深層学習による位相情報を考慮した音声合成の検討*

☆李権俊, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

テキスト音声合成 (Text-To-Speech) とは, 任意に与えられたテキストから対応する音声を合成する技術である. これまでテキスト音声合成を実現するための多くの手法が提案されており, 中でも隠れマルコフモデル (Hidden Markov Model: HMM) を用いたもの [1] が最も代表的なものとして挙げられる. 近年では深層学習による音声合成が HMM を用いた場合と比べてもより自然性の高い音声を合成できることが報告されている [2]. さらに, 平均声モデルを作成することで少量の学習データのみで合成音を作成できる適応技術 [3] や, 高い自然性を持つ合成音の作成が可能な WaveNet [4] など多くの改良がなされ続けている.

これまでの音声合成システムは特徴量として STRAIGHT [5] や WORLD [6] といった, ボコーダによって求められた音響特徴量を用いているものが大多数である. しかし, ボコーダを用いることで音声の劣化が起こってしまうという問題がある. この問題を解決するために高木らがボコーダを用いない, FFT スペクトルを用いた深層学習による音声合成を提案している [7]. 高木らの提案した手法では音声合成の際に用いる位相を, 得られた FFT 振幅スペクトルから Griffin/Lim 法 [8] によって求めている. しかし Griffin/Lim 法では完全な位相を求められることは保証されておらず, 位相情報の学習, 推定も行うことが, より高品質な音声の合成に有効であると考えられる. 本研究では複雑な位相情報を深層学習で学習, 推定するために, RPS(Related Phase Shift) という位相の表現手法を用い, 深層学習により位相情報を考慮したテキスト音声合成の検討を行う.

2 深層学習による音声合成

Fig. 1 に従来の深層学習による音声合成の概要を示す. 深層学習は入力データから対応する出力データが得られるように非線形関数を学習するモデルである. 音声合成の場合入力データとなる言語特徴量には, テキストから当該音素, 言語特徴量, アクセント型, フレーム位置などの情報をバイナリ値もしくは整数値で表現し連結したものを使用する. 出力データには入力テキストを発話した音声からボコーダによって抽出されたパラメータと, それらの動的特徴量を連結したものを使用する. 音声合成ではモデル内に動的特徴量を考慮する必要があるので, 出力データに各パラメータの動的特徴量を連結するか, もしくは LSTM などの RNN(Recurrent Neural Network) を用いて深層学習の構造内で時系列を考慮した学習, 推定を行う. 音声合成時には, テキストから言語特徴量を抽出し, モデルに入力して得られた音声特徴量からボコーダによって音声を合成する.

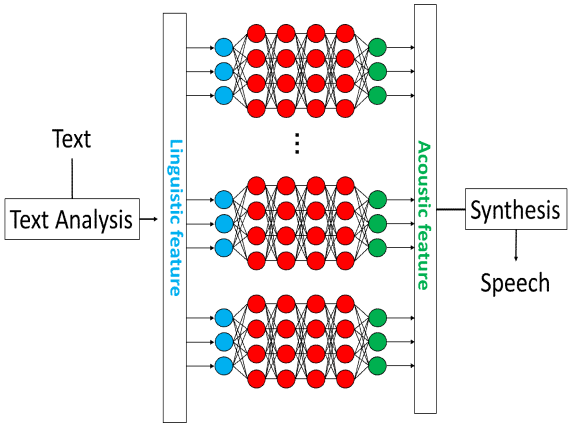


Fig. 1: A flow of speech synthesis using deep neural networks

3 提案手法

音声における人間の知覚に対して, 位相情報が及ぼす影響に関しては, 長年に渡って議論されてきた. 人間の聴覚では位相の変化を知覚できない, あるいは位相の変化が及ぼす影響はごく僅かで無視することができる, とする研究結果も存在したが [9], [10], 現在では位相が及ぼす影響は少なくないことが分かっており, 雑音除去などの分野では位相情報の重要性に関して研究がなされている [11].

STRAIGHT や WORLD などのボコーダでは, 最小位相応答として音声を合成することで, 位相

*Speech synthesis system using phase information with deep learning, by Konjun I, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

情報を構築することが一般的であり, この方法はより自然性の高い音声を合成することに貢献している.

本研究ではFFTスペクトルを利用した音声合成の品質を向上させることを目的に, 深層学習を用いて, FFT振幅スペクトルと合わせて位相情報の学習, 推定も行うモデルを提案する. しかし, 位相は複雑な構造をしており, それを学習, 合成することは容易なことではない. そこで本研究では位相を深層学習で効率的に学習するために, [12] らが提案している RPS(Relative Phase Shift) という位相の表現を特徴量として用いることにした.

3.1 RPS

音声データをフーリエ解析すると以下のような式が得られる.

$$y(t) = \sum_{k=1}^{N} \mathbf{A}_k e^{j\phi_k(t)} \tag{1}$$

$$\phi_k(t) = 2\pi k f_0(t) + \theta_k \tag{2}$$

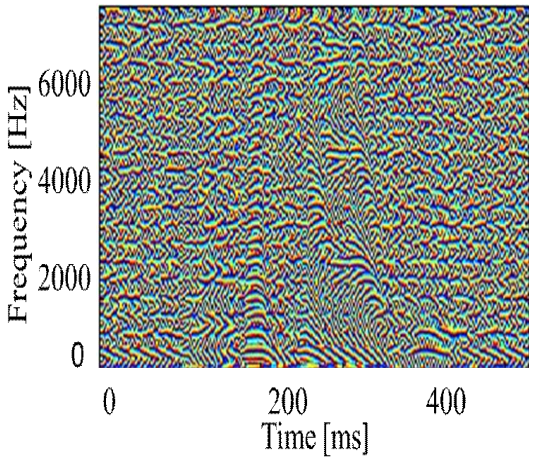
N, $\mathbf{A}_k$, ϕ_k, f_0, θ_k はそれぞれ周波数のバンド数, 振幅, 位相, 基本周波数, 初期位相を表す. 式 (2) から見て取れるように, 位相は時間と周波数に比例する項と第二項の初期位相との加算で表現できる. RPSは初期位相にのみ注目することによって, 複雑な位相の構造を捉えやすくするための位相の表現方法である. RPS $\psi_k(t)$ は瞬時位相から以下の式で求められる.

$$\psi_k(t) = \phi_k(t) - k\phi_1(t) \tag{3}$$

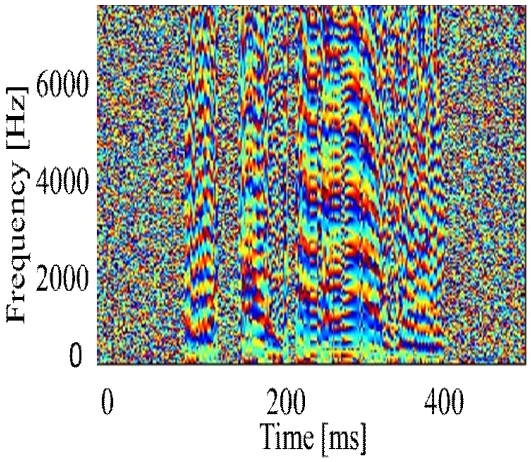
「いきおい」と発話したときの音声から求められる位相と RPS をそれぞれ Fig. 2a, Fig. 2b に示す. Fig. 2a の位相ではあいまいな発話部分での構造が Fig. 2b の RPS ではより捉えやすくなっていることが分かる.

3.2 RPSを使用した音声合成

深層学習で FFT 振幅スペクトルのみを学習, 推定し,Griffin/Lim 法で位相を求めた場合, 正しい位相を得られる保証がない. そこで本研究では RPS を特徴量に追加し, 推定された RPS から計算できる位相を, Griffin/Lim 法における位相推定の初期値として用いることで, より正確な位相を推定し, 合成音声の品質を向上させることに取り組んだ. 実験では次元数削減のために, RPS にメルフィルタバンクを適用した後離散コサイン変換し, 低次元にした特徴量を使用した.



(a) Phaesegram



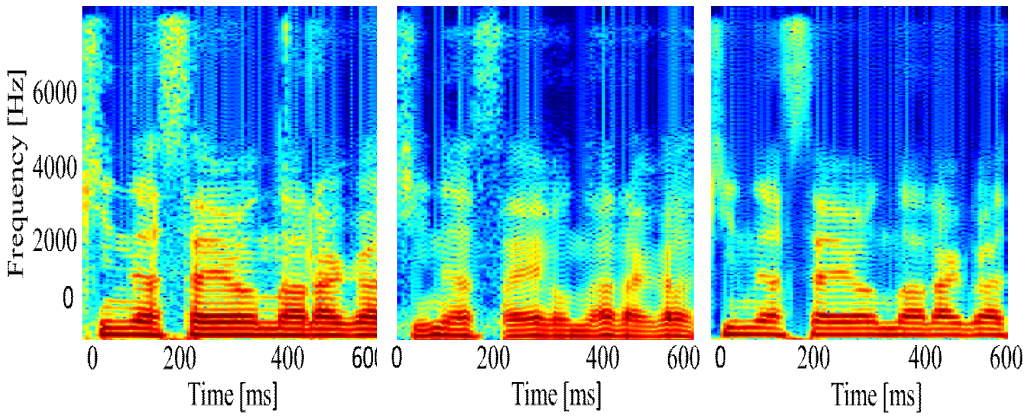
(b) RPS phasegram

Fig. 2: Phasegram and RPS phasegram for an utterance

4 評価実験

4.1 実験条件

ATR 音素バランス 503 文を用いて実験を行った. サンプリング周波数は 16kHz, フレームシフトは 5ms とした. 実験では, 言語特徴量にはコンテキストラベルに対して HTS 形式の Questin を適用して抽出したものを使用し, 音響特徴量には WORLD から抽出したパラメータを用いたもの, FFT スペクトルのみを用いたもの, FFT スペクトルに加えて RPS を用いたもの, 以上の 3 種類のモデルを構築した. 言語特徴量と音響特徴量はいずれも最小値 0, 最大値 1 に次元ごとに正規化を行った. WORLD から抽出したパラメータを使うモデル (WORLD) では, スペクトル包絡の 40 次元メルケプストラム係数, 対数 F0, 25 帯域 AP を音響特徴量とした. FFT 振幅スペクトルの



(a) Target (b) WORLD (c) FFT+RPS

Fig. 3: Comparison of spectrogram

みを使用するモデル (FFT) では, FFT 長 1024 で計算された振幅スペクトルを, FFT 振幅スペクトルに加えて RPS も使用するモデル (FFT+RPS) では, FFT 長 1024 で計算された振幅スペクトルに, RPS20 次元メルケプストラムと基本周波数, 基本周波数での位相を連結し音響特徴量とした. いずれのモデルでも LSTM を用いて学習を行った. 提案手法である FFT+RPS のモデルの略図を Fig. 4 に示す.

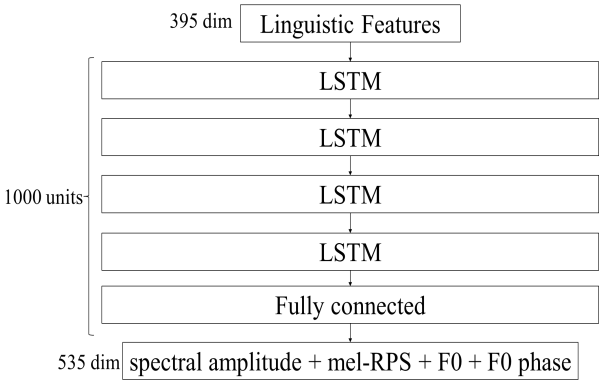


Fig. 4: FFT+RPS model

また合成した音声にはメルケプストラムに基づくポストフィルタ [13] を適用し, フォルマントの強調を行った.

提案手法の有効性を確かめるため, 音質と話者性の 2 つの観点において主観評価実験を行った. オープンなテキスト 10 文について, 3 種類の方法 (WORLD, FFT, FFT+RPS) で音声を合成し, ヘッドホンを用いて聴取実験を行った. 音質に関しては MOS (Mean Opinion Score) による 5 段階評価 (1:非常に悪い 2:悪い 3:普通 4:良い 5:非常に良い), 話者性に関しては 5 段階評価 (1:非常に遠い 2:遠い 3:普通 4:近い 5:非常に近い) である.

4.2 実験結果・考察

Fig. 3 に目標音声と手法 WORLD, FFT+RPS によって合成された音声のスペクトログラムの一部を示す. Fig. 3 より FFT+RPS がより目標音声に近いスペクトログラムを推定できていることが分かる.

Fig. 5, Fig. 6 に主観評価実験の結果を示す. 図中のエラーバーは 95%信頼区間を示している. Fig. 5, Fig. 6 より手法 FFT+RPS が最も高い音質, 話者性を持つことが分かる. また, FFT 振幅スペクトルのみを用いた手法 FFT が, 手法 WORLD より低い話者性を示しているのに対して, FFT 振幅スペクトルに加えて RPS を用いた手法 FFT+RPS は手法 WORLD を上回っていることから, 位相情報の学習, 推定は話者性の向上により有効に働いたといえる.

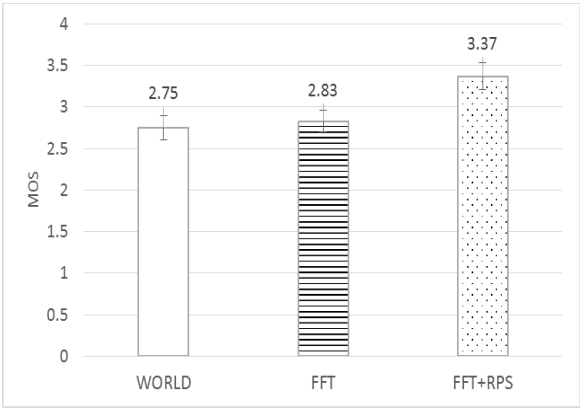


Fig. 5: Sound quality

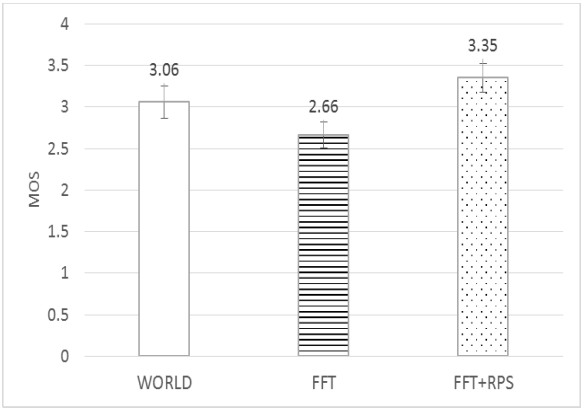


Fig. 6: Speaker similarity

5 おわりに

本研究では, 位相の表現方法の一種である RPS を用い, 深層学習を用いた位相情報を考慮したテキスト音声合成について提案を行った. 音質と話者性について主観評価実験を行い検証した結果, 提案手法が音質と話者性の両方で従来手法を上回る結果となり, 提案手法の有効性が示された. 今後は位相情報を考慮した声質変換などへの応用を検討していく.

謝辞 本研究の一部は, JST さきがけ JPMJPR15D2, JSPS 科研費 JP17H01995 の支援を受けたものである.

参考文献

[1] K. Tokuda *et al.*, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, 2000, pp. 1315–1318.

[2] H. Ze *et al.*, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*, 2013, pp. 7962–7966.

[3] J. S.Takaki, S.Kim, "Speaker adaptation of various components in deep neural network based speech synthesis," Speech Synthesis Workshop, 2016, pp. 167–173.

[4] A. v. d. Oord *et al.*, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[5] H. Kawahara *et al.*, "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, pp. 187–207, 1999.

[6] M. Morise *et al.*, "World: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877–1884, 2016.

[7] 高木信二 *et al.*, "DNN に基づくテキスト音声合成のための FFT スペクトルを用いた位相復元に基づく音声波形生成," 研究報告音声言語情報処理 (SLP), vol. 2016, no. 21, pp. 1–6, 2016.

[8] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984.

[9] H. Von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, Longmans, Green, 1912.

[10] S. P. Lipshitz *et al.*, "On the audibility of midrange phase distortion in audio systems," Journal of the Audio Engineering Society, vol. 30, no. 9, pp. 580–595, 1982.

[11] K.K.Paliwai *et al.*, "The importance of phase in speech enhancement," Speech communication, vol. 53, no. 4, pp. 465–494, 2011.

[12] I. Saratxaga *et al.*, "Simple representation of signal phase for harmonic speech models," Electronics letters, vol. 45, no. 7, pp. 381–383, 2009.

[13] K. Koishida *et al.*, "CELP coding based on mel-cepstral analysis," in *ICASSP*, 1995, pp. 33–36.