人を楽しませる音声合成の実現に向けて: 落語とディープランニングの融合

代表研究者　　山　岸　順　一　　情報・システム研究機構国立情報学研究所 准教授
共同研究者　　加　藤　集　平　　総合研究大学院大学複合科学研究科情報学専攻 博士課程学生

1 概要

テキストを音声に変換する音声合成は、コールセンターの自動応答、自動放送、音声アシスタントなど様々な分野で使用されている。しかし、これまでの音声合成は、音声の持つ情報伝達機能に主眼が置かれており、発話内容を正確に伝えることが重視されてきた。発話内容を正確に伝えるためには、音声の内容を正確に聞き取れた割合を示す了解度や、自然性を向上させることが必要であるが、これらは近年の深層学習の発展により大きな進歩を見せている。また、喜怒哀楽などの感情表現や様々な口調（発話スタイル）を実現する音声合成技術も研究されている。しかし、これらも発話者の情報や状態を伝える手段にすぎず、情報伝達機能の域を出ない。しかし、音声には情報伝達機能だけでなく、話芸に代表されるように人を楽しませるといった機能がある。単なる情報伝達を超えた、人を楽しませる音声合成を実現することで、音声の持つ人を楽しませる機能の解明に寄与するほか、新たな娯楽の提供、人間・コンピュータ間のコミュニケーションの発展が期待される。そこで本研究では、人を楽しませる音声合成の実現を目標に、代表的な話芸である落語を音声合成により行う研究を行った。まず落語家の最上位階級である真打の実演を、防音スタジオにて五日間にわたり、合計 20 時間収録した。次に本収録に対し、フィラーや言い澱みも含む書き起こし、登場人物、登場人物の状態、登場人物の上下関係、特殊音、噺の構造などを詳細に付与した。最後に、これらの情報を入力とする非線形自己回帰型ニューラルネットワークを学習し、音響特徴量系列を予測、予測された音響特徴量系列から音声波形を Wavenet という CNN により生成する落語音声合成システムを構築した。

2 落語音声合成

2-1 はじめに

音声合成の品質は年を追うごとに向上し,限られた条件下では自然音声と同等の平均オピニオン評点(MOS)を得るまでになっている[1].しかし,音声合成の研究で主に対象とされてきたのは読み上げ調の音声であり,感情表現をはじめとする多様な表現については，読み上げ調と比べると，まだ十分な品質を達成しているとは言い難い．また，読み上げ調の音声にせよ，感情表現のようなその他の表現にせよ，音声のメディアとしての役割，つまり情報伝達に主眼を置いて研究されてきたと私たちは考えている．すなわち，発話内容，感情，発話者の個人性あるいは発話の意図など，音声が伝える情報を可能な限り正確に再現することを目的としてきたのがこれまでの音声合成研究の主流であった．もちろん音声のメディアとしての役割の重要性は言うまでもないが，音声の持つ役割はそれだけだろうか．例えば話芸について考えてみる．話芸は「落語・漫談・講談など,たくみな話術で人を楽しませる芸」であるという[2].話術とはすなわち話の仕方であるから，つまるところ音声によって観客を楽しませる芸ということになる．観客は演者の発する音声を聞いて楽しむことになるが，ここで音声は単なる情報伝達の役割だけでなく，観客の感情を引き出す役割を果たしていると言えよう．それでは，話芸が観客の感情を引き出すのはなぜだろうか．一つの要因は間違いなく発話内容で，面白いことを言っているから楽しいのである．ところが，同じ発話内容でも演者によって巧拙があるのは明らかだ．例えば，落語の同じ演目を，入門したての前座が演じた場合と，確固たる地位を築いている真打が演じた場合では，いくら発話内容が同じでも，観客の楽しみ具合には相当の差があるだろう．つまり，表現の仕方も要因となる．落語が話芸であることを踏まえれば，特に音声表現の仕方が重要であると言えるだろう．このように，人は音声を使って単に情報伝達をするだけでなく，その表現を駆使して他の人を楽しませることができる．

では，音声合成はどうだろうか．音声合成に落語を演じさせたものは，人手で調整をしたものを含めて，動画投稿サイトに投稿された例がある[3].こういった動画を見て楽しめるかどうかは主観や価値観に依存するだろうが，少なくとも真打の演技には遠く及ばないと考えてよいだろう．この隔たりに，音声合成研究で解決すべき問題があると私たちは考えている．本稿では，落語を演じる音声合成の検討を行う．話芸の中で落語を題材に選んだのは，漫談と異なり観客との相互作用が比較的少なく，従来のテキスト音声合成の枠組

みを利用できるからであり，一方で，講談と異なり基本的に登場人物の会話で物語を進めていくことから，従来の音声合成が得意とする読み上げ調から比較的遠い表現をすると考えたからである（落語が「ハナス」ものであるのに対し，講談は「ヨム」ものとされる）[4]．枠組みとしては，一人の演者による落語音声をもとにニューラルネットワーク(NN)の音響モデルを学習し，テキスト音声合成を行い，出力音声の品質を評価する．また，テキストから解釈することが困難な情報をコンテキストに加え，どの情報が品質向上に寄与するかを分析する．

2-2 落語について

落語は必ずしも馴染みのあるものではないので，落語そのものについて簡単に説明する．落語は江戸時代に成立した話芸で，大きく分けて東京落語と，関西を中心に発達した上方落語があるが，本稿で扱うのは東京落語である．演者は高座と呼ばれる舞台に敷かれた座布団に座って一人で話をし，基本的に登場人物の会話によって物語を進めていく．東京落語の場合，小道具は扇子と手拭いのみであり，楽器などは用いない．物語は，マエオキ，マクラ，本題，オチ，ムスビの 5 つの小単位からなる[4]．このうちマエオキは必須ではなく，ムスビはオチのない演目（噺と呼ばれる）や，噺を途中で打ち切る場合に出現する．オチはいわゆる物語を締めくくる部分のことで，落ちる，落とすということが落語の特徴とされている．また，噺には古典落語と呼ばれる概ね昭和初期までに成立したものと，新作落語と呼ばれる後の時代に成立したものがある．職業として落語を演じる者ははなしか噺家や落語家と呼ばれる．東京落語には噺家の身分制度があり，下から順に前座，二ツ目，真打と呼ばれる．

2-3 実験

(1) 落語音声データベースの構築

本実験のために，落語音声データベースを構築した．演者は柳家三三（落語協会所属，真打）で，音声収録用のスタジオで録音した．録音時にスタジオ内にいたのは演者一人で，観客および観客からの反応は一切なかった．録音した噺は東京落語における古典落語 25 演目である．言い間違いや言い直しなどについては，演者本人が希望した場合を除いて再録音を行わず，そのままデータベースに収録した．録音した音声に対しては，筆頭著者が書き起こしを施した．言い間違い，フィラーあるいは笑い声などに対しても特別な表記は定義せず，聞こえたとおりに仮名で書き起こした．フィラーについては，長音に聞こえた部分について，表記上長短の区別は付けなかった．文中で間を空けて発声された箇所のみに読点を記した．文末は句点で表記したが，文末が上昇調で発声された場合のみ疑問符を用いた．また，書き起こしの他に，文ごとにその特徴を示すラベルを付けた．ラベルの詳細を表 1 に示す．なお，part について，本来マエオキとムスビに相当する文は，それぞれマクラとオチとして扱った．なお，文を区切る基準は，ラベルが一意に定まる範囲で，以下のとおりとした．

- 文法的に文の区切りと判断できる場合で，直後に間がある場合
- 上昇調で発声された単語の直後

表 1 ラベルの詳細

グループ	名前	説明	詳細
ATTR（発話者の属性）	role	役	男，女；子供，若者，壮年，老人；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	個性	間抜け
COND（発話者の状態）	condition	状態	通常，甘え，呆れ，怒り，息切れ，苛立ち，何かを飲んでいる，驚き，悲しみ，我慢，恐怖，興奮，小声，困惑，寒がっている，焦り，何かを食べている，体調不良，得意，泣き，眠気，不快，酩酊，喜び，力み，笑い
SIT（発話者の置かれた状	relationship	話し相手との関係性	地の文，独り言，目上，目下

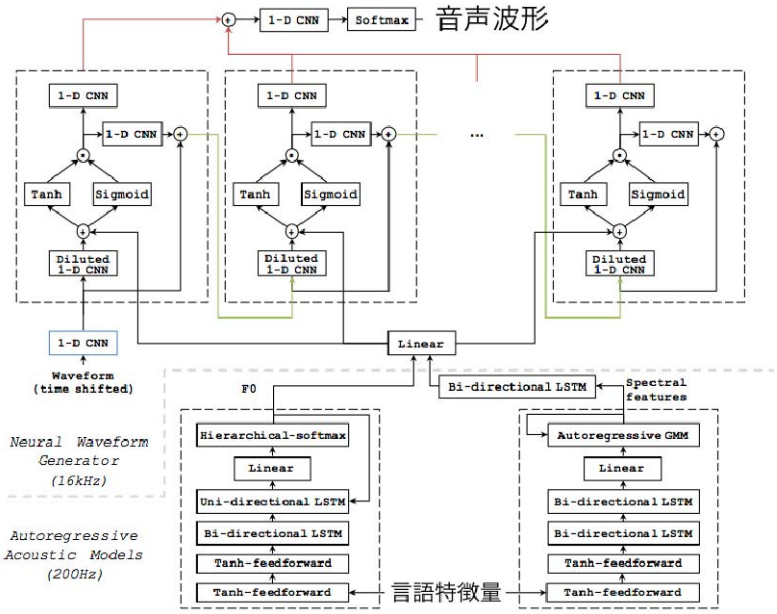
況）	n_companion	話し相手の人数	地の文，独り言，1 人，2 人以上
STR（物語の構造）	part	パート	マクラ，本題，オチ

(2) 使用した音声

本実験では，データベース整備の都合で，録音した 25 演目のうち 12 演目のみを使用した．合計時間は 6 時間 24 分 10 秒で，文数は 7,213 文であった．このうち 1 演目（298 文）をテストセット，他の 11 演目からおよそ 10 分の 1 の文を検証セット，残りを学習セットとした．

(3) モデルおよび特徴量

音響モデルは shallow autoregressive neural acoustic model (SAR)を使用した[5]．入力特徴量は quinphone，出力特徴量はメル一般化ケプストラム係数(MGC) 60 次元，基本周波数(F0)，voiced/unvoiced condition (VUV)，band aperiodicity (BAP) 25 次元を用いた．これらは読み上げ音声において、高品質な合成音声を生成可能であることが示されている。音素長の推定は行わず，自然音声から抽出した音素長を推定時に利用した．また，音響特徴量から音声波形を生成するために WaveNet ボコーダを使用した[6, 7, 8]．WaveNet ボコーダは、convolutional neural network による非線形自己回帰モデルであり、音声波形サンプルを直接予測でき、音声合成の肉声感を向上させることができる。なお，WaveNet ボコーダは[6]で使用した女性の（落語ではない）読み上げ音声で学習したものに基づいて，3.2 で述べた学習セットおよび検証セットによるデータを用いて転移学習を行った。これらを結合したネットワークの構造は以下の図の通りである。



なお、文と文の間のポーズは推定せず，一文ずつ合成した．

(4) 評価

聴取試験による主観評価は，今後の課題である．ただ，我々が試聴したところ，読み上げ音声とは明らかに印象が異なり，落語の雰囲気を十分に感じられるものとなっている．しかし、標準的な評価方法が決まっている読み上げ音声合成に比べ、落語音声合成の評価方法は全く過去に検討がなされておらず、指標および評価方法そのものを慎重に検討する必要がある。

2-4 おわりに

本稿では，人を楽しませる音声合成を目指し、落語音声合成用データベースの構築、および、必要なラベル情報の付与を行った．また、本データベースを利用して、深層学習による音声合成システムを構築した。今後の課題としては，落語音声合成の評価方法の検討、音素長の推定のほか，文と文の間のポーズの推定などが挙げられる．

【参考文献】

[1] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis and Yonghui Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," arXiv:1712.05884v2 [cs.CL], 2018.

[2] 新村出編, "広辞苑 第七版," 岩波書店, 2018.

[3] zky, 【初音ミク】ボカロ落語『野ざらし』, http://www.nicovideo.jp/watch/sm17066984, 2012.

[4] 野村雅昭, "落語の言語学," 講談社, 2013.

[5] Xin Wang, Shinji Takaki and Junichi Yamagishi, "An autoregressive recur-rent mixture density network for parametric speech synthesis," *Proc. ICASSP*, pp.4895-4899, 2017.

[6] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela and Junichi Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," *Proc. ICASSP*, pp.4804-4808, 2018.

[7] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda and Tomoki Toda, "Speaker-dependent WaveNet vocoder," *Proc. INTERSPEECH*, pp.1118-1122, 2017.

[8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior and Koray Kavukcuoglu, "WaveNet: a generative model for raw audio," arXiv:1609.03499v2 [cs.SD], 2016.

〈発 表 資 料〉

題　　名	掲載誌・学会名等	発表年月
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（発表予定）	日本音響学会 2018 年秋季研究発表会講演論文集	2018 年 9 月
音声合成は「落語」で人を笑わせられるのか？（取材協力）	ITMedia http://www.itmedia.co.jp/news/articles/1806/26/news016.html	2018 年 6 月