量）
Q1（自然性）	0.287	0.303	0.317	0.339
Q2（役の区別）	-	0.538	**0.486**	0.580
Q3（内容理解）	-	-	**0.597**	0.582
Q4（楽しめたか）	-	-	-	0.656

「どの程度楽しめたか」との相関は，「自然性」よりも，「役の区別」「内容理解」のほうが強い（前章と同じ傾向）．

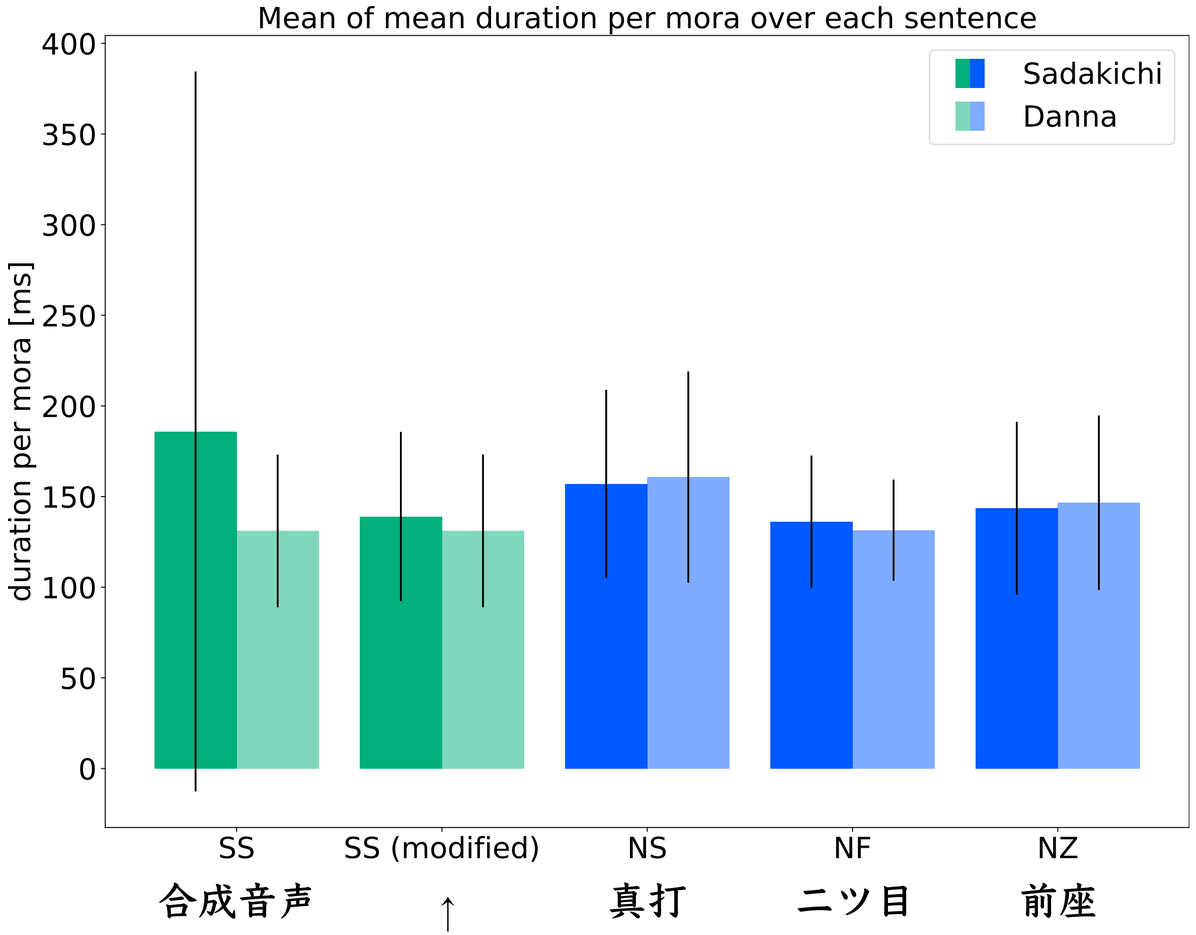
演者／役による基本周波数の違い



役による基本周波数の平均の差について，合成音声 (SS) はいずれの人間の落語家よりも小さい．

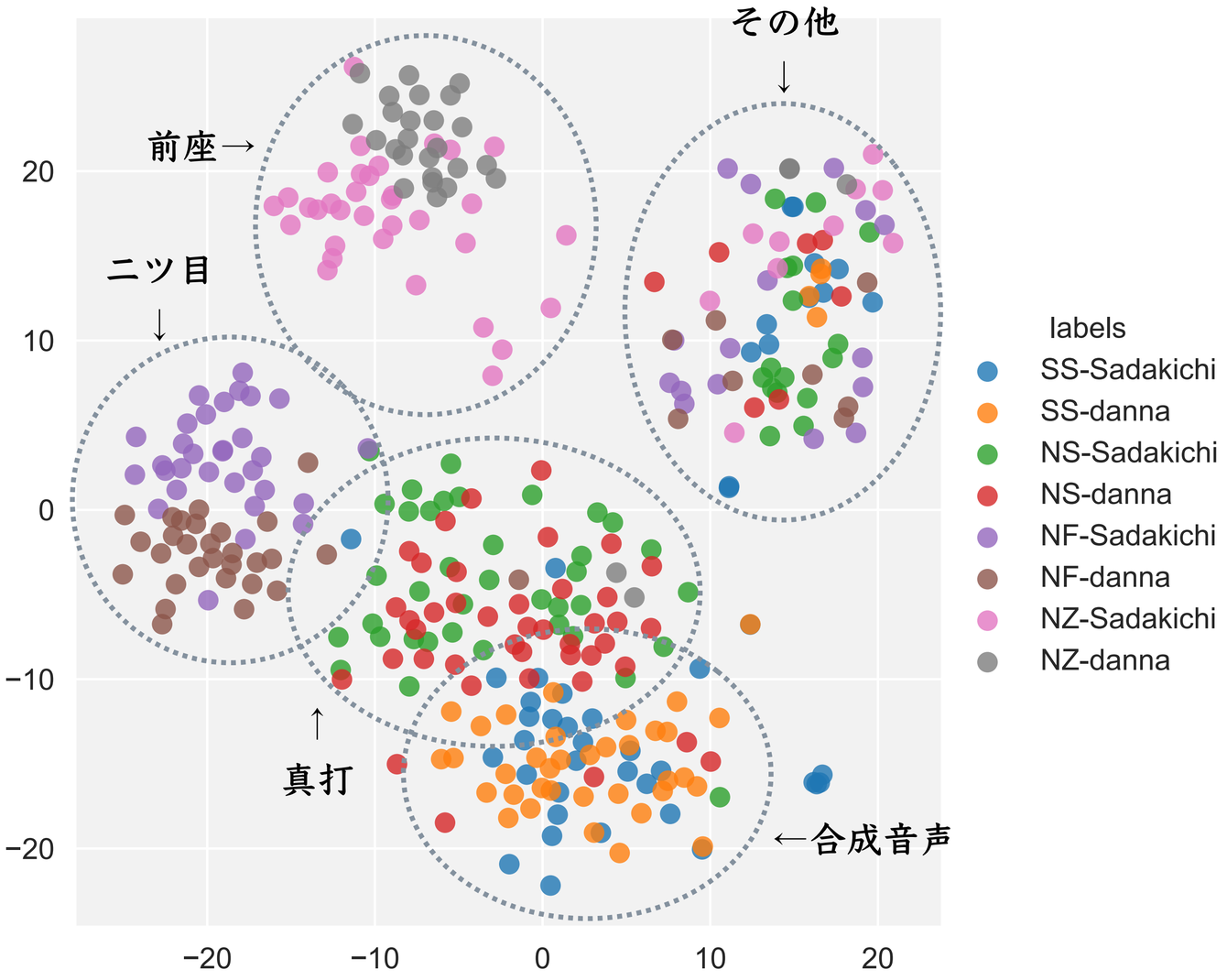
標準偏差については，合成音声および前座が役を問わず小さい傾向にある．

演者／役による話速の違い



話速については，明確に有利となるような条件は見られない．

x-vectorによる役の可視化



前座・ニツ目は，役ごとにクラスタが分かれている一方，真打・合成音声はそうでもない．

真打は**x-vectorでは捉えられない特徴**（時間方向にローカルな特徴あるいは韻律など）で役の区別を行っていると考えられる

（cf. 本実験で用いた合成音声には文ごとに付与したコンテキストラベルの情報を入力しており，グローバルな特徴を捉えている）．

本章のまとめ

- 江戸落語の3つの身分（前座・二ツ目・真打）の自然音声を比較対象として聴取実験を行い，現状の落語音声合成の水準を位置づけた．
- 現状の落語音声合成と人間の落語家の評価には有意差があった．
- しかしながら，**今後の改善に向けた多くの示唆を得ることができた．**

1. 聴取者をより楽しませるためには，**単に自然性を向上させるだけでなく，役の区別や内容理解の程度にも着目して，それらを向上させる必要がある**（前章と同じ傾向）．
2. 現状の落語音声合成は，**基本周波数を用いた役の区別に改善の余地がある．**
3. 役の区別をさらにつけるためには，**x-vectorで捉えられない特徴（時間方向にローカルな特徴など）もモデル化する必要がある可能性がある．**

結論

解決すべき課題 1 に対する回答

音声合成に適した落語音声データベースが存在しない.

- 落語のCDやDVDは数千種類が市販されている.
- 多くはライブ録音で，雑音や残響が多く音声合成には不適.
- **独自に音声データベースを構築する必要がある.**

回答

- 世界初の，**音声合成に適した落語音声データベースを構築**した.
- 書き起こしを行ったのみならず，コンテキストラベルを定義し，各文に付与した.

解決すべき課題２に対する回答

落語音声のスタイルは非常に多様である.

- 落語の本編は基本的に登場人物の会話から成り立っている.
  → **発話スタイルが多様.**

- 落語は演者がアドリブで，あるいは記憶を頼りに発話している.
  → **はっきり発音しているとは限らない.**

- 登場人物は，その性別・年齢・身分などによって異なる日本語を話す．→ **通常のテキスト処理が困難.**

- このような特徴により，通常の音声に比べて，**モデルの設計・学習がより困難**である.

回答

- End-to-end/sequence-to-sequence方式の音声合成モデルを用いることにより，落語音声をモデル化することに成功した.

- 課題は残るものの，GSTとコンテキストラベルにより，**多様なスタイルをある程度再現した.**

解決すべき課題 3 に対する回答

音声を聞いて，容易に登場人物の区別が付き，かつ内容が理解されるべきである．

- 落語の本編は基本的に登場人物の会話から成り立っている．
- 当然，**登場人物の区別が付くべき**である．
- （楽しんでもらうためには）**噺の内容が容易に理解されるべき**である．

回答

- 自然性・**役の区別・内容理解・どの程度楽しめたか**・技量について，聴取実験による主観評価を行った．
- その結果，課題の仮定のとおり，聴取者をより楽しませるためには，**従来音声合成の品質評価で重要視されてきた自然性を向上させるだけでなく，役の区別や内容理解の程度にも着目して，それらを向上させる必要がある**ことが示唆された．
- また，現状の落語音声合成は，**基本周波数を通じた表現に改善の余地がある**ことが示唆された．
- 以上のように，**従来の音声合成とは異なる方向性で改善を行う必要がある可能性が示された．**

解決すべき課題 4 に対する回答

合成された落語音声は，聞き手を楽しませるべきである．

- **人を楽しませる音声合成**として落語音声合成を開発している以上は，**聞き手が楽しめる**ような音声合成にするべきであるし，**どの程度楽しんだかを測定すべきである．**

回答

- 聴取実験では，**どの程度楽しめたかについて直接評価を行った．**
- また，**人間の落語家（前座・二ツ目・真打の各身分）との比較を行った．**
- 音声合成と人間の落語家との間にはまだ差があるものの，**将来の改善に向けた多くの示唆を得ることができた．**

残された課題

1. 合成音声の表現には改善の余地がある.

- 自然性のみならず，役の区別や内容理解の程度についても改善すべきである.
  - ただし，**落語の登場人物の属性は非常に偏っている**ため，モデルの設計には相応の工夫が必要である.
- 合成時に**文を超えた依存性を考慮**することも検討している（人間の落語家は当然考慮している）.

2. 文と文の間のポーズの長さを予測すべきである.

- 本論文では予測していなかったが，当然予測すべきである.

残された課題

3. 現状の音声合成ができなくて，人間の落語家がやっていることはまだたくさんある.

- 人間の落語家は，音声以外の多様な音表現を行う.
- 人間の落語家は，観客の反応次第で演じ方を変えることがある.
- 人間の落語家は，新しい噺を作ることがある（新作落語）.
- 人間の落語家は，音声だけでなく，身振り手振りを使って落語を演じている.