人を楽しませるための雑談対話システム向け笑い声合成

Laughter synthesis in a chat dialogue system to entertain people

三輪桃子 (Momoko Miwa)*

法政大学 情報科学部 ディジタルメディア学科
momoko.miwa.2k@stu.hosei.ac.jp

Abstract

A method for synthesizing laughter in a chat dialog system aimed at entertaining people is proposed. The conventional WaveNet method did not generate laughter with intonation, and sometimes generated laughter that was mostly unvoiced. In this study, the TTS system, VOICEVOX, is used to generate voiced laughter. In addition, speech transformation using the VAE model is used to enhance naturalness. OGVC is used because it is assumed that the training data contains a lot of laughter. Listening experiments were conducted to evaluate the naturalness of laughter. Furthermore, an experiment was conducted on the dialogue when laughter was inserted into the chat dialogue system to evaluate the excitement of the dialogue and the willingness to chat. The MOS of the original voice was 4.6, the MOS of laughter using the conventional method WanveNet was 1.8, and the MOS of laughter using the proposed method was 1.1. The dialogue excitement rating was 3.0, and the willingness to dialogue 2.3. The MOS of laughter using the proposed method was smaller than that of the original voice but close to that of laughter using WaveNet. The proposed method can generate natural laughter voices equivalent to those of conventional methods.

1 はじめに

医療福祉分野では，認知症予防やうつ病予防の手段としてカウンセリングや傾聴が有効とされている．しかし，少子高齢化による人手不足や，必要な時に身近に，かつ気軽に傾聴してもらえる相手がいるとは限らないという問題がある．これらに対して気軽に傾聴してもらえる傾聴対話システムが提案されている [1]．傾聴対話システムと同じく，雑談対話のような達成すべきタスクを設定しない非タスク指向型対話システムも多く提案されている．Starley 株式会社が開発した音声会話型おしゃべり AI アプリ Cotomo では笑い声が挿入されているが，笑い声が棒読みであり場面によっては馬鹿にされていると感じてしまうこともある．そのため対話システムとより楽しい会話が成り立つような笑い声が必要である．本研究では雑談対話システムに笑い声を挿入する目的で，自然な笑い声が合成できる手法を提案する．

本研究では音声合成を用いた雑談対話システムに着目し，そこで笑い声を挿入する場合の適切な笑い声の合成方法を検討する．雑談対話システムの発話に自然な笑い声を混ぜることで，機械と対話しているという意識が減り，対話がよりスムーズになったり，盛り上がったりすることが想定される．本研究で行うこととして，汎用性のある笑い声の数種類の合成を目指す．合成する笑い声が感情を読み取ることのできない機械らしいものでは，対話システム挿入評価時に合成音の品質に意識が向いてしまい，笑い声の挿入結果についての評価に支障をきたす可能性がある．そのため合成する笑い声は人間の笑い声に近い発音やイントネーションが再現できていることを目指す．本研究では作成した笑い声がより人間らしく自然であるか，笑い声を挿入したことによる雑談対話システムとの対話の盛り上がりや時間に変化がみられるか，2 つの評価を行う．

2 関連研究

そもそも笑い声の合成は難しいという問題がある．理由として，通常の TTS(Text to Speech) システムでは笑い声と発音辞書が異なるために笑いを表現しきれないこと，笑い声を収集することが難しく，学習用データが少なくなることが挙げられる．

有本ら [2] は膨張たたみ込みの積層に基づく自己回帰ニューラル波形モデルである WaveNet を使用して笑い声を合成している．主観評価である Mean Opinion Score（MOS）の結果は原音声が 4.0 に対し，生成された笑い声は 2.8 であった．無声音が多く含まれている笑い声が生成されていたため，雑談対話システムに挿入するには汎用性が少ない．

また，Tits ら [3] は雑談対話システムの対話内容にそのまま追加が可能である Seq2Seq を用いて笑い声を合成した．学習には通常発話 (150.5 分, 3299 発話) と演技音声 (1 話者分)

* 指導教員：伊藤克亘 教授

を使用している．MOS は原音声が 4.0 に対して 3.0 を獲得していた．しかし，使用する自作の TTS モデルは学習に多くのデータが必要であり，Tits らによって生成された笑い声は日本人の笑い声と音色が異なるという問題がある．

そこで本研究では学習する必要のない既存の TTS システムを使い，有声音の笑い声を生成する合成方法を提案する．

3　笑い声の合成

3.1　概要

雑談対話システムに笑い声を挿入するためには，どの笑いを挿入するのかという笑いの種類，笑い声の自然性をどのように保つのか，笑い声を挿入するタイミングの指定方法，雑談対話システムが対話の流れを把握する方法などを考慮すべきである．その中でも本研究では笑いの種類と笑い声の自然性に着目する．人間の笑いには笑い声単体と，笑いながら発話する喋り笑いが存在する．本研究では雑談対話システムの発話内容にそのまま追加が可能である笑い声単体を合成する．また，合成する笑いの種類について，雑談対話システムでの対話を楽しませることが目的であるため，快の笑いに限定する．笑い声の自然性は適切な合成手法を考えることで再現する．

想定する雑談対話システムでの笑い声挿入方法は，あらかじめ作成した笑い声辞書から笑いを選択し，対話文の前後に追加するというものである．これにより，挿入する笑い声の適切な種類とタイミングが把握できれば容易に笑い声を挿入することができるようになる．また，笑い声辞書の見出し語は笑い声のテキスト (例えば，「ははは」など) とする．

本研究では笑い声の合成手法として入力をテキストと長さ，出力を音声とする Seq2Seq のモデルを提案する．笑い声合成実装モデルのブロック図を図 1 に示す．笑い声の合成には，雑談対話システムに挿入しやすく有声音の多い笑い声を生成できるという理由から，既存の TTS システムを使用する．笑い声の構成要素ごとに音声を作成し最後に結合する手法では笑い声の構成要素の境目が顕著であったため，本研究ではよりひとまとまりの笑い声に聞こえることを目指す．作成する笑い声の人種は日本人であり，その中でも個人性を必要としない，汎用性のある笑い声を数種類合成する．TTS システムによって生成された合成音声をより原音声に近づけるため，音声変換を行う．通常の音声変換モデルの学習には，発話内容は同じであるが音色の異なるペアの音声ファイルが必要となる．しかし，笑い声は個人差があり，ペアとなる音声が存在しない．そこで本研究では音声変換モデルには目標音声のみを使用して訓練できる VAE モデルを使用する．VAE モデルの入力と出力にはスペクトルが使用される．VAE モデルから出力されたスペクトルから，音声波形を再合成する．

3.2　使用するデータセット

本研究では笑い声合成モデルの学習データ，評価用の笑い声の原音声として感情評定値付きオンラインゲーム音声チャットコーパス（OGVC）を使用する．これはヘッドセットマイクで録音されたオンラインゲームプレイ中の会話であり，話者 13 名，合計 9114 発話が収録されている．通常のコーパスと比較して笑い声が多く含まれていると考え採用する．収録音声は自発対話音声と演技音声であるが，本研究では自発対話音声を使用する．

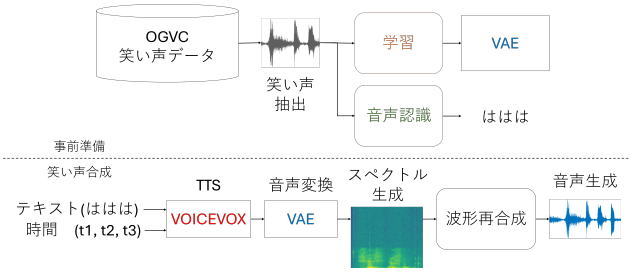

図 1　笑い声合成モデル

OGVC 話者の中でも，有本らが使用した男性話者 04_MSY の音声を使用する．

3.3　笑い声のテキスト化

まずはじめに TTS システムを使用するため笑い声をテキスト化する必要がある．OGVC の話者ごとの音声ファイルを，笑い声が登場する音声ごとに分割する．笑い声の判断には OGVC に元から含まれている発話テキストファイルで「{笑}」が含まれている箇所を用いる．分割された音声ファイルを音声認識システム Whisper を用いて笑い声の部分がテキスト化されるか試す．笑い声がテキスト化されたものはそのまま TTS システムに使用し，テキスト化されなかった笑い声は Montreal Forced Aligner(MFA)[4] を用いて笑い声の区間を推定する．Whisper を用いてテキスト化を図る．この際，通常の Whisper では笑い声の読みに対応した文字列にテキスト化されず，「笑」や「www」と認識されることがある．そのため，Whisper のファインチューニングを行った．

Whisper の tiny モデルでファインチューニングを行う．学習には OGVC の話者 04_MSY の笑い声を使用し，一度音声認識をかけた結果と実際に音声を聞いてテキスト化したものを正しいラベルとしたものを使用して学習する．その結果の文字誤り率を測定する．以下に OGVC の発話転記テキストファイルの内容サンプルを示す．数字は発話番号であり，テキストは発話内容である．

```
008，あーのランドジュー
009，うん
010，あ早いな。、 {笑}
011，俺まだ入ってねえよ。、{笑}
```

ファインチューニングを行った Whisper モデルに同じ音声を通すと，笑い声の部分が以下のように認識される．

```
008，あーのランドジュー
009，うん
010，あ早いな。、 ははは
011，俺まだ入ってねえよ。、ふふっ
```

OGVC の笑い声の認識結果は 7 カテゴリー 55 種類に分類された．

3.4　笑いの長さのモデル化

TTS システムの入力として，構成する笑いの音素の長さを指定する．入力する長さは OGVC の笑い声から統計をとり，モデル化する．

認識された音声ファイルとテキストファイルから MFA を用いることで音素の開始時間と終了時間が分かる．認識結果を得た音声を対象に全ての音素の開始時間と終了時間を記録する．音素の開始時間から終了時間までの長さを音素の継続時間とする．それぞれの音声ファイルに対して，笑い声全体の長さから音素の継続時間の割合を計算し，認識結果ごとに平均を取ることで，笑いの長さのモデル化を行う．

3.5　VAE

次の事前準備として音声変換に使用する VAE モデルの学習を行う．学習データセットには OGVC の笑い声ファイルを使用する．使用する VAE モデルは Valerio による generating sound with neural networks モデル [5] である．これを使用することで，OGVC の笑い声を学習し，TTS システムの生成した笑い声を OGVC 話者の笑い声へ変換できる．

VAE は入出力がスペクトログラムであり，音声波形を生成できない．そのため，波形再合成を別の手段で行う．本研究では振幅スペクトログラムから位相スペクトログラムを復元する手法である Griffin-Lim 法を採用した．

4　評価

4.1　笑い声の評価

作成した笑い声が雑談対話システムに導入する上で不自然でないかを評価する．生成された音声を従来手法である WaveNet を使用した笑い声・TTS システムで合成した笑い声・原音声と比較する．聴取実験では有本らの評価手法に一部準じて Mean Opinion Score(MOS) を使用することで，同じく MOS の評価方法を採用した従来研究と結果を比較・考察できるようにする．聴覚を使用しない評価方法として，コーパスの笑い声と合成された笑い声の波形やスペクトログラムを比較する．ただし，人間の聴覚システムの特性が音質に大きな影響を与えるため，聴覚による笑い声の評価を最重要視する．

聴取実験では作成した 3 種類の笑い声をノートパソコンにつないだヘッドフォンから再生し, 被験者に聞いてもらうことで主観評価を行う．被験者数は国際的な品質評価試験では通常 24 名以上の評価者を必要とするため，25 人を想定する．

4.2　対話の評価

雑談対話システムに笑い声を挿入したときの．ユーザーとシステムの対話について評価を行う．雑談対話システムに笑い声を挿入する方法として，そのシステムが完成しているかのように人間が動かす Wizard of Oz 法を用いる．雑談対話システムと通常の対話を行い，システムの笑い声挿入は人手で行う．生成する対話や笑い声を挿入するタイミングは実際の対話の録音の使用して作成する．

笑い声を挿入していない対話，録音した笑い声を挿入した対話，TTS システムの笑い声を挿入した対話，本研究で合成した笑い声を挿入した対話をそれぞれ，評価する．対話内容の評価は被験者に録音を聞いてもらったアンケートによる主観評価によって行う．評価軸は対話が盛り上がったと感じるかという「対話の盛り上がり」と，雑談対話システムと対話を続けたいと思うかという「対話意欲」である．これらの評価軸は対話システムに笑い声を挿入したときの評価として，重要な指標になると考え採用した．これらを最も悪い評価を 1，最も良い評価を 5 としたときの 5 段階で評価を行い，被験者の平均をとる．再生する対話の順番が結果に影響すると考え，順番はランダムに設定する．被験者は笑い声の評価と同じく 25 人を想定する．

4.3　実験

VAE モデルは OGVC の全話者の笑い声単体が収録された 1588 ファイルを 150Epochs 学習した．入力となるスペクトルのサイズを決める，サンプリング点数 (nfft) は 512，ホップ長は 128，音声の長さは 0.74 秒，サンプリング周波数は 22050Hz である．図 2 はモデルの学習中の Loss の推移をグラフにしたものである．Reconstruction Loss は平均二乗誤差による Loss であり，KL Loss は VAE（Variational Autoencoder）の潜在空間の分布が標準正規分布に近づくようにするためのペナルティを計算している．Loss は Reconstruction Loss と KL Loss を組み合わせた損失関数を計算している．

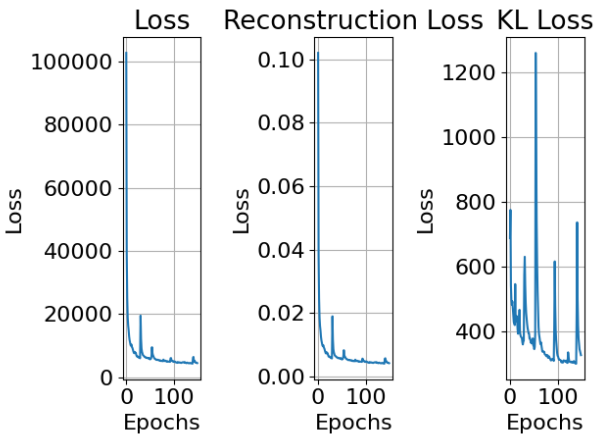

図 2　モデルの Loss の推移

本研究では笑い声に聞こえやすく，収録した笑い声の認識結果の音声である「ふふっははっ」,「ふふっ」,「ふははっ」の 3 種類のテキストから音声を生成した．長さは収録した原音声になるべく近づけた．また，本研究で使用した TTS システムは VOICEVOX である．

対話の評価実験用に作成した対話の内容を以下に示す．録音

の中で男性話者が短い時間の間に笑った回数が多く含まれているシーンを対象にした．男性話者の発話は VOICEVOX に変更している．V は VOICEVOX による発話，H は録音した女性の発話 (Human) を表している．対話は約 40 秒の音声ファイルになっており，VOICEVOX の発話である 3 か所の laugh を変更している．挿入した笑い声の種類は上から順に「ふふっはははっ」，「ふふっ」，「ふははっ」である．

V：え、ちなみにさ、何が一番好きなの？食べ物
H：うーん、え、甘いもの？それとも普通の甘くないやつ？
V：え、いやノルムノルム
H：ノルム？
V：ノルム
H：ノルムってなんだよ
V：{laugh}
H:{laugh}
V：総合値、総合値
H：{laugh}、総合値、ノルムってそうだっけ？
V：{laugh}、え、フロベニウスノルム。
H：うーん、えっとねー
V：{laugh}ガンスルー
H：マ、マ、マンゴーかな
V：ああはい、はいどうぞ
V：おー、マンゴー。私もねマンゴー好き
H：マンゴー好き？
V：うん、常夏のマンゴー
H：{laugh}常夏のマンゴー、えなんで
V：ココナッツじゃない、常夏、常夏
H：常夏ねそうだよ
V：そうそう

笑い声と対話の評価は 20 代〜50 代までの男女 30 名によって行われた．評価には音声再生方法の指定は行わず，Google フォームによるアンケート形式を採用した．

4.4　笑い声の主観評価結果

生成された笑い声はどの音声も男性話者が発声したように聞こえる笑い声となった．また，機械的な音声が混じっていることに加え，基本周波数が変化しないように聞こえる，例として「ふははっ」生成された笑い声のスペクトログラムを図 4 に示す．比較のため，OGVC 男性話者 04_MSY の「ふははっ」に聞こえる笑い声のスペクトログラムを図??に示す．0.25s〜0.30s あたりのスペクトログラムは濃い黄色部分の倍音構造がつぶれていることが分かる．この部分が強く聞こえるため，基本周波数が変化しないよに聞こえている．

生成された笑い声の聞こえ方について比較する．従来手法と比較した場合，無声音より有声音が多く含まれることで声であると認識しやすくなっている．VOICEVOX と比較した場合 VAE を用いて音声変換を行うことで，有声音の笑い声に加えて鼻からもれる息のような無声音が再現され，より人間らしい声質に変化している．しかし，VOICEVOX の方がよりクリアな発話であり，対話システムが笑ったと音声だけで判断しやすい．最後に原音声と比較した場合，VOICEVOX と同じく原音

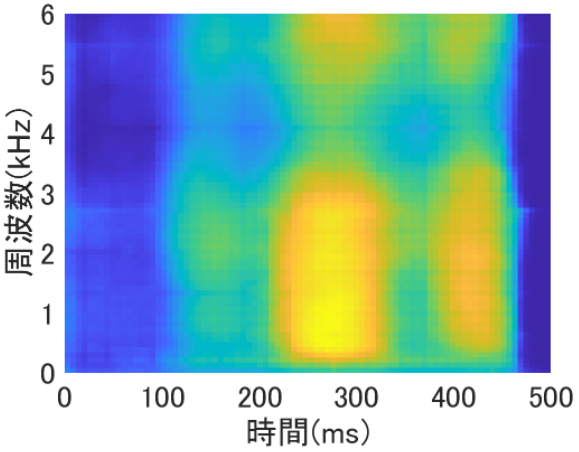

図 3　「ふははっ」を入力して生成されたスペクトログラム

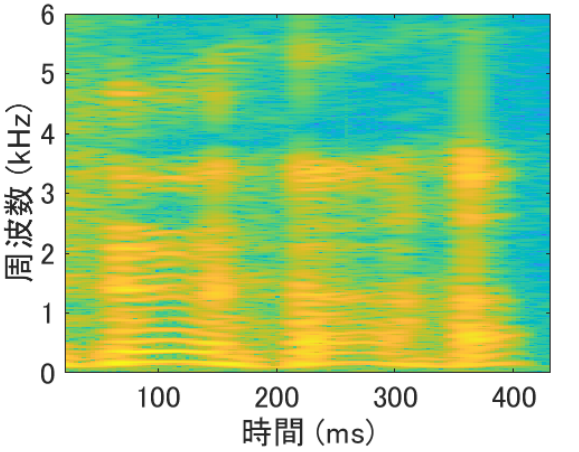

図 4　OGVC 話者 04_MSY の「ふははっ」に聞こえる笑い声のスペクトログラム

声の方がクリアな発話であり，かつ笑いの種類も多くあるため原音声の方が笑い声であると判断しやすい．また，基本周波数の変化は従来手法，提案手法，VOICEVOX，原音声の順で大きくなっている．

4.5　笑い声の聴取評価結果

生成された笑い声の聴取実験の結果を表 1 に示す．原音声は評価用対話の録音で得られた男性話者の実際の笑い声である．VOICEVOX の声は玄野武宏によって生成されたものである．VOICEVOX1 は「ふふっはははっ」を入力して出力されたものであり，提案手法 1 は VOICEVOX1 を VAE モデルで音声変換したものである．同じように，VOICEVOX2・提案手法 2 は「ふふっ」，VOICEVOX3・提案手法 3 は「ふははっ」によって作られたものである．

原音声が最も高く 4.02 という評価になった．VOICEVOX は平均値が 2.13 であり，VOICEVOX のなかで最も評価が高いものは VOICEVOX2 の 2.26 である．提案手法は平均値が

1.39 であり，提案手法の中で最も評価が高いものは提案手法 2 の 1.43 である．また，WaveNet の MOS の平均は 2.46 である．原音声の平均値と提案手法の笑い声の平均値の差は 2.63 である．VOICEVOX の平均値と提案手法の笑い声の平均値の差は 0.74 である．WaveNet の平均値と提案手法の笑い声の平均値の差は 1.07 である．

表 1　笑い声の MOS の平均値

laugh	MOS
原音声 1	3.66
原音声 2	4.43
原音声 3	3.97
VOICEVOX1	1.97
VOICEVOX2	2.26
VOICEVOX3	2.17
提案手法 1	1.40
提案手法 2	1.43
提案手法 3	1.34
Wavenet1	2.31
WaveNet2	2.60

4.6　対話の聴取評価結果

笑い声を挿入したときの対話の評価を表 2 に示す．全て対話の録音を聞いた被験者による評価の平均値を導出している．対話の盛り上がりと対話意欲のどちらの評価においても原音声の笑い声を挿入している時の対話が一番高い評価となった．反対に，笑い声を挿入していない時の対話はどちらの評価においても最も低い評価となった．提案手法の笑い声があるとき，対話の盛り上がり評価は VOICEVOX と同じく 2.66 であり，対話意欲は 2.35 が得られた．

表 2　笑い声挿入時の対話の評価

対話に挿入した笑い声	盛り上がり	対話意欲
笑い声なし	2.60	2.34
原音声	3.31	2.86
VOICEVOX	2.66	2.49
提案手法	2.66	2.35

4.7　考察

生成された笑い声は倍音構造がつぶれてしまい基本周波数が原音声に比べて変化していないように聞こえる．これは通常発話と比べて笑い声は発音のバリエーションが多く，VAE の潜在変数が対応しきれていないためである．また，生成された音声は人間が発声しないようなノイズによって機械的な音声が強く聞こえる．これは生成されたスペクトルから音声に戻す部分がうまくいっていないためである．現在は Griffin-Lim 法を使用して波形を復元している．しかし，Griffin-Lim 法では笑い声の位相が上手く予測されないため，原音声に比べて基本周波数の変化が少ない笑い声が生成されてしまう．これを解決するには別の手法で波形を復元する必要がある．また，従来手法や VOICEVOX と比較して提案手法の音声は波形再構築時の位相のずれによる機械的な音声のために，自然な笑い声であると評価されにくい音声となってしまっている．

笑い声を挿入したときとしていないときの対話について，盛り上がりと対話意欲に差が出た．笑い声を挿入していない時の対話はどの笑い声を挿入したときと比べても低い値となったことから，雑談対話システムの笑い声挿入は種類に関わらずユーザーの満足度向上に必要不可欠であると言える．加えて MOS の高い原音声を挿入したときの結果が最も良かったことから，自然な笑い声の挿入は対話の盛り上がりや対話意欲の向上につながる．提案手法による笑い声は対話システムから発声されたものだと分かるものの，機械的な音声が含まれているために不自然に聞こえてしまい，評価が低くなった．

5　まとめ

本研究では，自然な笑い声の適切な合成手法について提案した．従来の研究とは異なり，雑談対話システムに挿入しやすく有声音の多い笑い声を生成できるという理由から，既存の TTS システムを使用して音声を生成する手法を取り入れて研究を行った．その結果，笑い声の生成は可能であるが，基本周波数が変化していないように聞こえる音声が生成されるという結果を得られた．また，評価方法を用いることで雑談対話システムにおける笑い声の評価を行えることが分かった．今後の課題して，笑い声の評価サンプルを増やし，笑い声を変換した場合の対話の評価を行う．また，生成されたスペクトル音声波形を再合成する部分を改善する必要がある．

参考文献

[1] 下岡和也, 徳久良子, 吉村貴克, 星野博之, 渡部生聖．音声対話ロボットのための傾聴システムの開発．自然言語処理, Vol. 24, No. 1, pp. 3–47, 2017.

[2] 有本泰子, 今西利於, 森大毅．自然で表現豊かな笑い声合成に向けた感情情報からの笑い声の構成要素決定法．情報処理学会論文誌, Vol. 63, No. 4, pp. 1159–1169, 2022.

[3] El Haddad K. Tits, N. and T. Dutoit. Laughter synthesis: Combining seq2seq modeling with transfer learning. *Proc. Interspeech 2020*, p. 3401–3405, 2020.

[4] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. pp. 498–502, 2017.

[5] Valerio Velardo. generating-sound-with-neural-networks. https://github.com/musikalkemist/generating-sound-with-neural-networks/tree/main.