

第三話者読み上げコーパスによる音声ドキュメントの活用

小泉 悠馬¹ 塩出 萌子² 伊藤 克亘^{2,a)}

概要：音声ドキュメントの多くは作成時・生成時に制御することが不可能である。そのため、音声ドキュメントが潜在的に持つ可能性を十分に活用するのが困難である。本稿では、音声ドキュメントを活用のための第三者による読み上げコーパスを提案する。その応用として、第三話者コーパスを中間話者データとして利用したテレビアニメーションの話者変換を提案する。50文の第三話者コーパス二つを中間話者コーパスとして変換実験を行ったところ、明瞭度と音質の MOS 値はおおむね 3.5 以上を示し、最低限の変換を行えることを示した。

1. はじめに

音声言語には、言語情報だけでなく、パラ言語情報、非言語情報も含まれる。したがって、音声ドキュメントの活用方法は、潜在的に多様なはずである。

音声ドキュメントの活用方法を考えるために、言語の役割から考えてみる。言語使用には、二つの重要な役割があるといわれている [1]。一つは、interactional な機能と呼ばれるもので、人間が言語を用いて、社会的・感情的にどのようにお互いに作用しあうかにかかわっている。二つ目は、transactional な機能と呼ばれるもので、知識・技術・情報を伝達することにかかわっている。

もちろん、特に話し言葉の実際の言語使用においては、両方の側面を持つことが大半である。その中で、音声インタフェースの研究の多くは、interactional な機能に着目したものであり、音声ドキュメント検索に代表される音声ドキュメント処理の多くは、transactional な機能に着目しているといえる。

ところで、言語の発展につれて、重要な応用となってきたのが、映画や演劇、小説、音楽などの作品制作である。もちろん、これらの作品においても、制作者が観客や読者、聴衆に対してコミュニケーションを図るという意味では interactional だといえるし、コンテンツの内容を伝達する面を見れば transactional だといえ、言語使用としては、interactional, transactional 両方の側面を持つ。しかし、最も重要なのはそれらとは別に、制作者の表現である

う。表現は、例えば、言葉遣いや、声そのもの、話し方などにあらわれる。

表現に着目して、ドキュメント自身を作品制作に活用する試みは、引用という形で、古くから行われている。その引用は、技術の進歩につれて、直接的な引用により、素材として利用されるようになってきている。

例えば、ポピュラー音楽の分野では、レゲエなどの DJ のリミックスや、ヒップホップのトラックメイキングでのサンプリング、複数の楽曲によるマッシュアップなどがあげられる。CG でも、浮世絵の背景色をサンプリングして自分の絵に利用している例がある。

これらは、ドキュメントに含まれる質の高い表現の活用ととらえられる。これらを技術的に支援・解決するものとしては、声そのものや話し方の推定、抽出、転写、加工やそのための検索があげられる。これらを研究対象としようとした場合には、コーパスに代表されるようなデータの整備が不可欠である。しかし、この対象となるような質の高い表現は、その質の高さゆえに、ドキュメント利用者から見ると制御が困難である。したがって、コーパス化は困難である。

本稿では、アニメーションにおける声優の演技音声の声質を他の声優の声質に変換する問題を題材に、コーパス化が困難な部分をドキュメント利用者が第三話者として発話するコーパスを用いてカバーする方法を提案する。

2. アニメーションの演技音声の声質変換

日本アニメ(特にテレビ放映用)は、様々な制約より動きが少ない。したがって、それにあわせて、声優の声での表現は独特の豊かさを持つ。したがって、声のキャラクター性も強い印象を持つ。

¹ 法政大学大学院 情報科学研究科
3-7-2 Kajino-cho, Koganei, Tokyo 184-8584, Japan

² 法政大学 情報科学部
3-7-2 Kajino-cho, Koganei, Tokyo 184-8584, Japan

a) it@fw.ipsj.or.jp

しかし、長期間制作されている作品では、同じ登場人物に対する声優が交代してしまうことがある。交代した場合に、違和感を持つ人も少なくないようである。本研究では、このような違和感を解消するために、同一のキャラクターを別の声優が演じた発声をもう一方の声優が演じたように変換することを目的とする。つまり、あるアニメーション作品の交代後の声優(今後、元話者と表記する)の発話をその場面にそのまま使える形で、交代前の声優(目的話者と表記)が話したように変換する。

このような場合、発話のどの部分を変換すればよいのだろうか。本研究で想定する利用状況では、キャラクターの顔や動き、演技は同一である。もちろん、個々の発話における実際の話し方は、演技のバリエーションとして様々に変化する。実際に、同じ脚本のエピソードの台詞に対して、表現される感情が異なる演出がおこなわれた例があった。つまり、台詞を実現したときに含まれるパラ言語情報は演出レベルで異なるのである。

仮にこのレベルのパラ言語情報や、脚本レベルである言語情報で違和感があったとしても、本研究では、解消の対象とはしない。つまり、声に含まれる非言語情報での違和感を解消の対象とする。非言語情報は、大別すると、音源特性(声帯振動など)と声道特性があげられる。音源特性は、パラ言語情報を表現するために変動が非常に大きいので、まずは、声道特性を変換することを考える。

声質変換は、音素バランスを考慮したパラレルコーパスに基づく手法がよく用いられる。しかし、音声ドキュメント上の音声を対象とする場合、対象話者の発話テキストをコーパス制作者が制御できないため、パラレルコーパスの整備は困難である。

その問題を解決するために、変換における第三者による発話を整備した第三話者パラレルコーパスを利用する。このコーパスは、音声ドキュメントで利用したい発話のテキストを第三者が読み上げてパラレルコーパスを構築する。このパラレルコーパスを元話者と第三話者、第三話者と目的話者の二つの組(それぞれの発話テキストは異なる)について構築すれば、第三話者を中間話者として介することで、元話者から目的話者の声質変換が可能となる。

3. 中間話者パラレルコーパスを用いた声質変換

3.1 変換方法

ここでは、中間話者パラレルコーパスを用いた確率的な声質変換を次のように行なった。元話者の発声を x 、目標話者の発声を y として、中間話者の発声を z としたときに目標話者の発声を次の式で推定する。

$$\hat{z} = \operatorname{argmax}_z P(z|x) \quad (1)$$

$$\hat{y} = \operatorname{argmax}_y P(y|\hat{z}) \quad (2)$$

変換処理の概要を以下に示す。

Step 1. 元話者の音声データから声道特性を推定し、残差信号を分離する。

Step 2. 元話者の声道特性を変換モデル(元話者から中間話者)を用いて中間話者の声質に変換する。

Step 3. 中間話者の声質を変換モデル(中間話者から目標話者)を用いて目標話者の声質に変換する。

Step 4. 得られた声道特性を Step 1. で分離した残差信号に畳み込んで目的とする音声信号を合成する。

変換モデルとしては、学習データ量の少なさに頑健な手法として、変換の入出力音声対の結合確率密度推定に基づく学習法 [2] を用いる。学習データは、パラレルコーパスの発話対を動的計画法で対応付け、対応がとれたフレームのみを用いる。

変換には、変換モデルの静的特徴量を静的・動的特徴量に変換し、さらに系列内分散を考慮することで、統計処理の過度な平滑化を避ける [3]。

3.2 第三話者パラレルコーパス

中間話者パラレルコーパスは元話者の声優の第三話者パラレルコーパスと目的話者の声優の第三話者パラレルコーパスの二つのコーパスから構成される。

声優発声の部分は、アニメーションから収集する。BGMや効果音が重畳されておらず、他のセリフとも重畳していない発話を用いる。また、第三話者の発声作業を考慮し、驚きや悲しみ(泣き声)など喜怒哀楽の激しいセリフは除いた。このような発話はアニメーションにおいてはそれほど多くないため、コーパスの規模を大きくすることは困難である。

しかし、頑健な変換を実現するためには、学習用のコーパスが、なるべく音素バランスがとれていることが必要となる。そこで、ATRの音素バランスコーパス [4] の1セット(50文)の清音の音節の出現頻度を参照し、その頻度に達しない頻度の音節を含む文は必ず含むという方針で設計時に定めた文数になるまで選択する。

発話テキストが決定したら、第三話者は、声優発声の発話内容、口調(アクセント・イントネーションなど基本周波数(F0)の変化)、発話時間(話速)をできるだけ模倣して発声する。

4. 評価

提案法の変換精度を評価するために主観評価実験を行なった。

4.1 中間話者コーパス

変換対象は、アニメ作品「ドラえもん」とした。収集に用いたエピソード数と時間を表1に示す。

表1 収集データ

	旧声優	新声優
エピソード数	47	50
時間	613分42秒	513分17秒

これらを用いて、「ドラえもん」「のび太」「スネ夫」の3キャラクターの変換を行った。ドラえもんとのび太は女性声優、スネ夫は男性声優が演じている。収集できた発話の総数をキャラクタごとに表2に示す。

表2 収集した発話数

	旧声優	新声優
ドラえもん	136	146
のび太	203	257
スネ夫	60	49

これらの発話を息継ぎの部分などで分割し発声しやすい形に編集し、元話者(新声優)と第三話者、目標話者(旧声優)と第三話者、それぞれ50文の平行コーパスを整備した。50文の発話内容は異なる。発話テキストの例を示す。

えー(旧ドラえもん)

一年前もおんなじことやってたんじゃないか(同上)

はい(新のび太)

あーあ、一度でいーから高いまぐろをおなかいっぱい食べてみたいなー(同上)

あたり(旧スネ夫)

プラモをどっさり別荘におくってくれることになってるんだよ(同上)

図1 発話テキスト例

コーパスの規模(のべ音節数)を表3に示す。

表3 コーパスの規模(のべ音節数)

	旧声優	新声優
ドラえもん	544	543
のび太	642	575
スネ夫	619	536

第三話者は全てのキャラクターで同一で、演技経験などは特にない女子学生である。

4.2 変換処理

学習データのサンプリング周波数は16kHzとした。

学習データは、STRAIGHTスペクトルの24次のケプストラムの距離を用いて、Dynamic Time Warpingで対応付けた。フレーム長は15ms、フレームシフトは5msとした。

変換モデルの音響特徴量は、20次の線形予測係数を線スペクトル対に変換したものを用いた。モデルはEMアルゴリズムでGMMを64混合で学習した。分析窓長は32ms、フレームシフトは5msである。これらのパラメータは、実験的に定めた。

4.3 主観評価実験

変換音声の明瞭度、音質、話者性を被験者実験により評価した。被験者は学生10名である。

音声はヘッドホンを通じて提示し、聞き返しはなしとした。

評価に用いる音声データは、学習データに含まれない発話長が1秒以上のものとした。喜怒哀楽による偏りをなくすため怒りや悲しみを除く、感情表現の激しくない発声のものから選んだ。

4.4 明瞭度・音質評価

変換音声アニメ作品に利用可能な品質であるかを調べるため、発話内容の明瞭度と音質の劣化についてMean Opinion Score(MOS)実験により評価した。1キャラクターにつき、20文をランダムに選択し提示した。

明瞭度の結果を図2、音質の結果を図3に示す。

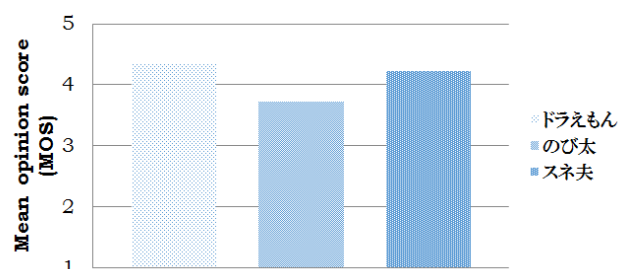


図2 明瞭度の評価結果

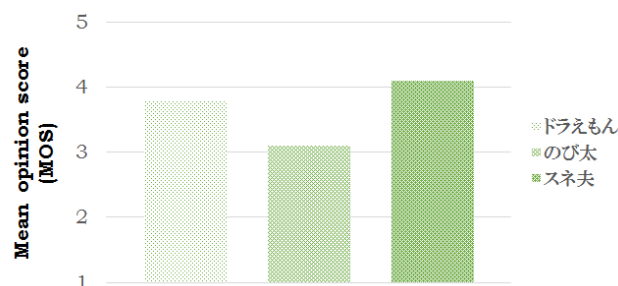


図3 音質の評価結果

のび太の音質を除いては、3.5以上を示しているため、許容可能なレベルである。しかし、アニメーションのセリフとしては、不十分な品質であろう。

4.5 話者性評価

話者性の評価には、変換音声が元話者と目的話者の音声のどちらに近いかを答えさせる ABX 評価を用いた。今回の実験の場合、元話者 A と目的話者 B の発話内容や発声方法を制御できないため、通常の ABX 評価と違い、AB とも同一の発話を使えるわけでない。声優の発声は、大きく F0 が変化するため、試行実験では、声質の変換の話者性に関係なく、F0 の値が近い方を選んでしまうことが多かった。そこで、発話ごとに F0 の平均を求め、A, B, X (変換音声) それぞれで昇順に並べて、組合せを決めた。変換音声は、MOS 実験と同じものを用いた。

結果は、「ドラえもん」が 31.0%、「のび太」が 60.5%、「スネ夫」が 50.5% であった。

4.6 考察

明瞭度評価の低かったサンプルを調べたところ、音質が悪いものが多かった。明瞭度と音質の 2 次元プロットを図 4 に示す。

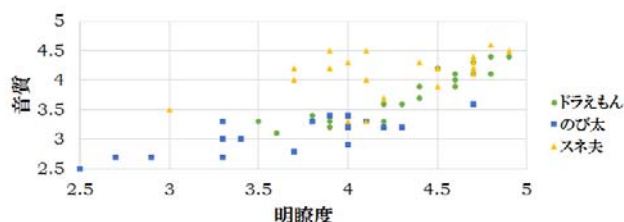


図 4 明瞭度と音質の関係

相関係数は、ドラえもんが 0.92、のび太は 0.80 と、高い値を示した。

また、音質評価の低かったサンプルの特徴として、発話データ中の子音部分で音質劣化が感じられる傾向にあった。これは、子音の学習データが不足していることに起因している。提案したパラレルコーパスの制作方法では、音声データを現状以上に増やすことは困難であり、音声強調や雑音除去で声優発声のデータ量を増やすことを考慮すべきである。

話者性評価が低かったサンプルには、語尾の言い回し(のびし方など)やイントネーションなどの言葉遣いが特徴的な入力データがあった。語尾の終助詞「よ」ののびし方が異なる例を次に示す。

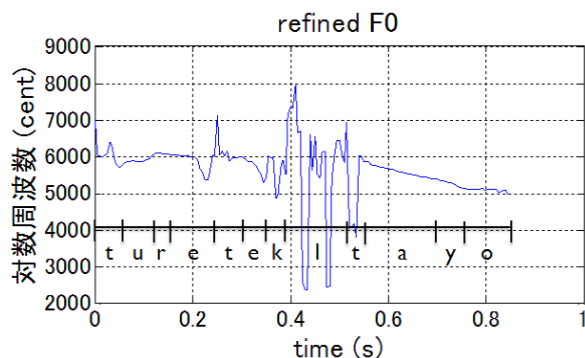


図 5 旧声優の発話の韻律の例

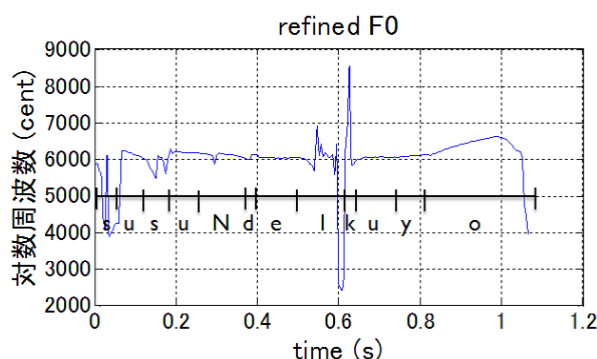


図 6 新声優の発話の韻律の例

この例では、新声優が、終助詞「よ」をのばして、F0 を上昇させていることがわかる。このような言い回しは、新声優ではしばしば聞かれたが、旧声優では余り聞かれなかった。

これは、特徴的な言葉遣いが特定の人物像を想起させる役割語 [5] となり、新声優の発声と旧声優の発声が別の発話キャラクタ [6] となってしまったと考えられる。

またドラえもんやスネ夫は、ダミ声のような発声をするがあった。しかし、ダミ声の特徴が新旧の声優で異なっていた。

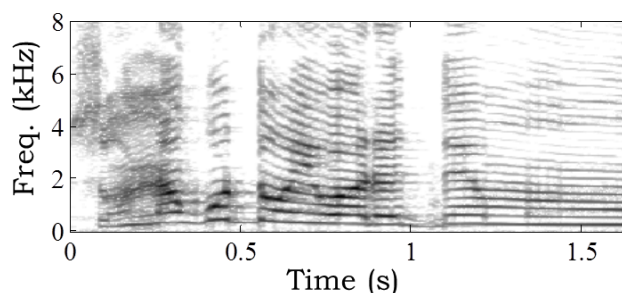


図 7 ダミ声の発話の例 (新声優)

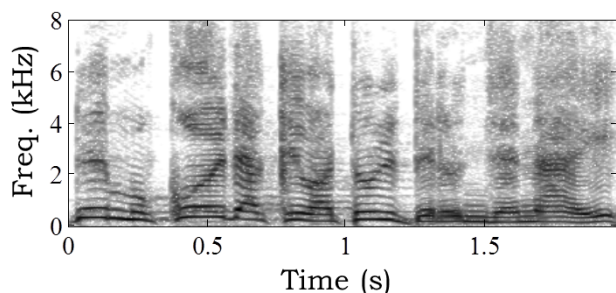


図8 ダミ声の発話の例(旧声優)

ダミ声は非調波成分の多く含まれる発声[7]で、声帯振動に関係している。旧声優は、図8で発話全体で非調波成分が強いことから、声質自体がダミ声であることがわかる。一方、新声優は、図7の冒頭0.3秒あたりまでに顕著のように、発話やフレーズの冒頭部分のみで非調波成分が強い。このような現象は、故意にダミ声を発声した場合に起きることが多い。この場合、印象が強くなるフレーズの冒頭部分のみダミ声で発声することで、発話の印象をダミ声となるようにしている。このような特徴の違いは、残差信号に残るため、元話者(新声優)の特徴が変換音声に残ってしまう。

5. おわりに

本研究では、同じキャラクターを演じる声優の声質変換を目的として、第三話者パラレルコーパス二つからなる中間話者パラレルコーパスを用いた声質変換手法を提案した。演技経験のない第三話者によるある程度音素バランスを考慮した50文ずつの中間話者パラレルコーパスを用いて変換した音声は、明瞭度・音質ともに、おおむね3.5を越える結果を示しており、最低限の品質は達成できた。しかし、話者性評価の値は最高でも0.6と余り高くなかった。しかし、明示的なモデル化の難しい声優の声質を扱う上で、第三話者コーパスの一定の可能性は示されたのではないだろうか。

最後に今後の課題をあげる。アニメーション音声のコーパス化に関しては、データ量を十分に確保するために、音声強調や雑音/BGM除去が不可欠である。

演技音声の話者変換のためには、言い回しの個性や、ダミ声などの発声方法のモデル化が必要である。

参考文献

- [1] Brown, G. et. al. "Discourse Analysis", Cambridge University Press, 1983.
- [2] Kain, A. et. al. "SPECTRAL VOICE CONVERSION FOR TEXT-TO-SPEECH SYNTHESIS," Proc.ICASSP, p.285-p.288, May 1998.
- [3] Tomoki Toda, et. al. "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory", IEEE Trans. AUDIO, Speech, Lang, Process., VOL.25, NO.8, pp.2222-2235, 2007

- [4] 磯, 他 "音声データベース用文セットの設計", 日本音響学会講演論文集, p.89-p.90, 1988.
- [5] 金水敏, "ヴァーチャル日本語 役割語の謎", 岩波書店, 2003.
- [6] 定延利之, "ことばと発話キャラクタ", 文学, 2006, p.117-p.129.
- [7] Eiji Yumoto, et. al. "Harmonics-to-noise ratio as an index of the degree of hoarseness", J.Acoust.Soc.Am, Vol.71, NO.6, pp.1544-1550, 1982