

# Moment-matching network に基づく音声合成における 音声パラメータのランダム生成

高道 慎之介<sup>1,a)</sup> 郡山 知樹<sup>2,b)</sup> 猿渡 洋<sup>1,c)</sup>

概要： 本稿では，moment-matching network に基づく音声合成における音声パラメータのランダム生成アルゴリズムを提案する．同一の言語情報・パラ言語情報を付与しようとしても，人間は同一の音声を生成することは不可能だが，典型的な統計的音声合成は入力コンテキストに対して完全に同一の音声を生成する．自然音声の発話間変動を合成音声に付与するために，本論文では，音声パラメータのランダムサンプリングを可能にする Deep Neural Network (DNN) 音響モデルを構築する．DNN は合成音声パラメータのモーメントを自然音声パラメータのモーメントに一致させるように学習される．音声パラメータ変動は低次元のシンプルな事前ノイズベクトルに圧縮されるため，音声パラメータの直接的なサンプリングと比較して計算量を抑えたサンプリングが可能となる．実験的評価では，音声パラメータのランダム生成が合成音声品質を劣化させるかについて調査する．評価結果より，最尤生成と比較して提案法による音質低下は生じないことを明らかにする．

## Random generation of speech parameters in speech synthesis based on moment-matching networks

TAKAMICHI SHINNOSUKE<sup>1,a)</sup> KORIYAMA TOMOKI<sup>2,b)</sup> SARUWATARI HIROSHI<sup>1,c)</sup>

### 1. はじめに

統計的音声合成 [1] は統計モデルを使用して音声を合成する方法であり，音声合成の最終目標の 1 つは人間の発話のように自然な音声を合成することである．音声品質は自然性の要素の 1 つであり，合成音声の品質向上のための様々な方法が提案されている [2], [3], [4]．特に，Deep Neural Network (DNN) に基づく音声合成 [5], [6] は，合成音声の品質を著しく向上させた．しかし，音声品質は自然性の要素の 1 つに過ぎず，合成音声の自然性は他の基準から評価される必要がある．

本稿は，新たな基準として同一コンテキストにおける

発話間変動 [7] を考慮する．従来の DNN 音声合成は最小誤差基準に基づいて合成音声を生成するため，Fig. 1 に示すように，入力コンテキストを固定した場合，合成音声は常に同一であり録音再生された音声に過ぎない．故に，従来の音声合成技術を利用した音声コミュニケーションシステムは，人間同士ではあり得ないワンパターンなコミュニケーションを行ってしまう．一方，人間の音声生成はランダム性を有するため，同一の言語情報・パラ言語情報を付与しようとしても人間は発話毎に異なる音声を生成する．本稿では，このような発話間変動を持つ音声コミュニケーションシステムの確立を見据え，発話間変動を合成音声に付与する方法を検討する．発話間変動を付与する直接的な方法は，同一の言語情報・パラ言語情報を持つよう繰り返し発話された音声データを用いて，発話間変動を明示的にモデル化することである．しかし，そのような音声データは統計的音声合成の典型的な学習データに含まれない．別の方法は，適切な確率分布から音声パラメータをランダムサンプリングする方法である．Shannon ら [8] は，トラ

<sup>1</sup> 東京大学 大学院情報理工学系研究科  
University of Tokyo, Engineering bldg. #6, 7-3-1 Hongo,  
Bunkyo-ku, Tokyo 113-8656, Japan.

<sup>2</sup> 東京工業大学  
Tokyo Institute of Technology, Japan.

a) shinnosuke\_takamichi@ipc.i.u-tokyo.ac.jp

b) koriyama@ip.titech.ac.jp

c) hiroshi\_saruwatari@ipc.i.u-tokyo.ac.jp

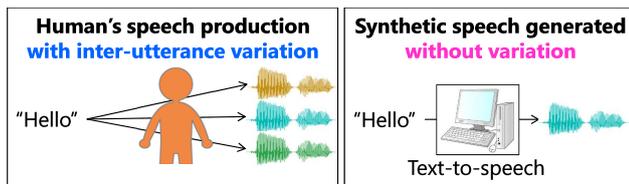


図 1 自然音声と典型的な合成音声の比較．入力コンテキストを固定した場合，自然音声は発話間ゆらぎを持つが，合成音声はゆらがない．

Fig. 1 Comparison of natural and synthetic speech. When input context is fixed, humans' speech has variation between utterances, but conventional synthetic speech does not.

ジェクトリ隠れマルコフモデル [9] を用いたランダムサンプリングを評価しており，ランダム生成された合成音声の品質が，最尤生成された合成音声の品質よりも著しく低下することを報告している．品質劣化の 1 つの理由は隠れマルコフモデルによる時間量子化 [10] であるため，トラジェクトリ DNN [11] または mixture density network [12] からのランダムサンプリングが品質劣化の緩和に有効であると期待される．しかしながら，そのような複雑な分布からのサンプリングは，計算コストが大きく実用には不向きである．

本稿では，moment-matching network を用いた音声パラメータのランダム生成法を提案する．DNN 音響モデルは，自然音声パラメータと合成音声パラメータのモーメントを一致させるように学習される．音声パラメータの変動は低次元のシンプルな事前ノイズベクトルに圧縮され，DNN はそのノイズを音声パラメータ変動に変形する．合成時には，ランダムサンプリングされた事前ノイズを用いて，合成音声パラメータをランダムサンプリングする．パラメータ変動はシンプルな事前ノイズとして表されるため，音声パラメータの直接的なサンプリングと比較して提案法の計算コストは小さい．本稿では，自然な発話間変動を持つ音声合成の構築に向け，提案するランダム生成法が合成音声品質を劣化させるかについて調査する．実験的評価では，最尤生成法とランダム生成法の音声品質を比較し，ランダム生成による音質劣化が生じないことを明らかにする．

## 2. 従来の統計的音声合成と音声パラメータのランダムサンプリング

従来の DNN 音声合成では，自然音声パラメータと生成音声パラメータの間の平均二乗誤差を最小にするように，DNN 音響モデルを学習する．この学習基準は，音声パラメータの確率分布を等方性ガウス分布（等方性共分散行列を有するガウス分布）とみなした最尤学習と等価であり，学習時にはガウス分布の平均ベクトルのみが推定される．合成時には，入力コンテキストが与えられた後，学習と同様に最尤基準に基づいて音声パラメータを生成する．した

がって，入力コンテキストを固定した場合，生成される音声パラメータは常に同一である．

適切な確率分布からのランダムサンプリングにより発話毎に異なる音声パラメータを生成できる．音声パラメータの時間遷移制約を有するトラジェクトリモデル（パラメータ系列長のサイズの全共分散正規分布）[8], [11] や，混合分布（例えば，混合正規分布）をモデル化できる mixture density network [12] は，等方性ガウス分布よりも適切な確率分布である．しかしながら，これらのような複雑な分布からのサンプリングは計算コストが高く，実用には不向きである．

## 3. Moment-matching network に基づく音声合成

本節では，moment-matching network を導入し，この DNN を用いた音声パラメータのランダム生成法を提案する．DNN の学習基準は，パラメトリックな分布（例えば，等方性正規分布 [5]，全共分散正規分布 [11] や，混合正規分布 [12]）ではなく，モーメント差を用いたノンパラメトリックなモデル化に基づく．

### 3.1 Moment-matching network

#### 3.1.1 Maximum Mean Discrepancy (MMD) の最小化 [13]

$y = [y_1^\top, \dots, y_t^\top, \dots, y_T^\top]^\top$  と  $\hat{y} = [\hat{y}_1^\top, \dots, \hat{y}_t^\top, \dots, \hat{y}_T^\top]^\top$  をそれぞれ，学習データに含まれるパラメータ系列，及び，DNN から生成されたパラメータ系列とする． $T$  は系列長である． $y_t$  と  $\hat{y}_t$  はそれぞれ，フレーム  $t$  における学習パラメータ及び生成パラメータである． $\hat{y}$  をランダムサンプリングする DNN は， $y$  と  $\hat{y}$  間のモーメントの差の二乗を最小化するように学習される．この学習基準は，(kernelized) Maximum Mean Discrepancy (MMD) の二乗として知られ，以下の式で示される．

$$L_{\text{MMD}}(y, \hat{y}) = \frac{1}{T^2} \{ \text{tr}(\mathbf{1}_T \cdot \mathbf{K}_y(y, y)) + \text{tr}(\mathbf{1}_T \cdot \mathbf{K}_y(\hat{y}, \hat{y})) - 2 \cdot \text{tr}(\mathbf{1}_T \cdot \mathbf{K}_y(y, \hat{y})) \}, \quad (1)$$

ここで， $\text{tr}(\cdot)$  は行列のトレース， $\mathbf{1}_T$  は全ての要素が 1 の  $T$ -by- $T$  の行列， $\mathbf{K}_y(y, \hat{y})$  は， $y$  と  $\hat{y}$  間のグラム行列であり，その  $t$  行  $\tau$  列目の要素は， $y_t$  と  $\hat{y}_\tau$  の分布間のカーネルである．カーネル関数としてガウスカーネルを使用する場合，無限次元までのモーメントの差を学習時に考慮する．低次元のノイズベクトル  $n$  を入力に持つ DNN は，損失関数  $L_{\text{MMD}}$  を最小化するように学習される．このノイズベクトルは既知のシンプルな確率分布からランダムサンプリングされる．Fig. 2 に示すように，ここで学習された DNN

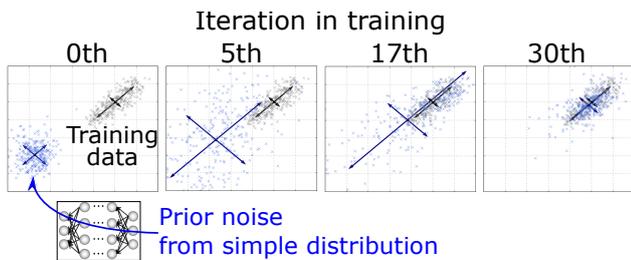


図 2 Moment-matching network の学習．学習データと生成データの分布およびそれらの 1 次と 2 次のモーメントを示す．DNN は，入力のシンプルなノイズの分布を学習データの分布に変形させる役割をもつ．

Fig. 2 Training of moment-matching networks. Distributions of training and generated data are drawn with their 1st and 2nd moments. Networks are trained to transform prior simple noise into data distribution.

はシンプルな確率分布を学習データの経験分布に変形する役割を持つ．

### 3.1.2 条件付き MMD の最小化 [14]

3.1.1 節の手法は，条件付き分布のモーメント差の最小化に拡張可能である． $y$  に対応する入力特徴量系列を  $x = [x_1^T, \dots, x_t^T, \dots, x_T^T]^T$  とすると， $\tilde{x} = [x^T, n^T]^T$  を入力とする DNN は，次式の条件付き MMD を最小化するように学習される．

$$L_{\text{CMMD}}(\tilde{x}, y, \hat{y}) = \frac{1}{T^2} \{ \text{tr}(G(\tilde{x}) \cdot K_y(y, y)) + \text{tr}(G(\tilde{x}) \cdot K_y(\hat{y}, \hat{y})) - 2 \cdot \text{tr}(G(\tilde{x}) \cdot K_y(y, \hat{y})) \}, \quad (2)$$

$$G(\tilde{x}) = \tilde{K}_x^{-1}(\tilde{x}) K_x(\tilde{x}) \tilde{K}_x^{-1}(\tilde{x}), \quad (3)$$

$$\tilde{K}_x(\tilde{x}) = K_x(\tilde{x}) + \lambda I_T, \quad (4)$$

ここで， $I_T$  は  $T$ -by- $T$  の単位行列であり， $\lambda$  は正則化の重みである． $K_x(\tilde{x})$  は  $\tilde{x}$  のグラム行列である．

ランダム生成時には，所望の入力特徴量  $x$  とサンプリングされた  $n$  を DNN に入力することで， $\hat{y}$  をランダム生成する．

## 3.2 Moment-matching network を用いた音声合成と音声パラメータのランダムサンプリング

Moment-matching network を使用した音声パラメータのランダム生成法を提案する．Fig. 3 に示すように，DNN は条件付き MMD を最小化するように学習される． $x$  と  $y$  は入力テキストのコンテキストベクトル系列と合成音声のパラメータ系列である．ノイズベクトル  $n$  はシンプルな分布からフレーム毎にサンプリングされる．この DNN は出力音声パラメータの静的・動的特徴量を予測し，最終的な  $\hat{y}$  はこれらの特徴を考慮して生成される [16]．合成時には，コンテキストベクトルとノイズベクトルを決定した後，通常の生成処理 [5] により  $\hat{y}$  をランダムサンプリングする．

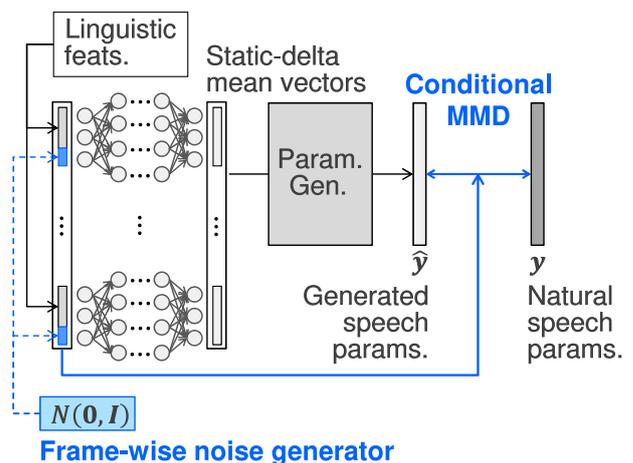


図 3 Moment-matching network を用いた音声パラメータのランダム生成，図示を簡単化するため，ここでは DNN の入力を言語特徴量（コンテキスト）としているが，実際には，別の DNN から得られた bottleneck 特徴量 [15] を入力としていることに注意する．

Fig. 3 Sampling-based speech parameter generation using moment-matching networks. Note that linguistic features are directly used in this figure for clear illustration, but bottleneck features [15] are used in place of linguistic features in actual implementation.

コンテキスト要素の大部分は 1-of- $K$  ホットベクトルであるため，コンテキスト要素間のカーネル関数は効果的ではない．そこで我々は，カーネルを計算するために，コンテキストベクトルの代わりに bottleneck 特徴量を利用する． $x$  から  $y$  を予測する別の Feed Forward neural network を，平均二乗誤差基準 [5] で学習する，カーネルは，特定の隠れ層の値を用いて計算される．

## 3.3 考察と従来法との比較

条件付き MMD はパラメトリックな分布を仮定しないため，提案アルゴリズムは mixture density network [12] やトラジェクトリ DNN [11] よりも複雑な分布をモデル化できる．さらに，音声パラメータ変動が低次元の事前ノイズベクトルに圧縮されるため，提案法は，上記のモデルからのサンプリングと比較して計算コストが小さい．

Generative Adversarial Network (GAN) [17] と条件付き GAN [18] は，提案法と同じく，複雑な分布をモデル化できる手法である．GAN の学習はミニマックス問題であるため，その最適化には経験的な知見が必要であることが知られている [19]．我々はこれまでに GAN を含めた音声合成法を提案している [4] が，この手法と比較して提案法の学習は容易である．これは，提案法の学習基準が条件付き MMD の単なる最小化問題であるためである．ここでさらに，GAN 及び提案法 (moment-matching network) と，従来の音声処理技術の関係性について説明する．GAN は，自然音声と合成音声の分布間の divergence (例えば，

Jensen-Shannon divergence [17] や  $f$ -divergence [20]) を最小化する．故に、音源分離におけるスパース性の議論で用いられる  $\beta$ -divergence [21], [22] などに関連する技術である．一方、moment-matching network はモーメントの差を明示的に使用する．故に、系列内変動 [23], 変調スペクトル [3], カートシス [24] に基づいた高次統計量復元・追跡などに関連する技術である．

発話間変動を付与する従来技術として、文レベルのコンテキストの付与 [25] がある．この手法は、発話者が意図的に付与した音声表現を合成音声に付与することに相当するが、提案法は、発話者の意図しないランダム性を付与することに相当する．

最後に、合成音声の品質基準としての音声なりすまし検出技術について述べる．声のなりすましを検出する Anti-Spoofing Verification (ASV) [26] を詐称することは、合成音声の品質基準となる [4]．ASV の技術のひとつに、提示された音声は自然音声か録音音声かを識別する replay-attack 検出技術 [27] がある．この技術では、事前録音音声と提示音声の一致度によって音声を検出する．従来の合成法で繰り返し合成された音声は常に同一であるため、従来の音声合成は replay-attack 検出技術によって容易に検出される．一方、提案法は発話間変動をもつため、検出を緩和可能である．

## 4. 実験的評価

### 4.1 実験条件

学習データは日本人女性 5 名による ATR 音素バランス 503 文 A-I セット 450 文 (計 2250 文) [28] であり、評価データは内 1 名による J セット 53 文である．学習データのサンプリング周波数は 16 kHz, フレームシフトは 5 ms とする．スペクトルパラメータとして STRAIGHT 分析 [29] による 0 次から 24 次のメルケプストラム係数, 音源パラメータとして  $F_0$ , 5 周波数帯域における平均非周期成分 [30], [31] を用いる．スペクトルパラメータには 50 Hz 変調周波数のトラジェクトリスムージング [32] を施す．コンテキストラベルは、音素などからなる 274 次元ベクトルと 5 次元の話者 ID [33] である．音響モデルの入力特徴量は、128 次元の bottleneck 特徴量, 平均 0, 分散 1 の正規分布に従う 3 次元のノイズベクトルである．音響モデルの出力特徴量はスペクトルパラメータの静的・動的特徴量 (75 次元) である． $F_0$ , 非周期成分, 継続長は自然音声の特徴量を使用する．音響モデルは、Feed-Forward neural network であり、隠れ層数は 3, 隠れ層の素子数は 512, 隠れ層及び出力層の活性化関数は、それぞれ ReLU と線形関数である．Neural network のコンテキスト特徴量及びスペクトルパラメータは、それぞれ平均 0, 分散 1 に正規化する．提案法における正則化係数  $\lambda$  は 0.01 とし、 $y$  に関するカーネル関数として、ガウスクーネル  $\exp\{-\|y_t - \hat{y}_\tau\|^2/\sigma^2\}$  を

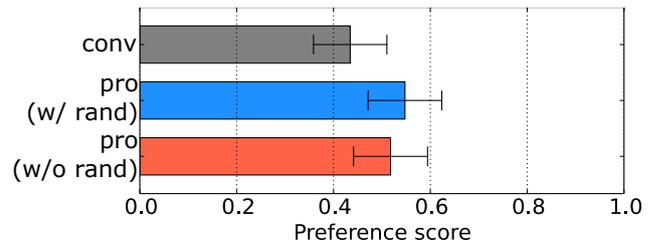


図 4 音質に関する主観評価結果 (エラーバーは 95%信頼区間)  
Fig. 4 Preference scores on speech quality with 95% confidence interval.

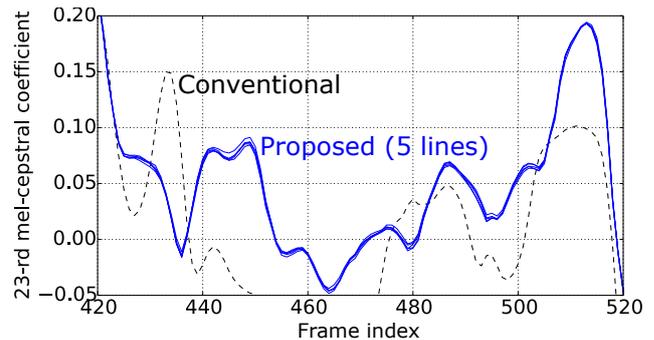


図 5 生成されたパラメータ系列の例．提案法を用いて 5 つの系列をランダム生成した．

Fig. 5 Example of generated speech parameter trajectories. We sampled five trajectories using proposed method.

使用する． $\sigma$  はガウスクーネルの指数部が  $-1$  以上になるように設定する [14]． $x$  に関するカーネル関数も同様に決定する．

本稿では、最尤生成と比較してランダム生成が音質を低下させるか [8] について調査する．評価する合成音声は以下の 3 つである．

- conv: 二乗誤差最小基準を用いる従来の音声合成 [5]
- pro (w/ rand): 提案法によるランダム生成
- pro (w/o rand): 提案法による最尤生成

“pro (w/o rand)” は、“pro (w/ rand)” と同様に学習されるが、生成時にノイズベクトルを最尤推定で固定する (すなわち、 $n = 0$ )．故に、“pro (w/o rand)” は発話間変動を有さない．

主観評価として、音質に関するプリファレンス AB テストを実施する．被験者数は 7 人である．

### 4.2 実験結果

Fig. 4 に主観評価結果を示す．Moment-matching network を用いた提案法において、音声パラメータを最尤生成した場合とランダム生成した場合で音質の劣化はみられない．故に、提案法は Fig. 5 に示すように生成毎に異なる音声パラメータ系列を生成しつつも、従来技術 [8] のような音質劣化を生じさせないことが明らかになった．また、提案法の音質は従来の音声合成の音質を上回ることが分かる．この改善は、従来の確率分布である等方性正規分布に

よるモデリングと、提案法のノンパラメトリックモデリングの違いによるものと思われる。

## 5. まとめ

人間は同じ言語情報・パラ言語情報を持つよう発話しても発話毎に異なる音声を生成するが、従来の統計的音声合成は、同一コンテキストに対して完全に同一の音声を生成する。本稿では、自然な発声間変動を合成音声に与えるために、moment-matching network を用いた音声パラメータのランダム生成法を提案した。Neural network は、学習データと生成データ間の条件付き maximum mean discrepancy を最小化するように学習される。実験の評価から、提案するランダム生成法は、最尤生成法と比較して音質劣化を生じさせないことを明らかにした。今後は、提案法について詳細な調査を行う。

謝辞: 本研究の一部は、JSPS 科研費 16H06681 及びセコム科学技術支援財団の助成を受け実施した。

## 参考文献

- [1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [2] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 239–250, 2014.
- [3] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.
- [4] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *Proc. ICASSP*, Orleans, U.S.A., Mar. 2017.
- [5] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013.
- [6] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," vol. abs/1609.03499, 2016. [Online]. Available: <http://arxiv.org/abs/1609.03499>
- [7] T. Inukai, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Investigation of intra-speaker spectral parameter variation and its prediction towards improvement of spectral conversion metric," in *Proc. SSW8*, Barcelona, Spain, Aug. 2013, pp. 89–94.
- [8] M. Shannon, H. Zen, and W. Byrne, "The effect of using normalized models in statistical speech synthesis," in *Proc. INTERSPEECH*, Florence, Italy, Jul. 2011, pp. 121–124.
- [9] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech and Language*, vol. 21, no. 1, pp. 153–173, Jan. 2007.
- [10] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: where do the improvements come from?" in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5505–5509.
- [11] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4455–4459.
- [12] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3872–3876.
- [13] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *Proc. ICML*, Lille, France, Jul. 2015, pp. 1718–1727.
- [14] Y. Ren, J. Li, Y. Luo, and J. Zhu, "Conditional generative moment-matching networks," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016, pp. 2928–2936.
- [15] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4460–4464.
- [16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. NIPS*, pp. 2672–2680, 2014.
- [18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2015.
- [19] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016. [Online]. Available: <https://arxiv.org/abs/1701.00160>
- [20] N. Sebastian, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," *Proc. NIPS*, pp. 271–279, 2016.
- [21] F. Cedric and J. Idier, "Algorithms for nonnegative matrix factorization with the  $\beta$ -divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, Aug. 2011.
- [22] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust TTS duration modelling using DNNs," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5130–5134.
- [23] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 816–824, 2007.
- [24] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2080–2094, Sep. 2012.
- [25] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015.
- [26] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, "ASVspoo 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, Dresden,

- Germany, Sep. 2015, pp. 2037–2041.
- [27] J. Lindberg and M. Blomberg, “Vulnerability in speaker verification - a study of technical impostor techniques,” in *Proc. EUROSPEECH*, Budapest, Hungary, Mar. 1999, pp. 1211–1214.
- [28] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, “ATR technical report,” no. TR-I-0166M, 1990.
- [29] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.
- [30] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” in *MAVEBA 2001*, Firentze, Italy, Sep. 2001, pp. 1–6.
- [31] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” in *Proc. INTER-SPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.
- [32] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, “The NAIST text-to-speech system for the Blizzard Challenge 2015,” in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.
- [33] N. Hojo, Y. Ijima, and H. Mizuno, “An investigation of DNN-based speech synthesis using speaker codes,” in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2278–2282.