Vol.2019-MUS-123 No.30
Vol.2019-SLP-127 No.30
2019/6/22

音素単位で話速制御を行うGAN-TTS

岡本 真由子[1,a)]　Sakriani Sakti[1,2,b)]　中村 哲[1,2,c)]

概要：テキスト音声合成（TTS）は入力された文から音声を生成する技術で, 学習に公開されている読み上げ音声データが多用されている. そのため, 合成された音声が自然発話と異なり単調になり, 対話システムなど自然発話を想定したシステムに用いる場合に違和感が生じてしまう. 本稿では, 話速などの非言語情報を考慮した上で音声の自然性を損なわない音声合成の実現を目指す. 自然発話では, 発話速度の変化は単語や句の単位のみでなく, 任意の音ごとに変化することがあるため, 本稿では音素ごとに話速を制御するため, 各音素に対して, 敵対的生成ネットワークを用いた音声合成（GANTTS）[1] を用いて話速情報を付与する方法を提案する. また, 独自に話速の異なる音声データ 6,792 発話を収録した。その結果, 提案手法は合成音声の波形を人工的に操作するよりも自然で, かつ音素単位で適切な話速変化を行うことができることを示した.

Phoneme Level Variable Speaking Rate Control in GAN-TTS

1. はじめに

近年, 音声合成の自然性は著しく向上し, 幅広い分野で用いられている. しかし, その多くは, 読み上げ音声をデータとして学習したものであり, 自然会話と比べ話し方が単調なため不自然に感じることがある. 文章の読み上げ音声と自然会話の間には様々な差異があり, その例として, エントレインメントが挙げられる. これは, 会話において相手の発話における言語的あるいは音響的な特徴に同調するものであり, その同調度合いが高いと対話に対して良い印象を抱くことが多いことが知られている [2]. このような, 相手を考慮するなどしその場に応じた適切な音声を出力することは, 自然な対話を構築する上で重要である. しかし, 従来の音声合成では, 入力テキストに対し最も自然な音声を出力するため不可能である. 一方で, 音素ごとに話速を制御するための音声合成システムも提案されている [3]. この手法では,Tacotron[4] を使用し very slow(150ms),slow(110ms),fast(70ms),very fast(150ms) の 4 種類の話速で音声を合成できる. しかし, この手法では, エントレインメントのことは考慮しておらず, ただ入力された継続長に基づいて音声を生成する. また, 話速の異なる自然データを学習に用いていない.

より自然な対話のためには, 話速, 声の高さ, 声の大きさなどの非言語情報を自然に, かつ相手に合わせて柔軟に表現することが必要である. 本稿では, 話速の異なる 6,792 発話の音声データを独自に収録した。本稿では少量のデータで高品質な音声合成を実現するため GANTTS を用いた. また, 発話内で音素ごとに話速を制御するために, 各音素に対して話速情報を付与する方法を提案する.

2. 敵対的生成ネットワークを用いた音声合成

音声合成の手法には, 隠れマルコフモデル (HMM), ディープニューラルネットワーク（DNN）などを用いた手法がある.HMM 音声合成は, 少ない学習データで高速に動作する反面,DNN 音声合成と比べ音質は低い.DNN 音声合成は近年急速に発展した技術であり, 波形生成手法である WaveNet[5] や, 文字から直接音響特徴量を生成できる Tacotron の提案によって, 容易に自然な音声を合成することができるようになった. しかし,DNN を用いた手法の欠点として, 学習に多くのデータが必要になるということが挙げられる. 話速を制御できる音声合成モデルを学習するために, 様々な話速の音声が含まれたデータが必要となる. しかし, これらのデータを大量に集めることは困難である. そこで, 本稿では比較的少ないデータでも学習を行える, 敵

[1] 奈良先端科学技術大学院大学
NAIST, Takayama-cho, Ikoma, Nara 630-0192, Japan
[2] 理化学研究所 革新知能統合研究センター,AIP
[a)] okamoto.mayuko.oi1@is.naist.jp
[b)] ssakti@is.naist.jp
[c)] s-nakamura@is.naist.jp

対的生成ネットワーク (GAN)[6] を用いた音声合成である GANTTS を用いることとした. GANTTS には, 生成器と識別器という二種類のニューラルネットワークがあり, 生成器は, 自身の生成結果を識別器が自然音声だと誤認識するような出力を生成できるよう学習し, 識別器は自然音声と生成器で生成された合成音声を正しく見分けられるように学習する.

3. 提案手法

音素単位での話速制御を行う GANTTS を実現するためには,GANTTS に話速の情報を入力する必要がある. 本稿ではこの話速情報の与え方を 2 手法提案する. 手法 1 は, 言語特徴量に含まれる音素記号を話速情報を元に拡張した. 例として,“ae” という音素がある時, 遅い発話では Slow を意味する S を音素の末尾に付与し “aeS”, 通常速度では Normal を意味する N を音素の末尾に付与し “aeN”, 速い速度では Fast を意味する F を音素の末尾に付与し “aeF” とする. 付与した “S”,“N”,“F” を本稿では話速タグと呼ぶ. 手法 2 は, コンテキストラベルの各音素に対する話速情報を通常音声を 100 とした際の比率で置き換え付与する. こうすることでどの音素も入力された四種類の継続長で直接出力する先行研究 [3] と比べ, 各音素で異なる最適な継続長を生成する. 遅い発話では 75, 通常発話では 100, 速い発話では 125 を新たな特徴量として与えた. 両手法において, 無音区間を意味する “pau” は, 話速に関わらず生成される音響特徴量に変化が無いため, 本稿では話速を考慮しない.

4. 話速の異なる音声データの構築

今回, 提案手法の検討のためデータセットを収録した. 本稿では, はじめに CMU ARCTIC[7] の単一話者の 1132 発話を元に, その話速を Normal とし,SoundExchange (sox)[*1] を用いて Normal の話速を 0.75 倍速に変化させた発話を Slow, 1.25 倍速に変化させた発話を Fast として人工的に話速を変化させたデータを作成した。次に, 女性 1 名, 男性 1 名の被験者に対して, 人工的に話速を変化させたデータを聞いてもらった上で,3 種類の速さで発話してもらった同じ内容の自然音声 (44100Hz,16bit) を収録した. これら,6,792 発話 (2 話者,3 種類の話速) をモデルの学習・評価に用いた.

5. 分析

本データセットの傾向を確認するため, いくつかの分析を行った. この分析で用いたデータは,4 章で男女 2 話者の協力の元収録した 3 種類の話速の自然音声と, その自然音声の内,Normal 話速の発話を sox を用いて話速を 0.75 倍,1.25 倍に人工的に変化させた音声を加えた 5 種類の発話について分析した. まず, 各音声の発話長を比較するため, 各話速

[*1] https://www.soundexchange.com

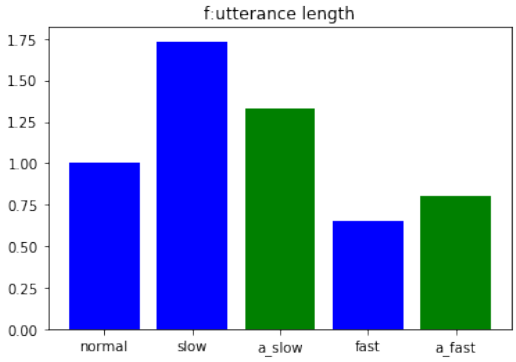

図 1 全体発話の継続長比 (女声)

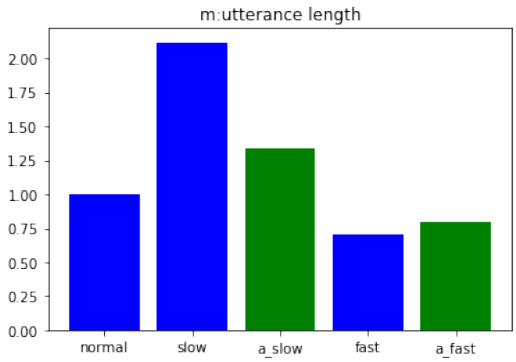

図 2 全体発話の継続長比 (男声)

における平均の発話長を示した図 1 (女声) と図 2 (男声) を示す. これらの図では Normal の発話全体の継続長を 1 とした場合に対し Slow と Fast の発話の継続帳を比率で表している. 尚, 人工的に変化させた Slow を “a_slow ” ,Fast を “a_fast” とした.

人工的に変化させた遅い発話である a_slow より自然音声である Slow の方が, より平均発話長が長く, 人工的に変化させた速い発話である a_fast より自然音声である Fast の方が, より平均発話長が短い.

続いて, 各音素の継続長を比較する. 話速が変化した場合, 母音と子音において, その継続長の変化がそれぞれ異なった挙動を示すかどうかの分析を行った. 図 3,4 (女声) と図 5,6 (男声) に母音と子音の平均継続長を示す. ここでも先述と同様に Normal の話速を 1 とした比率で表している.

これらの図から, 話速の変化によって, 母音, 子音とも平均継続長が変化するが, その変化はほとんど同じ傾向を示すことがわかった.

最後に, 音素のパワーを比較する. 各音素の平均パワーを示した図を, 図 7 (女声) と図 8 (男声) に示す. ここでも同様に Normal の話速のパワーを 1 とした比率で表している.

ここでは, 各話速においてパワーの変化に大きな差は見られない. 読み上げ音声においては, 話速の変化によって音量の変化は引き起こされないと言える.

今回の音声収録において, 話速の違う音声を用意し被験者に対してそれらの音声に対してエントレインメントを促

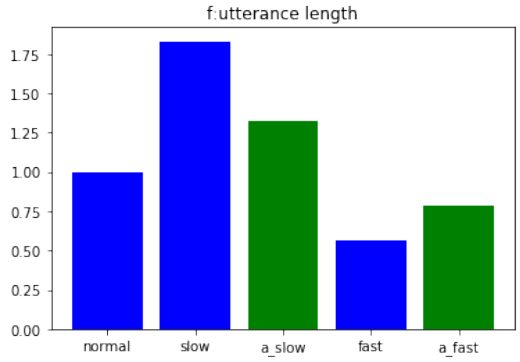

図 3 母音の平均継続長比 (女声)

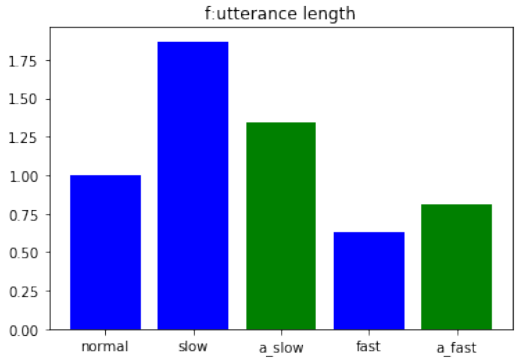

図 4 子音の平均継続長比 (女声)

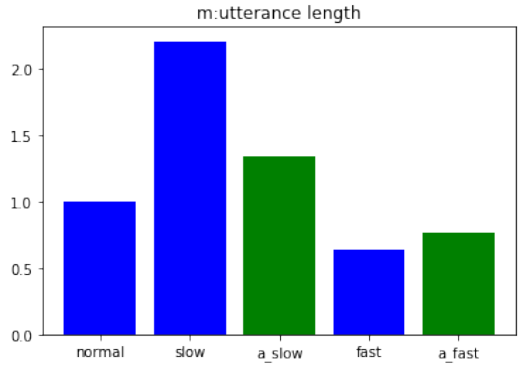

図 5 母音の平均継続長比 (男声)

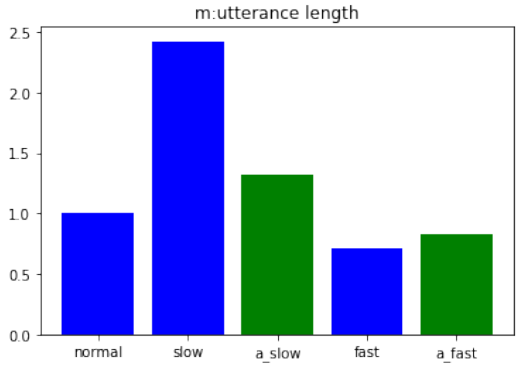

図 6 子音の平均継続長比 (男声)

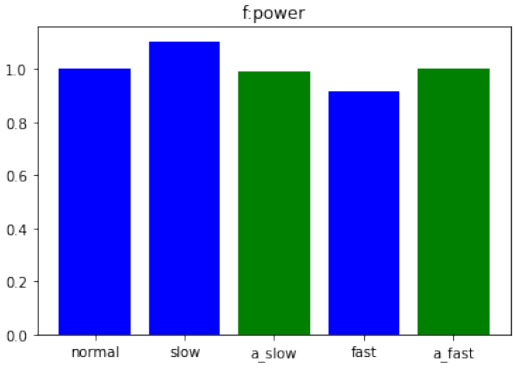

図 7 音素の平均パワー比 (女声)

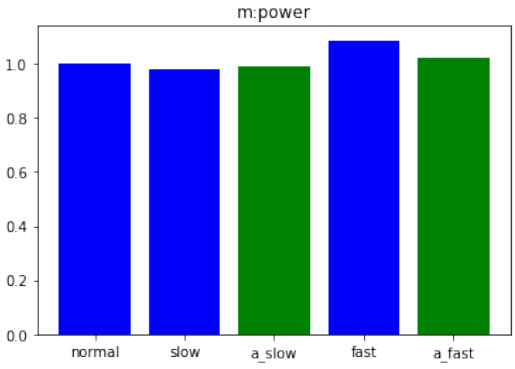

図 8 音素の平均パワー比 (男声)

した. その結果得られた音声は与えた音声の話速より, より短くまたは長く変化していることがわかった. また, 話速の変化によって母音と子音の平均継続長の変化傾向に違いがないことから, 英語音声において話速を制御する際, 母音と子音を区別して扱う必要がないと考えられる. さらに. 話速の変化によりパワーの変化は起こらないことから, 話速制御の際にパワーを考慮する必要が無いことを確認した.

6. 実験

HMM/DNN-based Speech Synthesis System (HTS)[8] を用いて入力文からコンテキストラベルを作成した. 得られたコンテキストラベルには, 現在および周辺の音素などが含まれる. 続いて, このコンテキストラベルに, 提案手法で述べた 2 種類の方法でそれぞれ話速情報を付与する.

本稿で提案した手法が有効なものであるかを確認するために,GANTTS に対して, 話速をもとに音素ラベルを拡張する方法（提案手法 1）, 話速情報を連続値として新しい特徴量を追加する方法（提案手法 2）に対し, ベースラインとしてコンテキストラベルのみで学習したモデルに対して sox を用いて話速を変化させた手法を比較した. 本稿では比較のため AB テストを用いた. sox で変化させる割合は,5 章で得られた, 自然音声間の比率を用いた. そのため, 合成された 3 手法の音声の話速とほぼ等しくなっている. 初めに, 音声の自然性に対して主観評価を行った. 被験者

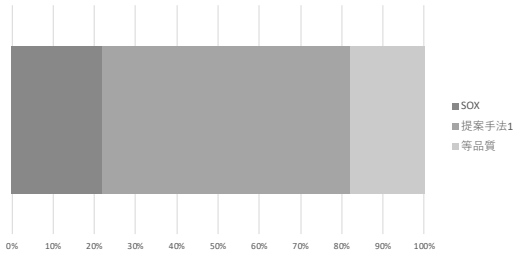

図 9 ベースラインと提案手法 1 の自然性 ABX 主観評価

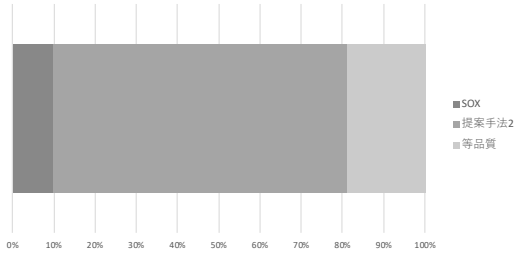

図 10 ベースラインと提案手法 2 の自然性 ABX 主観評価

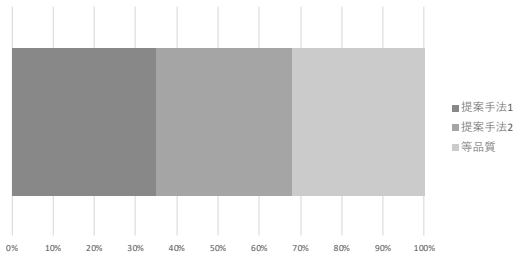

図 11 提案手法 1 と提案手法 2 の自然性 ABX 主観評価

に 2 つの音声を提示し, どちらの音声が自然であったかを回答してもらう. この音声対は, 再生順序において全ての手法において偏りがない.

図 9, 図 10 より, 提案手法 1,2 両方においてベースラインより自然であると言える. 図 11 より, 提案手法間では合成音声の品質に差は生じていない.

次に提案手法において, 音素単位で話速の制御が効果的に行えているかを確認するために,1 文章の中の特定の単語, フレーズのみに対し話速制御を行った結果を評価する. 評価方法として, 人手による強調認識評価を行った. 自然発話において会話の強調される部分においてパワーと話速が大きく変化する現象が存在する. そこで本稿は 1 文内で大きく話速が遅くなる部分を作り出し, その部分が強調しているように聞こえるか評価した. 正しく話速が制御されているならば, 該当箇所のみが強調されているように聞こえ, 部分的な話速制御が出来ていないならば強調箇所はわからない.

本稿では特定のフレーズのみを Slow とし, 残りの部分を Fast とした音声を提案手法で作成し,Normal 話速の音声,fast 話速の音声の 3 手法において, それぞれ AB テストを用いて比較を行った. また, 話速情報の与え方において,2

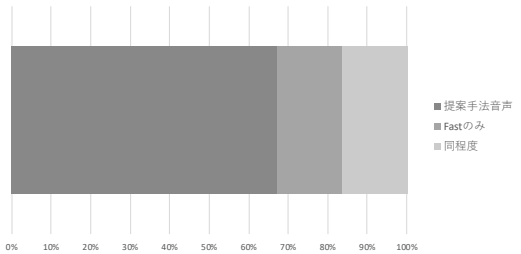

図 12 提案手法音声と Fast のみの自然性 ABX 主観評価

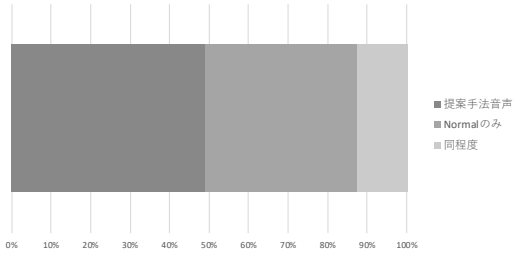

図 13 提案手法音声と Normal のみの自然性 ABX 主観評価

つの提案手法の間に品質の差が無いことから, 本実験は音素拡張を用いる方法にて行った. 被験者に対し, 強調箇所が太字になっている発話文を与える. 被験者は音声を聞いて, より強調されて聞こえる音声を回答してもらう.

図 12 より, 文章全てを Fast で生成した合成音声との比較では, 提案手法音声の方が, より特定フレーズが強調されて聞こえるとの回答が多い. これにより, 提案手法音声は文章中で話速を適切に変化させることができ, その結果文章中の特定フレーズを強調するような発話を行うことができたと言える. しかし図 13 より, 提案手法音声と Normal 音声間では差が無かった. これは,Normal 音声は自然発話よりも話速が遅く, 文章全体が強調されて聞こえたため, 提案手法との差が出なかった.

7. まとめ

本稿では, 発話文内で自由に話速の制御を行うことのできる, 音素単位での話速制御を行う GANTTS について提案した. この GANTTS を学習させるためのデータセットの構築および分析, そして提案手法の有用性の検証のために 2 種類の実験を行った. その結果, 提案手法は合成音声の波形を人工的に操作するよりも自然で, かつ音素単位で適切な話速変化を行うことができることを示した.

謝辞 本稿の一部は JSPS 科研費 JP17H06101 および JP17K00237 の助成を受けたものです.

参考文献

[1] Saito et.al., Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks 入手先 ⟨https://ieeexplore.ieee.org/abstract/document/8063435⟩ (2019.05.29)
[2] R. Nishimura, N. Kitaoka, and S. Nakagawa (2009).

Analysis of Factors to Make Prosodic Change in Spoken Dialog, Journal of the Phonetic Society of Japan, vol.13, No.3
[3] Park et.al., Phonemic-level Duration Control Using Attention Alignment for Natural Speech Synthesis, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
[4] Wang et.al., Tacotron: Towards End-to-End Speech Synthesis 入手先 ⟨https://arxiv.org/abs/1703.10135⟩ (22019.05.29)
[5] Oord et. al., WaveNet - A Generative Model for Raw Audio 入手先 ⟨https://arxiv.org/abs/1609.03499⟩ (2019.05.29)
[6] Goodfellow et.al., Generative Adversarial Networks 入手先 ⟨https://arxiv.org/abs/1406.2661⟩ (2019.05.29)
[7] CMU ARCTIC, 入手先 ⟨http://festvox.org/cmu_arctic/⟩ (2019.05.29).
[8] HTS. 入手先 ⟨http://hts.sp.nitech.ac.jp⟩ (2019.05.29).