Vol.2018-CVIM-212 No.10
2018/5/10

生成誤差最小基準と敵対的学習に基づく顔の向きを考慮したDNN発話動画像生成

石川 澄[1]　能勢 隆[1,a)]　佐藤 一樹[1]　伊藤 彰則[1]

概要：本稿では，DNN に基づくテキストからのパラメトリック発話動画像生成において再現性と自然性を向上させるため，生成誤差最小基準と敵対的学習を導入し，かつ顔の向きを考慮した手法を提案する．DNN に基づくパラメトリック発話動画像生成では AAM などにより顔画像をパラメータ化し，統計的音声合成と同様の枠組でフィードフォワードネットワークや LSTM 等によりパラメータ系列のモデル化・予測を行う．前者の場合には滑らかなパラメータ系列を予測するため，静的・動的特徴量を用い，この平均二乗誤差を最小とするようにネットワークを学習する．しかし，本来最小化すべき静的特徴量を直接最小化していないため，モデル化の精度が十分でないという問題があった．そこで我々は静的・動的特徴量から生成した静的特徴量を最小化する生成誤差最小基準を導入することにより精度の改善を図る．また，画像生成分野で広く用いられている敵対的学習を利用し，主観的な品質の向上に取り組むとともに，AAM の性質を利用した低次パラメータの置換により，土台となる動画の顔の向きを生成動画の顔に反映できることを示す．

DNN-Based Talking Movie Generation Using Minimum Generation Error Criterion and Adversarial Learning with Face Direction Consideration

Toru Ishikawa[1]　Takashi Nose[1,a)]　Kazuki Sato[1]　Akinoti Ito[1]

1. はじめに

テキストからその内容に応じた目標話者の発話動画像を生成する技術はトーキング・ヘッド，Visual Text-to-Speech(VTTS) などと呼ばれ，これまで様々な手法が提案されている [1], [2]．中でも顔画像をシェイプとアピアランスに分離し再合成する Active Appearance Model(AAM)[3] 等を用いる統計的パラメトリック発話動画像生成は，統計的パラメトリック音声合成 [4], [5] と同様の枠組でパラメータ系列のモデル化・予測を行うことで，比較的少量の学習データで滑らかで安定した発話動画像を生成できるという利点がある．

これまでの取り組みとしてはクラスタ適応学習 (CAT)[6] を用いて隠れマルコフモデル (HMM) に基づき複数の感情を表現することができる手法 [7] や同様の出力を多出力ディープニューラルネットワーク (DNN) により実現した手法 [8], Long Short-Term Memory[9] を用いた RNN(LSTM-RNN) に基づく手法 [10], [11] などが提案されている．フィードフォワード NN(FF-NN) を用いる場合には時間変化パターンを考慮し滑らかなパラメータ系列を予測するため静的・動的特徴量を用い，この平均二乗誤差を最小とするようにネットワークを学習する．しかし，本来最小化すべき静的特徴量を直接最小化していないため，モデル化の精度が十分でないという問題があった．これに対し，LSTM-RNN では時間変動をモデル化できることができ，HMM に基づく手法より自然性が向上することが報告されているが [11], FF-NN を用いた場合との比較は行われていない．

本研究では，静的・動的特徴量系列から尤度最大基準により生成された静的特徴量系列を最小化する生成誤差最小

[1] 東北大学 大学院工学研究科
Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan
[a)] tnose@m.tohoku.ac.jp

(Minimum Generation Error, MGE) 基準を導入することにより精度の改善を図る．MGE 基準による学習は DNN 音声合成においてその有効性が示されている [12], [13]．また，画像生成分野で広く用いられている敵対的学習の一つである Generative Adversarial Network(GAN)[14] を利用し，主観的な品質の向上に取り組む．さらに，シェイプパラメータの低次が顔の向きを表すことを利用し，低次パラメータの置換により土台となる動画の顔の向きを生成動画の顔に反映できることを示す．

2. DNN に基づく統計的パラメトリック発話動画像生成

DNN に基づく統計的パラメトリック発話動画像生成では，まず学習データとして目標となる話者の発話動画像データを用意し，これに対し AAM を適用しシェイプおよびアピアランスパラメータを抽出する．これらの学習用パラメータを用いて FF-NN や LSTM-RNN 等によりモデル化・生成を行い，最後に各フレームの画像を再合成することで動画像を生成する．本節ではこれらについて概説するとともに，その問題点について述べる．

2.1 AAM による顔画像の分析合成

AAM はシェイプ (顔の形状) とアピアランス (見え方) を，基準となる顔からの変動を元にパラメータ化することで任意の顔を表現する 2 次元顔モデルである．具体的にはあるフレームのシェイプおよびアピアランスはそれぞれ以下のようなバイアスベクトルおよび基底ベクトルの線形結合により表現される．

$$\boldsymbol{s} = \boldsymbol{s}_0 + \sum_{i=1}^{m} p_i \boldsymbol{s}_i \tag{1}$$

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \boldsymbol{A}_i(\boldsymbol{x}) \qquad \forall\, \boldsymbol{x} \in \boldsymbol{s}_0 \tag{2}$$

ここで，$\boldsymbol{s}_0$ は基準となる平均形状を表すバイアスベクトル，$\boldsymbol{s}_i$ は平均形状からの差分を表す基底ベクトルであり，重み係数 p_i がシェイプパラメータとなる．同様に，$\boldsymbol{A}(\boldsymbol{x})$ は平均形状内のあるピクセルの座標 $\boldsymbol{x}$ における色情報を表し，$\boldsymbol{A}_0(\boldsymbol{x})$ および $\boldsymbol{A}_i(\boldsymbol{x})$ はそれぞれバイアスベクトルおよび基底ベクトルであり，重み係数 λ_i がアピアランスパラメータとなる．シェイプパラメータはあらかじめ OpenFace[15] などを用いて得られた学習データの各フレームの顔特徴点に対し主成分分析を行うことで求めることができる．また，アピアランスパラメータは各フレームの顔特徴点から平均形状 $\boldsymbol{s}_0$ への写像を用いて各フレームのアピアランスを平均形状へと正規化した後，主成分分析を適用することで得られる．

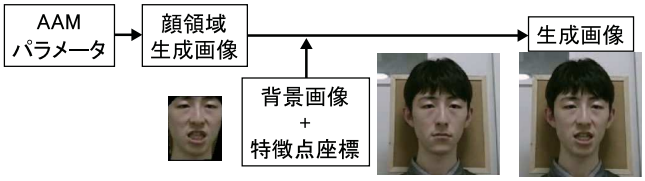

図 1 画像再構成の概要

2.2 DNN による AAM パラメータ系列のモデル化と生成

学習データの AAM パラメータ系列に対し DNN によりモデル化を行う．この際，音素情報と各フレームの音素内相対位置をコンテキストとして利用することで発話に伴う口唇形状を中心とした変動を表現する．FF-NN を用いる場合には時間変化パターンを考慮し滑らかなパラメータ系列を予測するため，AAM パラメータ系列 (静的特徴量) からその動的特徴量を計算しこれらを合わせた特徴量を DNN でモデル化する．通常，学習の際には，平均二乗誤差最小 (Minimum Mean Square Error, MMSE) 基準によりネットワーク重みを推定する．なお，LSTM-RNN を用いた場合には時間変動をモデル化できることができるため，動的特徴量は不要である．

発話動画像の生成時には，与えられた入力テキストからコンテキストを求め，ネットワークに入力することで AAM パラメータを生成する．この際，静的・動的特徴量を用いた場合にはこれらから HMM 音声合成 [16] で用いられる最尤パラメータ推定法 [17] により動的特徴量を考慮したパラメータを生成する．この AAM パラメータ系列から式 (1) および (2) によりシェイプおよびアピアランスを求め，アピアランスをシェイプに基づいて区分アフィン変換することで各フレームの顔画像データが得られる．ただし，ここで得られるのは顔領域部分のみであり，実際には，図 1 に示すように，土台となる発話動画像を予め用意し，それの顔領域部分に生成した顔画像を貼り付けて再構築することで最終的な発話動画像が得られる．

2.3 従来法の問題点

統計的パラメトリック発話動画像生成は比較的少量の学習データでも滑らかで安定した動画像が生成できるという特徴がある一方，いくつか問題点もある．まず，静的・動的特徴量を用いた場合にはこれらの結合特徴量について平均二乗誤差を最小化するように学習するため，実際の静的特徴量系列については必ずしも誤差が最小化されていないという点が挙げられる．この問題については以前から HMM 音声合成および DNN 音声合成においてそれぞれ指摘されており，これに代わり MGE 基準により学習を行うことでパラメータの予測精度と主観品質が改善することが報告されている [13], [18].

次に統計的学習による汎化作用の問題がある．これにつ

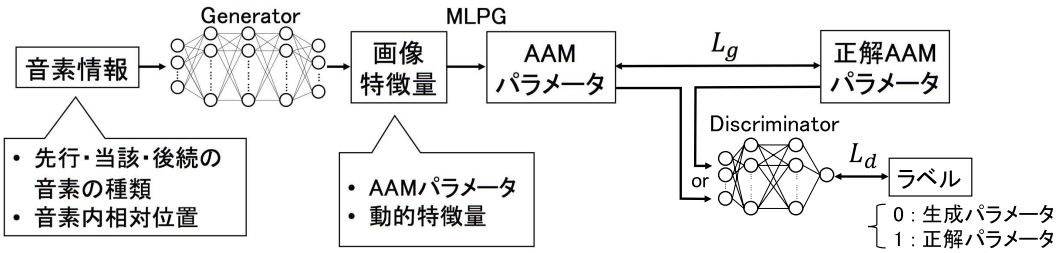

図 2 敵対的学習の概要

いても音声合成において過剰平滑化問題として知られており，画像においては変化の激しい口唇部分において，ぼやけた画像が生成され主観品質が低下する傾向がある．音声合成においては，この過剰平滑化問題は系列内変動を考慮したパラメータ生成 [19] やアフィン変換による分散補償 [20] などにより改善することが報告されているが，本研究では GAN を用いた手法 [21] を検討する．

最後に顔の向きの問題が挙げられる．従来の 2D を対象とした発話動画像生成では基本的に顔は正面を向いて固定されていることが前提となっている．しかし，実際の人間は話すときには顔の向きについてもある程度変動があることが普通であり，顔の向きが正面に固定された動画は自然性に欠ける．これについて本研究では AAM のパラメータ置換に基づく手法を検討する．

3. 自然で高精細な発話動画像生成

本研究では，2.3 節で述べた問題を軽減し，従来より自然で高精細な発話動画像を生成するため，MGE 基準によるモデル学習，敵対的学習，AAM パラメータの置換による顔の向きの考慮，の 3 つの手法を導入する．

3.1 MGE 基準によるモデル学習

従来の MMSE 基準による学習では，ネットワークの出力と正解の静的・動的特徴量との平均二乗誤差を最小とするようなネットワーク重みを学習していた．これに対し，MGE 基準による学習ではネットワークから出力された静的・動的特徴量を用いて発話毎に一旦発話全体のパラメータ系列を生成しそれと正解の静的特徴量系列との誤差を最小化する．これにより，学習時と生成時での基準の不一致が無くなり，より精密なモデル化が可能となる．具体的な損失関数は以下の式で表される．

$$L_{\mathrm{MGE}}(\boldsymbol{y},\hat{\boldsymbol{y}}) = -P(\boldsymbol{y}|\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{y}|\hat{\boldsymbol{y}},\boldsymbol{P}) \tag{3}$$

ここで $\boldsymbol{y}$ は静的特徴量系列，λ はネットワーク重みであり，$\boldsymbol{P}$ および $\hat{\boldsymbol{y}}$ はそれぞれ次式で与えられる．

$$\boldsymbol{P} = \left(\boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{W}\right)^{-1} \tag{4}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{\mu} \tag{5}$$

ここで $\boldsymbol{W}$ は静的特徴量系列から静的・動的特徴量系列を求める窓行列であり，$\boldsymbol{\mu}$ は各フレームにおけるネットワークの出力である静的・動的特徴量 $\boldsymbol{\mu}_t$ を発話単位で連結したパラメータである．同様に $\boldsymbol{U}$ は各フレームにおける静的・動的特徴量の共分散行列 $\boldsymbol{U}_t$ を発話単位で連結したパラメータであるが，文献 [22] 等では簡単のため，予め学習データ全体から求めた静的・動的特徴量系列の共分散行列を発話単位分複製し連結したパラメータで代用しており，本研究でも同様の手法を取る。一方，Mixture Density Network[23] 等を用いて共分散行列もネットワークの出力としてフレーム単位で直接モデル化する手法も提案されている [24], [25].

3.2 敵対的学習の導入

敵対的学習の一つである GAN[14] は生成モデルである生成器と識別モデルである識別器の 2 つのネットワークから成る．生成器は学習データにできる限り似たようなデータを生成し，識別器は入力が生成器によって生成されたものか学習データかを判定する．これらの 2 つのネットワークを敵対させるように学習を進めることで，最終的に生成器は学習データに区別がつかないほど似たデータを生成することが可能になる．GAN を音声合成に導入した先行研究 [21] では，GAN によりパラメータの予測精度が向上することで，従来の MGE 学習による音声合成手法と比較して生成パラメータの系列内変動 (Global Variance, GV) の平均が正解に近づき，合成音声の自然性も向上することが報告されている．

図 2 に GAN を用いた学習の概要を示す．生成器単体の損失関数 $L_{\mathrm{g}}(\boldsymbol{y},\hat{\boldsymbol{y}})$ は，式 (3) で示した MGE 学習における損失関数と同一であり，$L_{\mathrm{g}}(\boldsymbol{y},\hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y},\hat{\boldsymbol{y}})$ と表される．また識別器の損失関数 $L_{\mathrm{d}}(\boldsymbol{y},\hat{\boldsymbol{y}})$ は次の式で表される cross-entropy 関数として与えられる．

$$L_{\mathrm{d}}(\boldsymbol{y},\hat{\boldsymbol{y}}) = L_{\mathrm{d_o}}(\boldsymbol{y}) + L_{\mathrm{d_g}}(\hat{\boldsymbol{y}}) \tag{6}$$

$$L_{\mathrm{d_1}}(\boldsymbol{y}) = -\frac{1}{T_n}\sum_{t=1}^{T_n}\log D(\boldsymbol{y}_t) \tag{7}$$

$$L_{\mathrm{d_0}}(\hat{\boldsymbol{y}}) = -\frac{1}{T_n}\sum_{t=1}^{T_n}\log\left(1 - D(\hat{\boldsymbol{y}}_t)\right) \tag{8}$$

$L_{\mathrm{d}_1}(\boldsymbol{y})$，$L_{\mathrm{d}_0}(\hat{\boldsymbol{y}})$ はそれぞれ正解パラメータ，生成パラメータに対する損失であり，$D(\boldsymbol{y}_t)$ は入力パラメータが正解パラメータである事後確率を表す．すなわち識別器は入力が正解パラメータのときに 1 を，入力が生成パラメータのときは 0 をそれぞれ出力するように学習が行われる．

GAN の構造では，次の式で表される 2 つの損失の重み付き和を最小化するように生成器を学習する．

$$L_{\mathrm{GAN}}(\boldsymbol{y},\hat{\boldsymbol{y}}) = L_{\mathrm{g}}(\boldsymbol{y},\hat{\boldsymbol{y}}) + \omega\frac{E_{L_{\mathrm{g}}}}{E_{L_{\mathrm{d}}}}L_{\mathrm{d}_1}(\hat{\boldsymbol{y}}) \qquad (9)$$

ω は識別器に対する損失へ付与する重みを表すハイパーパラメータであり，$\omega = 0$ とすれば MGE 学習と等価になる．先行研究 [21] では，ω の値について検討を行った結果 $\omega \geq 0.2$ であれば生成結果に差が生じなかったことが報告されている．そこで本研究では ω の値を 1 に固定し実験を行った．また $E_{L_{\mathrm{g}}}$,$E_{L_{\mathrm{d}}}$ はそれぞれ生成器，識別器の損失関数 $L_{\mathrm{g}}(\boldsymbol{y},\hat{\boldsymbol{y}})$，$L_{\mathrm{d}}(\boldsymbol{y},\hat{\boldsymbol{y}})$ の期待値を表し，2 つの損失のスケーリングを行うパラメータである．式 (6) および (9) をそれぞれ最小化するように識別器と生成器を交互に学習させることで，最終的な生成モデルを構築する．

3.3 顔の向きを考慮した発話動画像生成

AAM パラメータは学習データに対する主成分分析を用いて計算されるため，シェイプパラメータの低次は学習データ中においてシェイプの動きの大きい部分を表している．このため，顔を正面に向けて固定しておらず自然な動きをしている際には，低次シェイプパラメータは顔の全体的な動きを表し，それ以外のまぶたや口などの局所的な動きは高次パラメータに含まれると考えられる．これを確認するため，6 節で使用する安倍首相の自然発話動画を学習データとして得られたシェイプの 1 から 6 次元目のパラメータについて，図 3 に平均形状における各顔特徴点からの向きと変動量を示す．図中の赤と青の線分はそれぞれの正負の変動を示しており，線分の方向は特徴点座標の変動の方向を，線分の長さはその変動の大きさを表している．図より，1 次元目は上下に頷くような顔の回転，2 次元目は首をかしげるような顔の回転，3 次元目は首を横に振るような顔の回転を指すような固有軸が見てとれる．4 次元目以降は顔全体というよりは唇や顎の変動が表現されており，口の動きを制御しているパラメータであることが推測される．すなわち，このデータにおいては，顔の向きの情報は主にシェイプパラメータの 1 から 3 次元目に含まれていると言える．

本研究では顔の向きを含む情報としてシェイプパラメータの低次成分に着目し，このパラメータの置換によって土台となる顔の向きを生成された顔動画像に反映する．具体的には低次については土台となる発話動画像のシェイプパラメータを，それ以外については生成されたシェイプパラメータを用いてシェイプの再構築を行う．

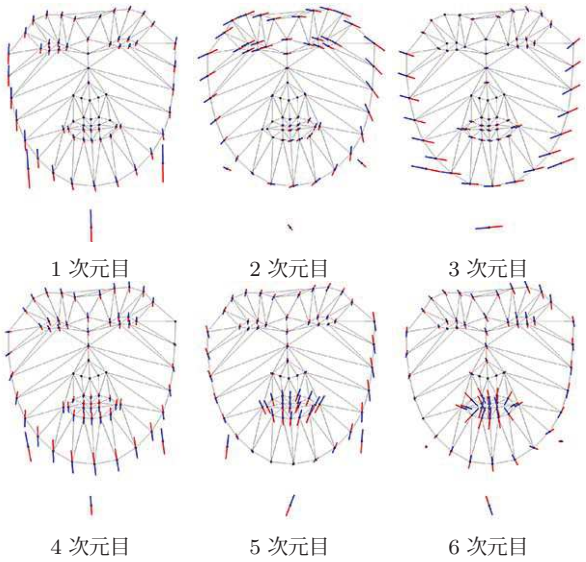

図 3 6 節で用いる学習データから得られた基底ベクトル

表 1 共通のネットワーク条件

入力層ユニット数	103
Optimizer	Adagrad[26]
活性化関数	Leaky ReLU[27]
バッチサイズ	5 文
Dropout 率	0.5

4. 学習基準の比較

まず，学習基準として以下の 3 つのネットワークによる手法を比較した．

- 2.2 節で述べた MMSE 基準による FF-NN(**MMSE**)
- 3.1 節で述べた MGE 基準による FF-NN(**MGE**)
- MMSE 基準による LSTM-RNN[10](**LSTM**)

4.1 実験条件

予備実験により表 1 に示した共通のネットワーク構造を用いた．**MMSE** および **MGE** では出力層のユニット数は 3168 であり，中間層数は 5，中間層のユニット数は 2048，エポック数は 30 とした．**LSTM** では，先行研究 [10] を参考にネットワークの層構造を全結合層 1 層，双方向 LSTM 層 2 層の計 3 層とし，各層のユニット数とエポック数についてのみ開発データセットを用いてチューニングを行った．ユニット数のチューニングは，ユニット数 64・128・256・512・1024・2048 のうち，開発データに対する損失の最小値を 3 回ずつ算出し，その平均値が最も小さくなるユニット数を選択した．またそのときのエポック数から実験に用いるエポック数を決定した．また出力ベクトルとして，アピアランスパラメータ 1024 次元とシェイプパラメータ 32 次元の計 1056 次元のベクトルに対し，平均が 0，分散が 1 になるよう正規化したベクトルを使用した．学習の際は **MMSE** と同様にミニバッチごとに正解パラメータとの平均二乗誤差を最小化するように誤差逆伝搬方でネットワー

表 2　学習基準についての客観評価結果

	RMSE		
	MMSE	MGE	LSTM
アピアランス	22.30	21.29	21.30
シェイプ	1.237	1.208	1.268

クの重みを更新した．

入力ベクトルとして，0 から 1 の間で表した音素内相対フレーム位置と，先行，当該，後続の音素の種類を 0 または 1 で表現したバイナリベクトル 102 次元の計 103 次元のベクトルに対し，最小値が 0.01，最大値が 0.99 になるように正規化したベクトルを使用した．また出力ベクトルとして，アピアランスパラメータ 1024 次元とシェイプパラメータ 32 次元，それらの一次，二次の動的特徴量を加えた計 3168 次元のベクトルに対し，平均が 0，分散が 1 になるよう正規化したベクトルを使用した．学習データセットは 17643 フレームで学習を行い，評価データセット 8050 フレームで以下に述べる客観評価・主観評価を行った．客観評価実験では 3 手法によって生成されたパラメータに対し，1 発話ごとの平均 RMSE を比較した．

主観評価実験では，3 手法によって生成された動画を用い，全組み合わせについて生成動画の再現性および自然性に関する対比較実験を行った．なおこれらの動画には全てデータセットに含まれる自然音声を付与した．再現性に関する対比較実験では，被験者はリファレンスの分析合成動画に続けて，各手法の動画をランダムな順で見た後，どちらの動画がリファレンス動画に近かったかを回答した．各組み合わせにおいて，2 手法のそれぞれ 10 動画に対して評価を行った．自然性に関する対比較実験では，被験者は各手法の動画をランダムな順で見たのち，どちらの動画が自然と感じたかを回答した．各組み合わせにおいて，2 手法のそれぞれ 10 動画に対して評価を行った．それぞれの主観評価実験において被験者は 8 名である．

4.2　客観評価結果

表 2 に各手法におけるアピアランス・シェイプの RMSE をそれぞれ示す．まず **MMSE** と **MGE** とを比較すると，それぞれのパラメータにおいて誤差が減少することがわかった．これは **MGE** において設定されている損失関数がより最終的なパラメータの誤差に近いことが影響していると考えられる．また **LSTM** を **MMSE**・**MGE** と比較すると，**LSTM** のアピアランスパラメータの RMSE は **MGE** と同程度の誤差であるのに対し，**LSTM** のシェイプパラメータの RMSE は **MMSE** よりも大きくなってしまうことがわかった．

4.3　主観評価結果

主観評価の結果を図 4，5 に示す．それぞれのグラフ内

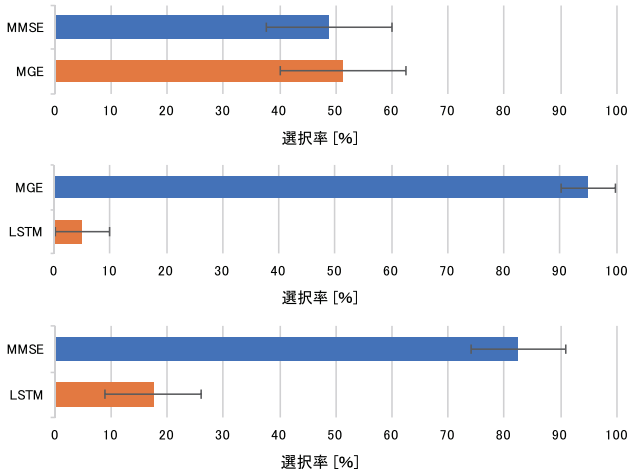

図 4　学習基準についての再現性の比較

のエラーバーは 95%信頼区間を表す．まず再現性に関する主観評価実験の結果から，**MGE** と **LSTM**・**MMSE** と **LSTM** をそれぞれ比較した際に有意差が現れているのに対し，**MMSE** と **MGE** とで比較した場合には有意差が生じなかった．一方で自然性に関する主観評価実験の結果から，自然性に関しては全ての手法間において有意差が生じ，**MGE** が最も自然性の評価が高くなることがわかった．**MMSE** で生成された動画は口が急峻に動く傾向があり，**MGE** で生成された動画では口の動きが **MMSE** よりも平滑化して見える傾向があった．本研究で用いたデータベースには比較的ハキハキとした口の動きが含まれているため，**MMSE** の口の動きが比較的リファレンス動画に近いと判断され **MGE** の再現性の評価と同等になり，自然性では逆に **MMSE** の口の動きの自然性が低いと判断され **MGE** の自然性の評価が有意に高くなったと考えられる．また **LSTM** で生成された動画は **MGE** よりもさらに口の動きが平滑化して見える傾向があったため，再現性・自然性ともに評価が低くなってしまったと考えられる．LSTM が性能を発揮できなかった要因の一つとして，学習データの少なさが挙げられる．先行研究 [10] では学習データとして約 470 文 (約 66000 フレーム) の動画を用いているのに対し，本研究ではその 1 割程度しか使用していないため，Bidirectional LSTM のように複雑なネットワーク構造を十分に学習することができなかったと考えられる．以上の結果から，発話動画生成においても MGE 学習が有効であることが示された．

5.　敵対的学習の効果

次に，敵対的学習の効果について比較実験を行なった．4 節の **MGE** に GAN を導入した手法を **MGE+GAN** と表記する．

5.1　実験条件

MGE については 4 節で使用した条件と同じものを用

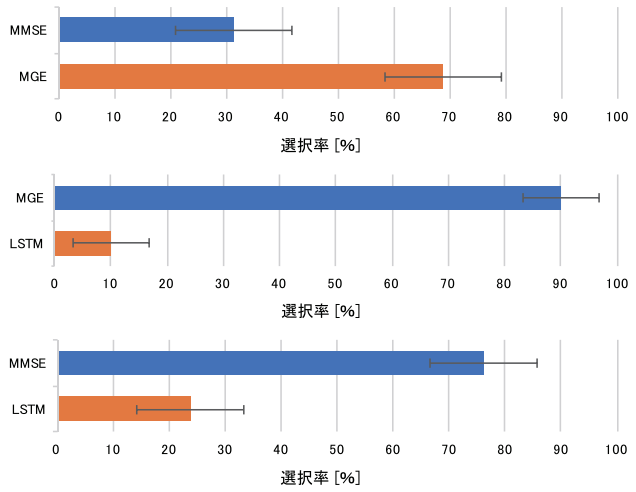

図 5 学習基準についての自然性の比較

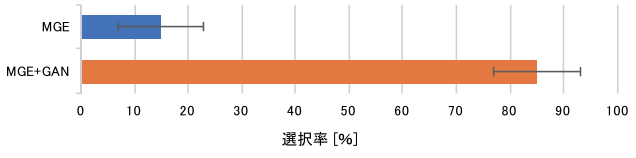

図 6 再現性における敵対的学習の効果

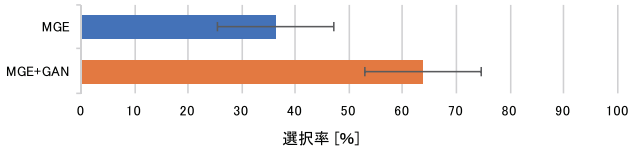

図 7 自然性における敵対的学習の効果

いた．**MGE+GAN** では，入力層および出力層のユニット数はそれぞれ 1056 と 1，中間層のユニット数は 1024 とした．それ以外の条件は表 1 と同じである．用いたデータセットや特徴量なども 4 節と同じとした．学習の際は，まず生成器をエポック数 30 で事前に学習しておき，その後生成器と識別器を組み合わせてエポック数 15 で学習を行った．

5.2 主観評価結果

4 節と同様に対比較による主観評価を行なった．2 手法のそれぞれ 10 動画に対して評価を行った．それぞれの主観評価実験において被験者は 8 名である．主観評価の結果を図 6，7 に示す．図中のエラーバーは 95%信頼区間を表す．結果より，再現性と自然性のいずれにおいても，敵対的学習を導入することで有意に主観品質が改善することがわかる．

6. 顔の向きの考慮の有無の影響

最後に 3.3 節で述べた顔の向きを考慮した場合について，考慮しない場合との比較を行なった．

6.1 実験条件

前節までに用いた動画像データは正面向きで顔を固定して収録したものであったため，本実験では新たに Youtube にて公開されている安倍首相の約 2 分のビデオメッセージ動画を用いた．この動画では安倍首相が身振り手振りを交え，時に顔をやや傾けるなど動かしながら話している様子を正面から撮ったものである．フレームレートは 29.97fps で，発話が途切れるタイミングで 25 発話に分割をして順にインデックスを付け，5 で割った剰余が 0 の 5 発話を評価データ，剰余が 4 の 5 発話を開発データ，それ以外の 15 発話を学習データとした．各フレームの画像は顔を中心とした部分について 500x360 のサイズでトリミングしたものを用いた．予備実験によりアピアランスとシェイプの次元数はそれぞれ 256 と 16 とし，中間層数は 3，中間層と出力層のユニット数はそれぞれ 256，816 とした．生成器はエポック数 20 で，識別器は 10 で事前に学習を行い，その後エポック数 30 で敵対的学習を行なった．それ以外の条件は前節と同じとした．

6.2 生成結果

図 8 に生成例を示す．上から順にオリジナルの動画像，5 節で評価した **MGE+GAN**，**MGE+GAN** に 3.3 節で述べた顔の向きの考慮を導入した **MGE+GAN+FD** である．図から，オリジナルの動画像には比較的大きい顔の動きが含まれており，**MGE+GAN** では顔の向きの情報が考慮せずに貼り付けを行なっているために顔領域内外で顔の向きのずれが生じていることがわかる．一方，**MGE+GAN+FD** ではオリジナルの動画像における顔の動きが反映され，貼り付けのずれが解消された．今回はデータセット全体で約 2 分という小規模なものであったが最低限口の動きが見てとれる発話動画像生成を行うことができた．

7. おわりに

本稿では従来の MMSE 基準による FF-NN による発話動画像生成に (1)MGE 基準による学習，(2) 敵対的学習，(3) 低次 AAM パラメータの置換による顔の向きの考慮，の 3 手法を導入し，その効果を評価した．主観評価により，MGE 基準の導入は主観的な自然性の向上に寄与することがわかった．また敵対的学習は再現性，自然性の両方を向上させる結果となった．また，顔の動きを伴う動画像については顔の向きを考慮しない場合には土台となる動画像との顔の向きの不一致の問題が生じることを確認し，低次 AAM パラメータの置換によりこの問題が低減されることを示した．今後の課題としては，学習データ量を増やした場合の評価，および系列内変動の考慮による口唇部分の再現性の向上が挙げられる．

謝辞 本研究の一部は，科学研究費補助金 (課題番号

オリジナル

MGE+GAN

MGE+GAN+FD

図 8 生成例

JP15H02720, JP16K13253, JP17H00823) の助成を得た.

参考文献

[1] Ostermann, J. and Weissenfeld, A.: Talking faces—technologies and applications, *Proc. the 17th International Conference on Pattern Recognition (ICPR)*, Vol. 3, pp. 826–833 (2004).

[2] Mattheyses, W. and Verhelst, W.: Audiovisual speech synthesis: An overview of the state-of-the-art, *Speech Communication*, Vol. 66, pp. 182–217 (2015).

[3] Cootes, T. F., Edwards, G. J. and Taylor, C. J.: Active appearance models, *European Conference on Computer Vision*, pp. 484–498 (1998).

[4] Zen, H., Tokuda, K. and Black, A.: Statistical parametric speech synthesis, *Speech Commun.*, Vol. 51, No. 11, pp. 1039–1064 (2009).

[5] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. M. and Deng, L.: Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends, *IEEE Signal Processing Magazine*, Vol. 32, No. 3, pp. 35–52 (2015).

[6] Gales, M.: Cluster adaptive training of hidden Markov models, *IEEE Trans. Speech Audio Process.*, Vol. 8, No. 4, pp. 417–428 (2000).

[7] Anderson, R., Stenger, B., Wan, V. and Cipolla, R.: Expressive visual text-to-speech using active appearance models, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3382–3389 (2013).

[8] Parker, J., Maia, R., Stylianou, Y. and Cipolla, R.: Expressive visual text to speech and expression adaptation using deep neural networks, *Proc. ICASSP*, pp. 4920–4924 (2017).

[9] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[10] Fan, B., Wang, L., Soong, F. K. and Xie, L.: Photo-real talking head with deep bidirectional LSTM, *Proc. ICASSP*, pp. 4884–4888 (2015).

[11] Fan, B., Xie, L., Yang, S., Wang, L. and Soong, F. K.: A deep bidirectional LSTM approach for video-realistic talking head, *Multimedia Tools and Applications*, Vol. 75, No. 9, pp. 5287–5309 (2016).

[12] 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一：ニューラルネットワークに基づく音声合成における系列内変動を考慮したトラジェクトリモデル学習，日本音響学会秋季研究発表会講演論文集，pp. 237–238 (2015).

[13] Wu, Z. and King, S.: Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 7, pp. 1255–1265 (2016).

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014).

[15] Baltrušaitis, T., Robinson, P. and Morency, L.-P.: OpenFace: an open source facial behavior analysis toolkit, *IEEE Winter Conference on Applications of Computer Vision* (2016).

[16] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthe-

sis, *Proc. Eurospeech*, pp. 2347–2350 (1999).
[17] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. and Imai, S.: An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features, *Proc. Eurospeech*, pp. 757–760 (1995).
[18] Wu, Y.-J. and Wang, R.-H.: Minimum generation error training for HMM-based speech synthesis, *Proc. ICASSP*, pp. 889–892 (2006).
[19] Toda, T. and Tokuda, K.: A Speech parameter generation algorithm considering global variance for HMM-based speech synthesis, *IEICE Trans. Inf. Syst.*, Vol. E90-D, No. 5, pp. 816–824 (2007).
[20] Nose, T.: Efficient Implementation of Global Variance Compensation for Parametric Speech Synthesis, *IEEE/ACM Trans. Audio, Speech and Language Process.*, Vol. 24, No. 10, pp. 1694–1704 (2016).
[21] Saito, Y., Takamichi, S. and Saruwatari, H.: Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 84–96 (2017).
[22] Zen, H., Senior, A. and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, *Proc. ICASSP*, pp. 7962–7966 (2013).
[23] Bishop, C. M.: Mixture density networks, Technical report (1994).
[24] Zen, H. and Senior, A.: Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis, *Proc. ICASSP*, pp. 3844–3848 (2014).
[25] 橋本佳，大浦圭一郎，南角吉彦，徳田恵一：DNN 音声合成における音響特徴量系列とその時間構造の同時モデル化，電子情報通信学会技術研究報告，Vol. 116, No. 414, pp. 71–76 (2017).
[26] Duchi, J., Hazan, E. and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *Journal of Machine Learning Research*, Vol. 12, No. Jul, pp. 2121–2159 (2011).
[27] Maas, A. L., Hannun, A. Y. and Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models, *Proc. icml*, Vol. 30, No. 1, p. 3 (2013).