Vol.2020-MUS-127 No.19
Vol.2020-SLP-132 No.19
2020/6/6

# テキスト音声合成における劣化音声を活用したデータ拡充に関する検討

五来 丈瑠[1,a)]　須田 仁志[1,b)]　齋藤 大輔[1,c)]　峯松 信明[1,d)]

**概要**：本稿では，テキスト音声合成において，非可逆圧縮音声などの劣化音声を用いた学習データの拡充を行う試みとして，非負値行列因子分解を用いた手法を提案する．実験の結果，主観評価により提案手法の有効性が示された．

**キーワード**：テキスト音声合成, 非可逆圧縮音声, 非負値行列因子分解

## 1. はじめに

音声合成とは，音声を人工的に生成する技術であり，テキスト音声合成はテキストからそれに対応する音声を生成する技術をさす．近年，テキスト音声合成において深層ニューラルネットワーク (DNN) に基づく統計的パラメトリック音声合成 [1] や WaveNet[2] など深層学習を用いた手法が提案され，合成音声の自然性は大きく向上した．それらの手法では大量の音声データを必要とするが，高品質のデータを大量に集めるのはコストがかかる．そのため，比較的入手しやすい低品質音声を効果的に活用することが求められる．ここで低品質音声とはノイズや非可逆圧縮などによって劣化した音声を指し，特に MP3 音声などの非可逆圧縮をされた音声はインターネット上に大量に存在する．

音声の非可逆圧縮では，MP3 (MPEG-1 Layer3) や AAC (MPEG-2 Advanced Audio Coding) などのコーデックが用いられており，それらでは人間の聴覚心理を利用することで音質の劣化をあまり伴わずに情報量を削減することが可能になっている．具体的には，周波数ごとの音の聞こえやすさや大きな音が鳴った際にその直前，直後や近い周波数の音が聞こえづらくなるマスキング効果を考慮して，周波数ごとに量子化ビット数を適応的に割り当てている [3]．そのため非可逆圧縮音声を学習データとして用いた場合，音声から言語情報などを抽出する音声認識では影響は出にくいが，音声合成をする際には合成音声の品質が劣化することが予想される．[4] では，HMM を用いたボコーダーの学習に MP3 音声を用いたときの影響が検証されている．しかし，深層学習に基づく音声合成については分析が十分でない．

そこで本研究では，まずテキスト音声合成の学習データとして MP3 などの非可逆圧縮音声を用いることの影響を実験的に評価する．次に評価実験の結果をもとに，このような非可逆圧縮音声をなるべく品質低下を伴わずにテキスト音声合成に適用する方法について検討する．本研究では，非負値行列因子分解 (NMF) に基づく手法を 2 種類提案する．一つ目は，劣化音声とクリーンな音声のパラレルなデータセットを利用する手法であり，エンコーディングが既知であることが必要である．二つ目は，未知の劣化に対しても適用できる手法である．提案手法により非可逆圧縮音声を効果的に利用できることを評価実験において示す．

## 2. 音声合成における NMF の応用

### 2.1 NMF の原理

振幅スペクトルなどの非負値の観測ベクトル ($K$ 次元) が $N$ 個与えられたとし，$\boldsymbol{y}_1, ..., \boldsymbol{y}_N$ とする．今，式 (1) のように各ベクトルを $M$ 個の基底ベクトル $\boldsymbol{h}_1, ..., \boldsymbol{h}_M$ の非負結合で近似することを考える．

$$\boldsymbol{y}_n \simeq \sum_{m=1}^{M} \boldsymbol{h}_m u_{m,n} (n = 1, ..., N) \qquad (1)$$

ここで，観測ベクトルを並べた行列を $\boldsymbol{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]$，基底ベクトルを並べた行列を $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_M$，$\boldsymbol{U}] = (u_{m,n})_{M \times N}$ とおくと，式 (1) は

$$\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U} \qquad (2)$$

と表され，非負の観測行列を非負の行列の積で近似する問

[1] 東京大学大学院工学系研究科
Graduate School of Engeneering, The University of Tokyo
[a)] gorai@gavo.t.u-tokyo.ac.jp
[b)] hitoshi@gavo.t.u-tokyo.ac.jp
[c)] dsk_saito@gavo.t.u-tokyo.ac.jp
[d)] mine@gavo.t.u-tokyo.ac.jp

題と見ることができ，これを非負値行列因子分解 (NMF) と呼ぶ [5]．ここでは，$\boldsymbol{H}$ を基底，基底の重み $\boldsymbol{U}$ をアクティビティと呼ぶ．NMF の非負制約により，アクティビティ $\boldsymbol{U}$ はスパースになることが知られている．

式 (2) は，$\boldsymbol{Y}$ と $\boldsymbol{HU}$ の乖離度 $\mathcal{D}(\boldsymbol{Y} \mid \boldsymbol{HU})$ が最小になるような $\boldsymbol{H}$ と $\boldsymbol{U}$ を推定する問題と言い換えることができる．NMF を解くアルゴリズムでは，補助関数法による反復計算を行うことで各行列の非負値性を保ちつつ $\mathcal{D}(\boldsymbol{Y} \mid \boldsymbol{HU})$ を単調に小さくしている．乖離度の規準 $\mathcal{D}$ には，二乗誤差，I ダイバージェンス，板倉斎藤擬距離などが用いられ，それぞれ

$$\mathcal{D}_{EU}(y \mid x) = (y - x)^2 \tag{3}$$

$$\mathcal{D}_{KL}(y \mid x) = y \log \frac{y}{x} - y + x \tag{4}$$

$$\mathcal{D}_{IS}(y \mid x) = \frac{y}{x} - \log \frac{y}{x} - 1 \tag{5}$$

で与えられる．

式 (4) の I ダイバージェンス規準を用いる場合，式 (6), (7) に示した更新を反復的に行う．
ただし，$x_{k,n} = \sum_m h_{k,m} u_{m,n}$ とおいた．

$$h_{k,m} \leftarrow h_{k,m} \frac{\sum_n y_{k,n} u_{m,n} / x_{k,n}}{u_{m,n}} \tag{6}$$

$$u_{m,n} \leftarrow u_{m,n} \frac{\sum_k y_{k,n} h_{k,m} / x_{k,n}}{h_{k,m}} \tag{7}$$

### 2.2 NMF による声質変換

NMF には，音声情報処理における様々な応用方法が提案されている．その一つに NMF を用いた統計的声質変換があげられる [6]．以下の手法で声質変換を行っている．

学習時は入力話者 A と出力話者 B のパラレルな音声を用意し，それらの振幅スペクトログラムをそれぞれ $X_A$, $X_B$ とする．次の手順でパラレルな基底を生成する．

1) $X_A$ に NMF を適用し，

$$X_A \simeq H_A U_X \tag{8}$$

となるように分解する．

2) 1) で求めた $U_X$ を固定して $X_B$ に NMF を適用し，

$$X_B \simeq H_B U_X \tag{9}$$

となるように分解する．

変換時にはパラレルな基底 $H_A$，$H_B$ を用いて，与えられた入力話者のスペクトログラム $Y_A$ を出力話者のスペクトログラム $Y_B$ に変換する．手順を以下に示す．

1) $H_A$ を固定した状態で $Y_A$ に NMF を適用し，

$$Y_A \simeq H_A U_Y \tag{10}$$

となるように分解する．

2) 出力話者の基底 $H_B$ と，1) で得られた $U_Y$ の積をとることで以下のように $Y_B$ を得ることができる．

$$Y_B = H_B U_Y \tag{11}$$

最後に得られた $Y_B$ からボコーダーを用いて波形生成を行う．

[6] では，I ダイバージェンス規準が用いられている．

### 2.3 NMF を用いたテキスト音声合成

NMF のアクティビティを音響特徴量として用いた統計的パラメトリックテキスト音声合成が提案された [7]．この手法 (以下 NMF-TTS と呼ぶ) では次のような手順で TTS を行う．

まず，訓練用の音声データの振幅スペクトログラム X に対して NMF を適用し，基底行列 H とアクティビティ $U_X$ に分解する．次に言語特徴量と生成された $U_X$ の対応を DNN で学習する．テキストから音声を合成するときは，学習した DNN を用いて言語特徴量からアクティビティ $U_Y$ を推定する．訓練用の音声から得た $H$ と $U_Y$ を掛け合わせることでテストデータのスペクトログラム Y を得る．最後にボコーダーを用いて波形を生成する．

従来の統計的パラメトリック音声合成では，メルケプストラム係数やスペクトル包絡を音響特徴量としていたのに対し，この手法ではアクティビティを用いている．アクティビティはスパースな特徴量のため誤差関数に二乗誤差を用いるのは適切ではない．そこで次のように誤差関数が設計されている．

まず，各フレームのアクティビティを $\boldsymbol{u}'$ とする．$\boldsymbol{u}'$ をその L1 ノルム c で割ることで和が 1 となるように正規化されたアクティビティ $\boldsymbol{u}$ を得る．$\boldsymbol{u}'$ の KL ダイバージェンスは次のように展開できる．

$$\begin{aligned}\mathcal{D}_{KL}(\boldsymbol{u}' \mid \hat{\boldsymbol{u}}') &= \sum_m (c u_m \log \frac{c u_m}{\hat{c} \hat{u}_m} - c u_m + \hat{c} \hat{u}_m) \\ &= c\{-\sum_m u_m \log \hat{u}_m + (\frac{\hat{c}}{c} - \log \frac{\hat{c}}{c} - 1) \\ &\quad + \sum_m u_m \log u_m\} \\ &= c\{\mathcal{D}_{CE}(\boldsymbol{u} \mid \hat{\boldsymbol{u}}) + \mathcal{D}_{IS}(\hat{c} \mid c) - \mathcal{D}_{CE}(\boldsymbol{u} \mid \hat{\boldsymbol{u}})\}\end{aligned} \tag{12}$$

$\mathcal{D}_{CE}$ はクロスエントロピー，$\mathcal{D}_{IS}(\hat{c} \mid c)$ は板倉齋藤距離で観測値と推定値を入れ替えた双対板倉齋藤距離 (D-ISD) である．そこで正規化アクティベーション $\boldsymbol{u}'$ のクロスエントロピーとパワー c の D-ISD の和を誤差関数とする．

NMF-TTS では，音響特徴量にメルケプストラムを用い

た場合と比べ，よりスペクトルの微細構造を保持しながら音声合成でき，論文中の評価実験では，従来の統計的パラメトリック音声合成とと同等以上の品質の音声が生成できることが示されている．NMF-TTS では，他の NMF を応用した音声処理技術とテキスト音声合成を組み合わせることが期待できる．[7] では，その一つとして NMF を用いた帯域拡張 [8] を TTS に応用した手法が述べられている．概要を以下に示す．

まず，訓練データとして狭帯域の音声と一部広帯域の音声が与えられているとする．広帯域，狭帯域のパラレルな音声を用いて NMF 声質変換と同様の方法でパラレルな基底を生成する．次に狭帯域音声から抽出したアクティビティを用いて DNN 音響モデルを学習し，テストデータの言語特徴量に対してアクティビティを推定する．最後に得られたアクティビティに広帯域の基底をかけることで，狭帯域の音声と少量の広帯域の音声から広帯域の音声を合成することができる．論文中の実験では，サンプリング周波数が 48kHz と 16kHz の音声で行われており，データの量は同じで 48kHz の音声のみを用いた場合に近い品質での音声合成ができることが示されている．

## 3. NMF を用いた劣化音声の活用

学習データとしてクリーンな PCM 音声と非可逆圧縮音声が与えられている状況を考える．本稿では 2 節で述べた NMF による帯域拡張や声質変換に着目したテキスト音声合成法を 2 通り提案する．

第 1 の手法は，NMF 声質変換と同じ方法で MP3 音声を PCM 音声に変換し，それを学習データとしてテキスト音声合成を行う手法である．この手法は，劣化音声のエンコーディングが既知であり，PCM と MP3 のパラレルな音声が作れることを前提としている．

第 2 の手法は，NMF-TTS を応用した手法である．こちらはパラレルな音声を用意する必要がないため，エンコーディングが未知の場合にも適用できる．

### 3.1 パラレルな音声を用いた手法

前述の通り，今設定している状況では劣化音声のエンコーディングを既知としているので，半分の PCM 音声を圧縮することでパラレルな MP3 (32kbps) 音声を生成することができる．そのため NMF 声質変換と同じ方法で圧縮済みの MP3 音声を PCM 音声に近づけることができる．手順を以下に示す．

まず，パラレルな基底を生成するため，PCM 音声とそれを圧縮した MP3 (32kpbs) 音声からスペクトル包絡を抽出し，それぞれ $X_{pcm}$，$X_{mp3}$ とする．$X_{mp3}$ に NMF を適用し，

$$X_{mp3} \simeq H_{mp3} U_X \tag{13}$$

と分解する．次に $X_{pcm}$ に対し，アクティビティを $U_X$ に固定して NMF を行う．

$$X_{pcm} \simeq H_{pcm} U_X \tag{14}$$

このようにして，パラレルな基底 $H_{pcm}$，$H_{mp3}$ を得る．MP3 音声のスペクトログラムを先に分解したのは [6] で入力話者の方を先に分解していることに対応している．

次に残りの MP3 音声のスペクトル包絡 $Y_{mp3}$ に対し，次式のように基底を $H_{mp3}$ に固定した状態で NMF を行い，アクティビティを抽出する．

$$Y_{mp3} \simeq H_{mp3} U_Y \tag{15}$$

$H_{pcm}$ と $U_Y$ の積をとることで，$Y_{mp3}$ を PCM に近づけた ${Y'}_{pcm}$ を得る．$X_{pcm}$ と ${Y'}_{pcm}$ を音響特徴量としてテキスト音声合成を行う．

### 3.2 NMF-TTS を用いた手法

前節で提案した手法では，劣化音声のエンコーディングが既知であることを利用した．しかし，実際には劣化を再現できないような状況も多く存在する．例えば，非可逆圧縮音声でエンコーディングが複数回されている場合である．そこで，エンコーディングが未知の場合でも適用できる手法について検討する．

手法の概要を説明する．まず，PCM 音声のスペクトル包絡 $X_{pcm}$ と $Y_{mp3}$ に対し，同時に NMF を行う．

$$[X_{pcm}, Y_{mp3}] \simeq H_{mix} [U_X, U_Y] \tag{16}$$

このとき生成された基底 $H_{mix}$ は，PCM と MP3 の中間のような基底になると予想される．次に $X_{pcm}$ に対しアクティビティを $U_X$ に固定して NMF を行い，PCM 音声の基底 $H_{pcm}$ を抽出する．

$$X_{pcm} \simeq H_{pcm} U_X \tag{17}$$

次に $U_X, U_Y$ を用いて DNN を学習し，テストデータのアクティビティを推定する．最後に $H_{pcm}$ との積をとり，得られたスペクトル包絡からボコーダーにより波形生成を行う．

## 4. TTS において非可逆圧縮がもたらす影響の評価

ここでは DNN に基づく統計的パラメトリック音声合成において，非可逆圧縮音声を用いた場合に生じる品質低下を評価した実験について述べる．

### 4.1 音響特徴量としてメルケプストラムを用いることの問題点

MP3 や AAC などの非可逆圧縮では ビットレートに応

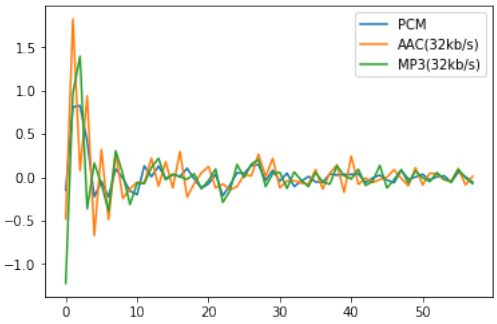


**図 1** あるフレームにおける非可逆圧縮音声のメルケプストラムの例

じてローパスフィルタが適用され，スペクトルに不連続な段差が生じる．

ボコーダを用いたテキスト音声合成ではスペクトル包絡を表現したパラメータとしてメルケプストラム係数が用いられるが，求める過程で対数スペクトルに対し，逆フーリエ変換が行われる．非可逆圧縮によるスペクトル領域での段差がケプストラム領域では標本化関数の畳み込みとなって現れるため，メルケプストラムで不適切で大きな歪みが生じる．図 1 に例を示す．これが音響特徴量の推定精度を下げる一因になりうる．

### 4.2 実験条件

本実験では，音声データベースとして ATR 日本語音声データベース [9] に収録された音声（男性話者の発話音声）を用いた．[9] は A-J セット，合計 503 文の発話で構成される，大きな雑音のない環境下で収録されたクリーンな音声である．音声フォーマットは PCM，サンプリング周波数は 48kHz である．

このデータセットをベースに，1)PCM (450 文) (*pcm_full*)，2)MP3 (450 文) (*mp3_full*)，3)PCM (225 文) (*pcm_half*)，4)PCM (225 文)+MP3 (225 文) (*pcm_mp3*) の 4 種類を訓練データとしてそれぞれ TTS を行った．いずれも J セット (53 文) を評価用として用いた．MP3 音声はサンプリング周波数は変えずに，MP3 で一度圧縮してから PCM に戻して生成した．圧縮時のビットレートは 32kbps とした．

音響モデルの入力特徴量には HTS[*1] のフルコンテキストラベルを元にフレームごとに計算された 425 次元の言語特徴量を [0.01,0.99] の値を取るように正規化したものを用いた．出力特徴量は音声から抽出した基本周波数，メルケプストラム係数，非周期性指標及びそれらの一次，二次動的特徴量と有声/無声のフラグを用い，平均が 0，分散が 1 となるよう正規化した．音響モデルには FeedForward 型の DNN を用いた．隠れ層は 6 層・1024 次元，活性化関数は tanh とした．これらは DNN 音声合成ツール Merlin[12] のデフォルトの値である．また，誤差関数は二乗誤差とした．DNN の出力から最尤パラメータ生成により静的な音響特徴量を推定した．音響特徴量であるスペクトル包絡は 1025 次元，非周期性指標は 5 次元とした．音声の分析・合成には，音声分析変換合成システム WORLD[10] (D4C edition [11]) を用いた．

[*1] http://hts.sp.nitech.ac.jp/

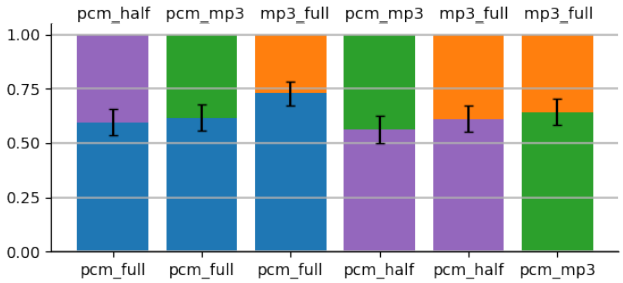


**図 2** 主観評価結果．エラーバーは 95 %信頼区間を示す．

上記の手順で 4 種類の合成音声を生成した．これらの音声の自然性を主観的に評価するため，プリファレンス AB テストを行った．このテストでは，比較したい 2 つの同一発話の音声でペアを作り，被験者は音質・自然性が高いと感じた方を選択する．一つのペアにつき被験者は 25 人とし，一人の被験者は J セットの 53 文からランダムに選ばれた 10 文を評価した．

### 4.3 実験結果

主観実験の結果を図 2 に示す．*pcm_mp3* は，*mp3_full* より品質が良かったものの，*pcm_full* より低くなっていた．また，*pcm_half* と比べても品質が低く，PCM と MP3 (32kbps) の音声が 225 文ずつある場合に MP3 (32kbps) の方は用いない方が良いという結果を示した．これは，先に述べたようなメルケプストラム領域での歪みが生じた音声とそうでない音声が混在していることにより，メルケプストラムの推定が適切に行われなくなったためだと推測される．また，*pcm_full* と *pcm_half* では学習データ量の大きい *pcm_full* の方が有意に品質が高いという結果を示した．

## 5. 劣化音声を活用するための手法の実験的評価

### 5.1 パラレルな音声を用いた手法

#### 5.1.1 実験条件

前節と同じデータセットを用いてテキスト音声合成を行い，品質を評価した．

[6] や [7] と同様，NMF の適用対象として振幅スペクトログラムを利用した．NMF は I ダイバージェンス規準で行い，基底数は 200 とした．また初期値はランダムに設定した．正則化は行わなかった．テキスト音声合成は 4 節のものと同じ手法・条件で行った．また，基本周波数と非周期性指標は同一のものを利用した．

合成音声に対し，プリファレンス AB テストによる主観

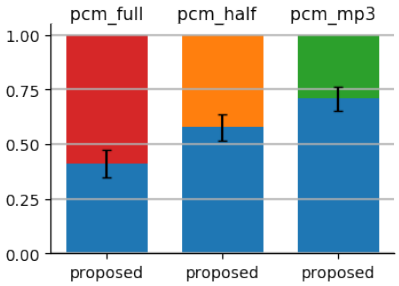


**図 3** パラレルな音声を用いた手法の主観評価結果．エラーバーは 95 %信頼区間を示す．

**表 1** 実験に用いた音声フォーマット

| 訓練用音声 | 合成手法 | 合成音声の表記 |
|---|---|---|
| PCM (450 文) | NMF-TTS | *pcm_full* |
| PCM (225 文) | NMF-TTS | *pcm_half* |
| PCM (225 文)+MP3 (225 文) | NMF-TTS | *pcm_mp3* |
| PCM (225 文)+MP3 (225 文) | 提案手法 | *proposed* |

評価を行った．4 節の実験で合成した *pcm_full*，*pcm_half*, *pcm_mp3* の 3 つと比較した．

#### 5.1.2 実験結果

主観評価実験の結果を図 3 に示す．提案手法の合成音声の自然性は *pcm_full* と比べるとやや劣っていたものの，*pcm_half* や *pcm_mp3* からは改善が見られた．

*pcm_full* からの品質劣化の要因として，NMF による基底生成や $H_{MP3}$ によるアクティビティの抽出の過程で劣化が生じたことと，PCM のスペクトル包絡への変換が完全には再現できなかったことが挙げられる．しかし，*pcm_mp3* から品質が改善されたことから，基底の交換により MP3 の圧縮に起因する劣化が改善され，メルケプストラム領域でデータが混ざることによる学習効率の低下を軽減する効果が確認された．さらに *pcm_half* よりも品質が高くなっていたことから，MP3 音声による訓練データの拡充が実現できることが示された．

### 5.2 NMF-TTS を用いた手法

#### 5.2.1 実験条件

PCM 音声と MP3 (32kpbs) 音声が 225 文あるようなデータセットに対し，3.2 で述べた手法を適用した．また，比較のため PCM (450 文)，PCM (225 文)，PCM (225 文)+MP3 (225 文) の 3 種類のデータセットに対して NMF-TTS を適用した．すべての条件において，[9] の J セット (53 文) を評価用データとして TTS を行った．以上により表 1 に示した 4 種類の条件でそれぞれ評価用音声を生成した．それらに対し，主観評価実験による品質の比較を行った．

すべての手法において，NMF はパラレルな音声を用いた手法の実験と同じ条件で行った．アクティビティ推定を行う DNN は中間層は 6 層・各 1024 次元の feedforward 型であり，中間層の活性化関数は tanh とした．入力は 675 次元の言語特徴量である．出力は 2.3 節に述べたように，重みの和が 1 となるように正規化されたアクティベーションとパワー項の計 201 次元である．出力関数はアクティビティが softmax，パワー項が softplus とした．誤差関数はアクティビティが cross-entropy，パワーが双対板倉齋藤擬距離とした．分析・合成には WORLD を用い，波形生成時の基本周波数と非周期性指標は 4 節の実験で生成されたものを用いた．

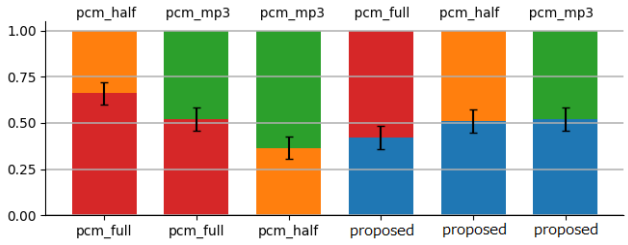


**図 4** 提案手法及び NMF-TTS による合成音声の主観評価結果．エラーバーは 95 %信頼区間を示す．

主観評価実験では，表 1 の 4 種類の条件での音声から総当たりでプリファレンス AB テストにより品質を比較した．

#### 5.2.2 実験結果

主観評価実験の結果を図 4 に示す．*pcm_full* の品質は *pcm_half* の品質よりも有意に高かった．NMF-TTS における学習データ量の差が合成音声の品質の差に現れたことを示している．これは，3 節の実験と同様の結果である．一方，*pcm_mp3* は，*pcm_full* と同程度の品質になるという結果を示し，3 節の実験結果とは異なる傾向が見られた．4.1 で述べたようにメルケプストラムは非可逆圧縮音声の影響を受けやすく，メルケプストラムを音響特徴量として用いた TTS で PCM と MP3 (32kpbs) 音声を混ぜると学習効率が低下していた．それに対し，NMF-TTS の場合はエンコーディングの依存度が低いアクティビティの領域でデータを混合したので，学習効率の低下がほとんど起こらなかったということが示された．

*proposed* は *pcm_half* と比較して品質に改善がみられたものの，*pcm_mp3* と *pcm_full* の品質に有意差が生じなかったため，PCM と MP3 (32kpbs) の音声を混合しそのまま NMF-TTS を適用した方法に対する提案手法の優位性は確認できなかった．

## 6. おわりに

本研究では，テキスト音声合成におけるデータの拡充を目的として，テキスト音声合成の学習データに非可逆圧縮音声を用いた場合の影響を実験的に調査し，それらを効果的に活用する方法についていくつか検討した．

TTS に非可逆圧縮音声を用いることで起こる品質低下を評価した実験では，ある程度ビットレートが低い場合，MP3 の非可逆圧縮により，合成音声の品質が劣化している

ことが明らかになった．また，エンコーディングが混ざっている状態でそのまま TTS を行うと合成音声の品質が低下することが実験により示された．

上記の問題を改善するため，NMF によるスペクトルモデリングを導入した手法を 2 つ提案し，PCM 音声と MP3 (32kpbs) 音声が半分ずつ存在する実験系で評価実験を行った．パラレルな音声を用いた手法の評価実験では，MP3 (32kpbs) を効果的に利用でき，MP3 音声を用いたデータ拡充に成功していることが示された．

また，エンコーディングが未知の場合にも適用できる，NMF-TTS を応用した手法の評価実験では，有用性を確認することはできなかった．しかし，NMF-TTS ではアクティビティを音響特徴量として用いていることに起因して，メルケプストラムを用いた TTS の実験で見られたエンコーディングが混ざっている場合の影響が大きく軽減されるという知見を得た．

今回は PCM 音声と非可逆圧縮音声が半分ずつ存在するような状況で実験を行った．そこで今後の課題として劣化音声の割合や量を変化させたときの各合成手法における品質の評価が求められる．また劣化の程度や量の割合がどの程度であれば劣化音声を加えた方が良いかについて検討が求められる．

## 参考文献

[1] H. Zen, A. Senior and M. Schuster, “Statistical parametric speech synthesis using deep neural networks”, IEEE international conference on acoustics, speech and signal processing, pp. 7962-7966, 2013.
[2] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, “WaveNet: a generative model for raw audio. ” arXiv:1609. 03499, 2016.
[3] BRANDENBURG. K, “mp3 and AAC explained”, AES 17th International Conference on High-Quality Audio Coding, 1999.
[4] Bajibabu Bollepalli, Tuomo Raitio and Paavo Alku “Effect of MPEG Audio Compression on HMM-based Speech Synthesis”, 2014.
[5] D. D. Lee and H. S. Seung: “Algorithms for non-negative matrix factorization”, Advances in neural information processing systems, pp. 556-562, 2001.
[6] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng and H. Li, “Exemplar-based voice conversion using non-negative spectrogram deconvolution”, 8th ISCA Speech Synthesis Workshop, pp. 201-206, 2013.
[7] S. Goto, D. Saito, and N. Minematsu, “DNN-based statistical parametric speech synthesis incorporating non-negative matrix factorization”, Asia-Pacific Signal and Information Processing Association, 2019.
[8] D. Bansal, B. Raj and P. Smaragdis, “Bandwidth expansion of narrowband speech using non-negative matrix factorization. ”, Eurospeech’05, 2005.
[9] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis”, Speech Communication, pp. 357-363, 1990.
[10] M. Morise, F. Yokomori and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications”, IEICE transactions on information and systems, pp. 1877-1884, 2016.
[11] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis”, Speech Communication, pp. 57-65, 2016.
[12] Z. Wu, O. Watts and S. King, “Merlin: an open source neural network speech synthesis system” 9th ISCA Speech Synthesis Workshop (SSW9), 2016.