Vol.2020-SLP-133 No.15
2020/10/23

落語音声合成は人間の落語家にどれだけ迫れるのか？

加藤 集平[1,2,†1,a)]　安田 裕介[1,2]　Xin Wang[2]　Erica Cooper[2]　山岸 順一[2]

概要：私たちは，人を楽しませる音声合成として，落語音声合成の研究に取り組んでいる．これまでに，独自の音声合成向け落語音声データベースの構築ならびに落語音声の分析，そのデータベースを用いたend-to-end (sequence-to-sequence) 落語音声合成システムの構築，聴取実験による落語音声合成システムの評価を行ってきた．これまでの研究においては，音声合成との比較対象として，モデル学習に使用した落語家の音声のみを使っており，「音声を聞いてどれだけ楽しめたか」などの評価指標について，（モデル学習に使用した）落語家にはまだ及んでいないという結果を得ている．ところで，一連の研究の対象としている江戸落語には身分制度があり，身分は下から順に前座，二ツ目，真打と呼ばれる．モデル学習に使用した落語家は最高位の真打であるが，音声合成がその水準に達していないとすれば，果たして前座，二ツ目とくらべてどの程度の水準にあるのだろうか．これを明らかにするために，本稿では，同一の演目の音声を用いて，前座，二ツ目，真打，そして音声合成を比較する聴取実験を行った．聴取実験の結果，音声合成は前座，二ツ目，真打いずれのレベルにも達していなかったものの，「音声を聞いてどれだけ楽しめたか」など一部の評価項目については，前座との差は他よりも小さなものであった．また，「音声を聞いてどれだけ楽しめたか」の評価値は，「演者は人間だと思うか」の評価値との相関は比較的弱く，「登場人物の役が区別できたか」の評価値もしくは「内容がどれだけ理解できたか」の評価値との相関が比較的強かった．このことから，私たちが構築した落語音声合成は音声としてある程度高い自然性を有しているものの，役の区別などの表現のモデリングに不足があり，結果として内容理解ひいては十分に楽しむことが難しいものであることが示唆された．

How close is rakugo speech synthesis to human rakugo performers?

1. はじめに

私たちは，人を楽しませる音声合成としての落語音声合成の研究に取り組んでいる．ここでいう人を楽しませる音声合成とは，単純な情報伝達にとどまらない音声合成のことである．音声はメディアとして，発話内容，話者の感情，個性，意図などを聴取者に伝えている．従来の音声合成研究は，これらの情報の正確な伝達に主眼を置いてきたと言ってもよく，特に，より人間に近い自然な発声をする音声合成が追求されてきた．これらの目的は限られた条件下ではあるが既に達成されており，mean opinion score (MOS) が人間の音声と同等の音声合成システムが存在する [1], [2].

一方，落語を含む話芸においては，演者は音声というメディアを通じて聴取者（観客）を楽しませる．言い換えると，音声を通じて聴取者の感情を喚起しているとも言え，音声の持っている役割が単に情報を伝達するだけではないことが分かる．しかしながら，従来の音声合成研究においては，この「人を楽しませる」という点は必ずしも重視されてこなかった．また，本稿で取り扱っている落語についてはニコニコ動画にいくつか音声合成を用いた作品が投稿されている [3], [4], [5] が，少なくとも著者の主観としては非常に単調であり，人間の口演と比べて質が遥かに劣るものであるという印象である．このように機械（音声合成）と人間との間にはまだ大きな隔たりが存在する．私たちは，機械と人間のコミュニケーションをより深いものにするための一つの方策として，このような隔たりを埋めるべきで

[1] 総合研究大学院大学
The Graduate University for Advanced Sciences (SOKENDAI), Hayama, Miura, Kanagawa 240–0193, Japan
[2] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[†1] 現在，株式会社 RevComm
Presently with RevComm Inc., Shibuya, Tokyo 150–0002, Japan
[a)] shuhei@shuheikato.info

あると考えている．そこで私たちはこれまで，人を楽しませる音声合成としての落語音声合成を開発し，評価を行ってきた．

私たちがこれまでに行った研究業績は，独自の音声合成向け落語音声データベースの構築ならびに落語音声の分析，そのデータベースを用いた end-to-end (sequence-to-sequence) 落語音声合成システムの構築，聴取実験による落語音声合成システムの評価である [6], [7], [8], [9]．聴取実験では，モデル学習に使用した落語家の音声に対して，「音声を聞いてどれだけ楽しめたか」などの指標について評価を行い，高品質な音声を生成可能ではあるものの，合成音声の表現力は，モデル学習に使用した落語家にはまだ及ばないという結果を得ている．

ところで，私たちの一連の研究では江戸落語を対象としている．江戸落語には身分制度があり，身分は下から順に前座，二ツ目，真打と呼ばれる．モデル学習および上記の聴取実験での比較に使用した落語家は最高位の真打である．つまり，落語音声合成はまだ真打の水準に達していないと推察される．それでは，落語音声合成は，果たして前座，二ツ目とくらべてどの程度の水準にあるのだろうか．これを明らかにし，分析を行うことが本稿の目的である．この目的を実現するため，私たちは前座，二ツ目，真打のそれぞれについて同一の演目を収録し，さらに同演目を合成された音声も加えて聴取実験を行った．本稿では，その聴取実験の方法および結果について説明し，私たちが構築した落語音声合成の水準と，その水準にある理由について考察する．

2. 江戸落語における身分制度

落語は発達してきた地域によって大きく二つに分けられ，それぞれ**江戸落語**（江戸・東京で発達）および**上方落語**（京都・大阪で発達）と呼ばれる．本稿で対象としているのは，江戸落語である．

江戸落語には身分制度があり，下から順に**前座**，**二ツ目**，**真打**と呼ばれる．落語家を志す者は，まず真打である師匠に入門し，見習い期間を経て前座となる．前座は，まだ正式には落語家とは認められていない身分であり，仕事としては，寄席（落語を主に上演する劇場）の楽屋における雑務のほか，師匠の身の回りの世話などを行う．このような仕事を行いながら，自らの師匠や他の真打から噺（演目）を教わり，練習を重ねる．前座のうちは高座（舞台）に上がることはないが，文字通り寄席の番組（上演プログラム）の前座として落語を演じることがある．

前座としての修行を終えて昇進が認められると，二ツ目という身分になる．二ツ目になると，上記の雑務や師匠の世話といった仕事からは解放され，一人の落語家として高座に上がることができる．一方で，真打と異なり弟子を取ることはできず，寄席で主任（トリ）を務めることもない．この間も，自らの師匠や他の真打から噺を教わり，練習を重ねることに変わりはない，

二ツ目としての修行を終えて認められると，真打という身分になる．真打は先述の通り，弟子を取ることができ，寄席で主任を務めることができる．

人によって差はあるものの，通常，前座から二ツ目に上がるまでは 3 年から 5 年，二ツ目から真打に上がるまでは 10 年程度の期間を要する．2020 年現在，江戸落語では（前座を含め）およそ 600 名が落語家として活動している [10], [11], [12], [13].

3. 落語音声の追加録音

落語音声合成は，（前座を含めた）落語家と比べて，どの程度の水準にあるのだろうか．これを明らかにするために，前座，二ツ目，真打のそれぞれについて，同一の演目[*1]を録音し，聴取実験に用いることとした．

録音は 2020 年 1 月に行われた．演者は，前座が柳家小ごと[*2]，二ツ目が柳亭市童，真打が柳家三三（音声合成モデルの学習に用いた演者と同一）である．録音条件は，音声合成モデルの学習に用いた落語音声データベース [9] の録音時と同一で，録音ブースの中で演者がそれぞれ一人で落語を演じた．なお，観客は一人もいなかった．そのため，観客からの反応は一切なく，収録も行わなかった．

収録演目は，上記の落語音声データベース中の演目のマクラ等に含まれる 22 の小噺と，やはり同データベースに収録されている『味噌豆』という噺である．収録時間は，前座，二ツ目，真打の順に，小噺が 20 分 37 秒，18 分 3 秒，20 分 42 秒，『味噌豆』が 2 分 31 秒，2 分 40 秒，4 分 14 秒であった．なるべく自然な流れの落語音声を収録するために，演者が希望した場合を除いては，言い淀みや言い直しがあっても再録音は行わなかった．ただし，前座の録音には真打が立ち会い，必要に応じて指導を行った．

4. 音声合成モデル

音声合成モデルは，私たちの過去の研究 [9] において，聴取実験で最もよい評価を得た SA-Tacotron-context を用いた．特徴量，ネットワーク構造，学習条件などについては [9] を参照されたい．ただし，[9] との相違点として，モデル学習に用いる学習・検証セットは『味噌豆』のデータを除いて作成し直し，学習セットとして 6,362 文（3 時間 40 分 19 秒），検証セットとして 706 文（24 分 59 秒），テストセットとして 273 文（13 分 18 秒）を使用した．また，音声合成モデルの出力するスペクトログラムおよび WaveNet [14] ボコーダ [15], [16] で生成した音声波形のサンプリング周波数を 24 kHz に変更した．それに伴い，音

[*1] 台本があるわけではないので，言い回しなどは演者によって異なる．
[*2] 2020 年 2 月廃業.

表 1　音声合成モデルおよび WaveNet ボコーダの音響分析条件
Table 1　Acoustic analysis conditions of speech synthesis model and WaveNet vocoder.

Sampling rate	24 kHz
Frame shift	12 ms
FFT size	2,048

表 2　各質問に対する MOS 値の相関係数
Table 2　Correlation coefficients of MOSs between questions.

	Q2	Q3	Q4	Q5
Q1	0.287	0.303	0.317	0.339
Q2	-	0.538	0.486	0.580
Q3	-	-	0.597	0.582
Q4	-	-	-	0.656

声合成モデルおよび WaveNet ボコーダの音響分析条件を**表 1** のように変更した．

5. 聴取実験

5.1 聴取実験の目的

私たちが構築した落語音声合成が，人間の落語家（前座，二ツ目，真打）と比べてどの程度の水準にあるかを明らかにするため，3 で録音した音声および 4 により合成した音声を用いて聴取実験を行った．

5.2 実験条件

聴取実験には，演目『味噌豆』の音声を用いた．以前の研究では小噺の音声を用いて評価を行っていたが，落語音声の水準を適切に評価するには不十分であり，短いながらも噺（演目）と呼べる内容で比較を行うことがより適切であると考えたからである．そこで，録音協力者である真打・柳家三三の助言に基づき，短いながらも噺として成立する『味噌豆』を採用することにした．音声の合成は文単位で行い，文と文の間のポーズ長は予測しなかった[*3]．聴取者は，文単位ではなく，噺全体を聞いて評価を行った．また，全ての音声について，噺全体の音量が −26 dBov になるように，sv56 [17] を用いて正規化した．

評価は MOS 試験により行った．聴取者は，人間の音声（前座，二ツ目，真打）あるいは合成音声の 4 つの音声のうちいずれか 1 つの音声を聞いて，以下の 5 つの質問について，それぞれ 5 段階で評価を行った．

1) 演者は人間だと思うか？
2) 登場人物の役の区別はどの程度付いたか？
3) 噺の内容はどの程度理解できたか？
4) 音声を聞いてどの程度楽しめたか？
5) 演者の落語家としての技術はどの程度だと思ったか？

落語が話芸であり，観客にとっては娯楽であることを鑑みると，最も重要な質問は 4 である．また，落語音声合成の噺家としての水準を推測するためには，質問 5 の回答を利用することができる。その他の質問は，これらの質問の結果に与える要因を探るための質問である．

なお，聴取者は合計 292 人であった．

[*3] 当然，本来は予測すべきであるが，本稿の主旨ではない．

5.3 結果

結果を**図 1** に示す（SS: 音声合成，NS: 真打, NF: 二ツ目，NZ: 前座．他図においても同様）．統計分析として，Brunner-Munzel 検定 [18] を行い，Bonferroni 法によって補正した．その結果，いずれの質問についても，音声合成は人間（前座，二ツ目，真打）の水準に及ばず，有意差が観測された．

しかしながら，有意差の傾向には差が見られた．まず，質問 1（演者が人間だと思うかどうか）については，音声合成のスコアは平均 4.0 と高かった．これは合成音声の自然性が高いことを裏付けている．真打のスコアとの間の p 値は $0.01 < p < 0.05$ であった．それに対して，質問 2 から質問 4 における音声合成の平均スコアは 3 と 4 の間であり，前座，二ツ目，真打それぞれのスコアと大きな差があった．p 値も，$p < 0.005$ もしくは $p < 0.001$ であった．一方，質問 3 および質問 4 については，音声合成と前座のスコアの差に関する p 値は，音声合成と二ツ目および真打のスコアの差に関する p 値よりも小さかった．

噺家としての水準を推測するために設けた質問 5 のスコアについては，期待通り，真打 > 二ツ目 > 前座となった．なお，音声合成のスコアは，前座，二ツ目，真打それぞれのスコアよりも低かった．

6. 考察・議論および関連する分析

残念ながら，現在の落語音声は人間（前座，二ツ目，真打）の水準に及んでいないことが分かった．では，それはどのような要因が影響しているのだろうか．これを考察するために，いくつか分析を行った．

まず，各質問に対する MOS の相関係数を**表 2** に示す．この表を見ると，質問 1（演者が人間だと思うかどうか）は，他の各質問との相関係数が比較的小さかったことが分かる．特に，質問 4（音声を聞いて楽しめたかどうか）との質問は，質問 5（演者の落語家としての技術）を除いては，質問 1，質問 2（役の区別），質問 3（内容理解）の順に相関係数が大きくなっている．また，質問 2 と質問 3 の相関係数も比較的大きい．これらと図 1 をあわせて考えると，私たちが構築した音声合成による落語音声は，**音声としては比較的人間に近い自然なものである**一方で，**役の区別が十分に付いていないために内容理解が不十分**で，結果として**十分に楽しめていない**と推察できる．つまり，音声

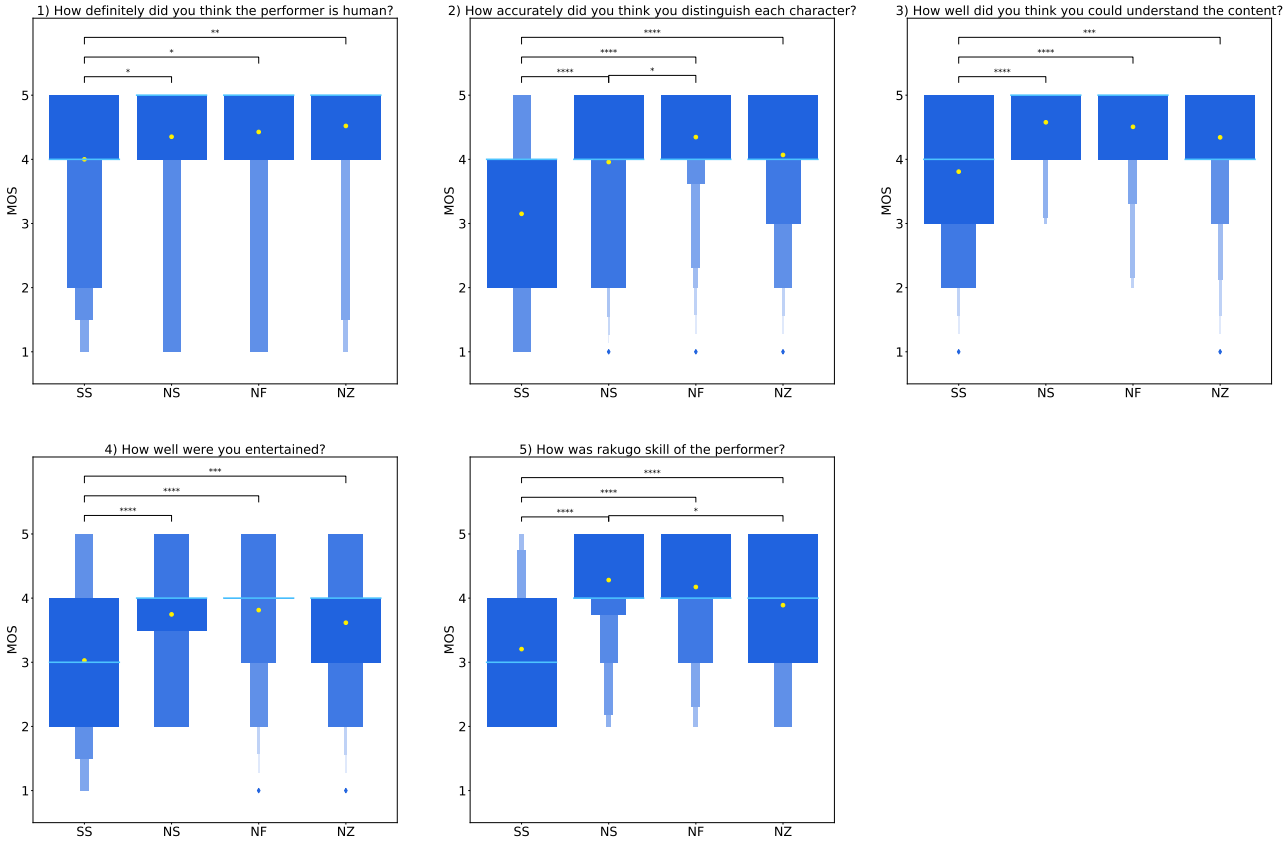

図 1 聴取実験の各質問に対する評価結果の boxen plot．水色の線は中央値，黄色の点は平均値を示す．*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$.

Fig. 1 Boxen plots for each questions of listening test. Light blue lines and yellow dots represent medians and means, respectively. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$.

合成が人を楽しませるためには，自然性だけを追求していても不十分であり，落語の場合，具体的には役の区別を十分につける必要がある可能性がある．

それでは，役の区別はどのような要因で不十分なのだろうか．これを考察するために，役の区別に影響があると考えられる対数基本周波数 ($\ln(f_o)$) および話速（1 モーラあたりの継続長）について，文ごとに平均および標準偏差を計算し，それらを役別に噺全体で平均したものを計算した．結果を**図 2** および**図 3** に示す．なお，Sadakichi（定吉）は小僧（丁稚），Danna（旦那）は主人である．

図 2 より，音声合成による落語音声は，役による f_o の平均の差が人間の音声のものよりも小さいことが分かる．特に，二ツ目の音声の差の付け方とは大きな開きがある．役の区別をどれだけ大袈裟に付けるかは演者の個性でもあり，モデルの学習に用いた演者である真打（柳家三三）は役の区別をあまり大袈裟に付けないとのことである [19]．その（自然）音声と比べても，合成音声の役による f_o の平均の差は小さい．このことから，私たちが構築した落語音声合成は，f_o による役の区別の付け方が不十分であると言ってもよいだろう．

話速についてはどうだろうか．図 3 より，どの人間の落

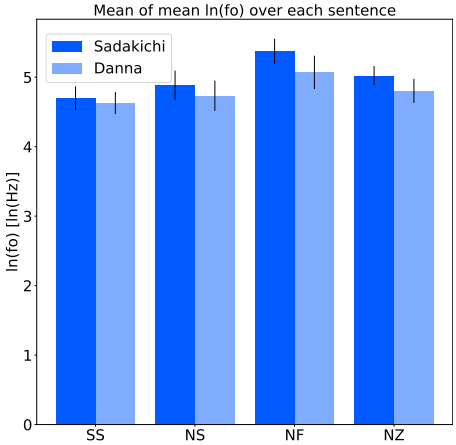

図 2 各文の対数基本周波数 ($\ln(f_o)$) の平均および標準偏差を，役別に噺全体で平均したもの．

Fig. 2 Means per character of means and standard deviations of logarithmic fundamental frequency ($\ln(f_o)$) over each sentence.

語家（前座，二ツ目，真打）も，（この噺『味噌豆』においては）話速を大きく変えることによって役の区別を付けていないことが分かる．これは，音声合成も同様である．しかし，人間の落語家は，噺によっては話速を大きく変える

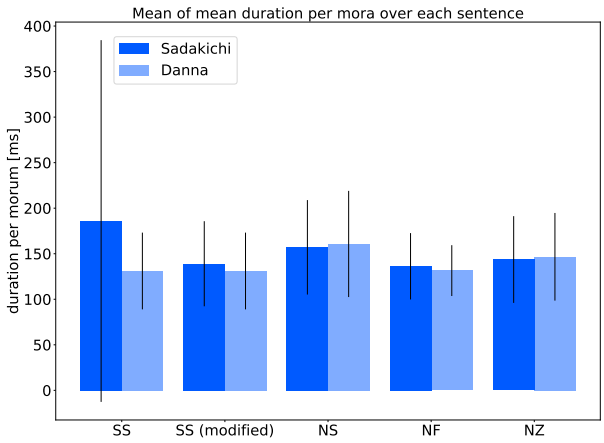

図 3 各文の 1 モーラあたりの継続長の平均および標準偏差を，役別に噺全体で平均したもの．SS (modified) は，継続長を過剰に長く予測した 2 つの文を除いた結果である．

Fig. 3 Means per character of means and standard deviations of duration per mora over each sentence. SS (modified) is calculated based on sentences excluding two sentences which duration was estimated too long.

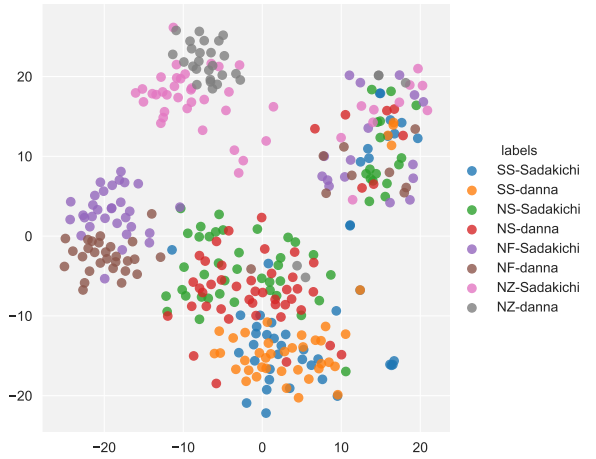

図 4 文単位で求めた音声の x-vector を t-SNE で可視化したもの

Fig. 4 Visualization of x-vector for each sentence using t-SNE.

ことによって役の区別を付けることがあることに注意したい [6]．図 3 からは，私たちが構築した落語音声合成が話速による役の区別の付け方が不十分であると言うことはできないが，以前の研究 [9] で聴取実験に用いた小噺の中には，自然音声との話速の差の付け方の違いが大きなものがあった．

また，話者性についてはどうだろうか．**図 4** は，音声合成および人間の落語家それぞれに対して，文ごとに音声の x-vector [20] を求め，それを役別に分類した上で，t-SNE [21] を用いて平面上に可視化したものである．

図 4 では，「前座」，「二ツ目」，「真打（自然音声）および音声合成」，「全システム」の 4 つのクラスタが観測できる．「前座」および「二ツ目」のクラスタについては，さらに役ごとに概ねクラスタリングされている．一方，「真打（自然音声）および音声合成」のクラスタについては，真打と音声合成でさらにクラスタリングされているものの，それぞれ役ごとにはクラスタリングされていない．このことから，そもそも真打は役ごとに（x-vector で表現されるような）話者性の差を大きく付けておらず，真打と同一の話者の音声を元に学習した音声合成モデルも，話者性について差を大きく付けるようには学習していなかったと言える．

以上より，私たちが構築した落語音声合成は，少なくとも f_o による役の区別の付け方が不十分であることが分かった．今後はこれを改善することにより，役の区別がより明確になり，結果として内容がより理解しやすくなり，落語音声合成をより楽しめるようになる可能性がある．

なお，本聴取実験の枠組みの正当性についても議論しておきたい．表 1 において，質問 5（演者の落語家としての技術）では，人間の落語家の身分が高い順（真打，二ツ目，前座）に MOS 値の平均は低下していたが，有意差は真打と前座の間 ($p < 0.05$) にしか見られなかった．また，質問 4（楽しめたかどうか）に至っては，人間の落語家の間では有意差が見られなかった．落語家の昇進は年功序列的な側面はあるものの，通常その技量がある程度反映されていると考えられる．3 で前座，二ツ目，真打のそれぞれについて録音を行ったのは，技量，ひいては観客がそれぞれの落語音声を聞いて楽しむ程度に差があることを期待したからである．

質問 4 および質問 5 の評価結果がこの意図とは裏腹であった理由は様々なものが考えられるが，聴取者の属性として，落語を普段聞いているかどうかを問わなかったことは影響しているかもしれない．著者の中でも筆頭著者や最終著者は一連の研究を通じて落語音声を比較的よく聞いており，特に筆頭著者は音声データベースのアノテーション作業で細部に至るまで落語音声を聞いたほか，研究のために寄席や落語会にも足を運び，また落語についての文献も多く読んでいる[*4]．そのような筆頭著者や最終著者の耳には，明らかに技量や楽しむ程度に差を感じていた．聴取者を，落語を普段聞いている人に限定することが，さらに有意義な結果を得るために必要かもしれない．

また，落語は話芸と言いながら，本来は視覚と聴覚のマルチモーダルで表現を行う芸である．3 の録音時にはビデオ撮影を行ったものの，現在の落語音声合成は視覚情報を合成しないため，撮影した映像は聴取実験に使用していない．今後，視覚情報を交えた評価を行うことで，評価結果が変わる可能性もあるだろう．

7. おわりに

本稿では，落語音声合成が人間の落語家（前座，二ツ目，真打）と比べてどの水準にあるかを明らかにすることを目的として，落語音声の録音およびそれらを用いた聴取実験

[*4] さらに，近いうちにプロに習うつもりである．

を行った．結果としては，音声合成による落語音声の水準は人間の落語家による音声のそれに及ばないことが判明した．しかしながら，少なくとも f_o による役の区別をより明確にすることで，その差を縮めることができる可能性を示した．

今後は，役の区別をより明確に表現できるような音声合成モデルの設計や学習の枠組みを検討する予定である．ただし，役の属性（性別，年齢，身分など）の，データベース中の存在割合は非常にアンバランスであり，極端に存在例が少ないものもある（武家奉公をしている女性など）．そのため，十分に役を区別できるような音声を合成するためには，単純な複数話者モデリングだけでは不十分であり，相応の工夫が必要と考えられる．また，現在は予測していない文と文の間のポーズ長を予測する，視覚情報を合成するなどの課題にも今後取り組んでいく予定である．

謝辞 本研究は，JST CREST (JPMJCR18A6, VoicePersonae project) ならびに JST AIP チャレンジ，および科研費 (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051) の支援を受けたものである．

参考文献

[1] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y. and Wu, Y.: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, *Proc. Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, pp. 4779–4783 (2018).

[2] Li, N., Liu, S., Liu, Y., Zhao, S. and Liu, M.: Neural Speech Synthesis with Transformer Network, *Proc. AAAI Conf. Artif. Intell. (AAAI-19)*, Honolulu, HI, USA (2019).

[3] MSS: 噺家ミクの特選落語　まんじゅうこわいですぅ (2009), 入手先 ⟨https://www.nicovideo.jp/watch/sm5899050⟩.

[4] 目つき悪いP：【ボーカロイド落語】看板のピン【目つき悪いミク】 (2011), 入手先 ⟨https://www.nicovideo.jp/watch/sm13959846⟩.

[5] zky: 【初音ミク】 ボカロ落語 『野ざらし』 (2012), 入手先 ⟨http://www.nicovideo.jp/watch/sm17066984⟩.

[6] 加藤集平，高木信二，山岸順一，Wang, X.：WaveNet を用いた落語音声合成の検討およびコンテキストの分析—人を楽しませる音声合成に向けて—，日本音響学会 2018 年秋季研究発表会講演論文集，大分県大分市，pp. 1139–1142 (2018).

[7] 加藤集平，高木信二，山岸順一，安田裕介，Wang, X.：落語音声合成における Tacotron およびコンテキスト特徴量の使用とその評価，電子情報通信学会技術研究報告，Vol. 118, No. 497, 長崎県長崎市，pp. 161–166 (2019).

[8] Kato, S., Yasuda, Y., Wang, X., Cooper, E., Takaki, S. and Yamagishi, J.: Rakugo Speech Synthesis Using Segment-to-Segment Neural Transduction and Style Tokens — Toward Speech Synthesis for Entertaining Audiences, *Proc. 10th ISCA Speech Synthesis Workshop (SSW10)*, Vienna, Austria, pp. 111–116 (2019).

[9] Kato, S., Yasuda, Y., Wang, X., Cooper, E., Takaki, S. and Yamagishi, J.: Rakugo Speech Synthesis and Its Limitations: Toward Speech Synthesis That Entertains Audiences, *IEEE Access*, Vol. 8, pp. 138149–138161 (online), DOI: 10.1109/ACCESS.2020.3011975 (2020).

[10] 落語協会：芸人紹介 入手先 ⟨https://rakugo-kyokai.jp/variety-entertainer/⟩.

[11] 落語芸術協会：協会員プロフィール 入手先 ⟨https://www.geikyo.com/profile/⟩.

[12] 東京かわら版：東都寄席演芸家名鑑，東京かわら版 (2018).

[13] 落語立川流：立川流の落語家たち 入手先 ⟨http://tatekawa.info/member/⟩.

[14] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, arXiv:1609.03499 [cs.SD] (2016),

[15] Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L. and Yamagishi, J.: A Comparison of Recent Waveform Generation and Acoustic Modeling Methods for Neural-Network-Based Speech Synthesis, *Proc. Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, pp. 4804–4808 (2018).

[16] Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K. and Toda, T.: Speaker-dependent WaveNet Vocoder, *Proc. INTERSPEECH*, Stockholm, Stockholm, Sweden, pp. 1118–1122 (2017).

[17] Int. Telecommun. Union: Recommendation G.191: Software Tools and Audio Coding Standardization (2005),

[18] Brunner, E. and Munzel, U.: The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation, *Biometrical J.*, Vol. 42, No. 1, pp. 17–25 (2000).

[19] 東京弁護士会：インタビュー 落語家 柳家三三さん，*LIBRA*, Vol. 11, No. 11, pp. 22–25 (2011).

[20] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: X-Vectors: Robust DNN Embeddings for Speaker Recognition, *Proc. Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, pp. 5329–5333 (online), DOI: 10.1109/ICASSP.2018.8461375 (2018).

[21] van der Maaten, L. and Hinton, G.: Visualizing Data Using t-SNE, *J. Mach. Learn. Res.*, Vol. 9, pp. 2579–2605 (2008).