差分スペクトル法に基づく広帯域声質変換のためのサブバンドリフタ学習*

☆佐伯 高明, 齋藤 佑樹, 高道 慎之介, 猿渡 洋 (東大院・情報理工)

1 はじめに

声質変換の実応用のためには，品質だけではなく，リアルタイム性や省計算リソース性が求められる．かねてより，CPU によるリアルタイム声質変換手法 [1, 2] が提案されているが，変換可能な音声の帯域幅が狭いことや，ボコーダによる品質劣化が生じるなどの課題がある．本研究では，CPU を用いたリアルタイム・広帯域声質変換を実現するため，より高品質かつ計算効率の比較的高い声質変換手法である差分スペクトル法 [3] を用いる．差分スペクトル法は，変換元話者と変換先話者のスペクトル包絡の差分を与えるようなフィルタを推定し，それを変換元話者の音声波形に対して直接適用することによって変換を行う手法である．これまでに我々は，狭帯域声質変換のための差分スペクトル法に対し，フィルタ打ち切りを考慮したリフタ学習法を提案した [4]．この方法では，実ケプストラムを変換するモデルのみならず，実ケプストラムから位相を復元するリフタをデータドリブンに学習する．このリフタ学習法により，16 kHz サンプリング音声の変換において，品質を劣化させずにフィルタのタップ長を 1/8 にまで短縮し，変換に要する計算量を大幅に低減できることを示した．

差分スペクトル法をそのまま広帯域声質変換に適用した場合，高域のランダムな変動によりモデル化性能が低下するという問題が生じる．また，音声の周波数帯域が拡大するほど，フィルタリングに要する計算量が増大することも，無視できない問題である．そこで本稿では，サブバンド信号処理 [5] とリフタ学習を組み合わせた差分スペクトル法を提案する．具体的には，サブバンドマルチレート信号処理により広帯域音声を複数の帯域に分割した後，最低域に対して，フィルタ打ち切りを考慮したリフタ学習を適用する．高周波域のモデル化を避けることで，ランダム変動に起因する変換音声の品質の低下を回避し，さらに，狭帯域信号に対してのみリフタを学習することで，広帯域化による計算量の増加を緩和する．実験的評価では，48 kHz サンプリングの音声の変換に対して提案法を適用し，フィルタリングの計算量を削減しながら品質を大幅に改善できることを示す．

2 従来法: 最小位相フィルタを用いた差分スペクトル法に基づく広帯域声質変換

差分スペクトル法に基づく声質変換では，最小位相フィルタを用いることによって，MLSA (Mel-Log Spectrum Approximation) フィルタを用いた場合よりも高品質な変換音声が得られることが知られている [6]．この節では，最小位相フィルタを用いた差分スペクトル法を広帯域声質変換に適用する場合の，学習時・変換時の処理について述べる．

2.1 学習時

まず，変換元話者の音声波形を短時間フーリエ変換することにより，複素スペクトル系列 $\boldsymbol{F}^{(\mathrm{X})} = [\boldsymbol{F}_1^{(\mathrm{X})\top}, ..., \boldsymbol{F}_t^{(\mathrm{X})\top}, ..., \boldsymbol{F}_T^{(\mathrm{X})\top}]^\top$ を得る．ただし，t はフレームインデックスで，T は総フレーム数である．以降，フレーム t のみに着目して議論する．$\boldsymbol{F}_t^{(\mathrm{X})}$ から低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{X})}$ を抽出し [7]，これを DNN (Deep Neural Network) の入力として，差分フィルタの低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{D})}$ を推定する．このとき，変換音声の低次実ケプストラム $\hat{\boldsymbol{C}}_t^{(\mathrm{Y})}$ は $\hat{\boldsymbol{C}}_t^{(\mathrm{Y})} = \boldsymbol{C}_t^{(\mathrm{X})} + \boldsymbol{C}_t^{(\mathrm{D})}$ と書け，フレーム t に関する損失関数 L_t は式 (1) のように求まる．

$$L_t = (\boldsymbol{C}_t^{(\mathrm{Y})} - \hat{\boldsymbol{C}}_t^{(\mathrm{Y})})^\top (\boldsymbol{C}_t^{(\mathrm{Y})} - \hat{\boldsymbol{C}}_t^{(\mathrm{Y})}) \tag{1}$$

ただし，$\boldsymbol{C}_t^{(\mathrm{Y})}$ は変換先話者の自然音声の低次実ケプストラムである．このとき，式 (2) の損失関数 L を最小化するように DNN のパラメータを学習する．

$$L = \frac{1}{T}\sum_{t=1}^{T} L_t \tag{2}$$

2.2 変換時

変換時は，まず学習済みの DNN によって差分フィルタの低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{D})}$ を推定する．さらに，$\boldsymbol{C}_t^{(\mathrm{D})}$ の高次の項を 0 埋めし，式 (3) によって表される最小位相化のためのリフタ係数 $\boldsymbol{u}_{\min}$ [8] を掛けてヒルベルト変換を行うことにより，差分フィルタの複素スペクトル $\boldsymbol{F}_t^{(\mathrm{D})}$ を得る．

$$\boldsymbol{u}_{\min}(n) = \begin{cases} 1 & (n = 0, n = N/2) \\ 2 & (0 < n < N/2)\,, \\ 0 & (n > N/2) \end{cases} \tag{3}$$

ただし，N は周波数ビン数である．この $\boldsymbol{F}_t^{(\mathrm{D})}$ をフーリエ変換することにより，時間領域での差分フィルタを得る．その後，この差分フィルタを変換元話者の音声波形に対して畳み込むことにより変換を行う．ここで，変換時に単純なフィルタ打ち切りを行う場合，計算量が削減できる代わりに音質が劣化する．

3 提案法: サブバンドリフタ学習による広帯域声質変換

差分スペクトル法を広帯域音声の変換にそのまま適用した場合，高域のランダム性によりモデル化性能が低下する．このため，帯域拡張したにも関わらず変換音声の品質は大きく向上しない．そこで，サブバンドマルチレート処理によって変換元話者の音声を帯域分割し，低域にのみ差分フィルタを適用する．ここではまず，提案法の要素である，フィルタ打ち切りを考慮したリフタ学習 (3.1 節) とサブバンドマル

*Sub-band lifter-training method for full-band voice conversion using spectral differentials by Takaaki Saeki, Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari (The University of Tokyo)

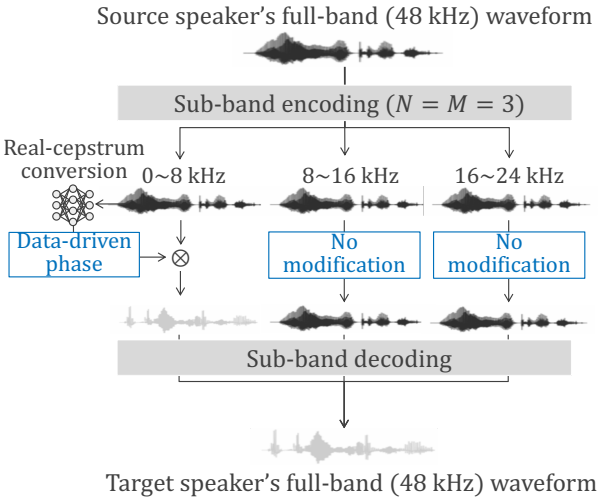

Fig. 1　サブバンドリフタ学習の処理フロー．サブバンドマルチレート処理の分析によりサブバンド信号を取り出し，0-8 kHz のみリフタ学習法を適用する．その後，再度合成することによって最終的なフルバンド変換音声を得る．ここでは，音声のサンプリング周波数を 48 kHz，帯域分割数 N を 3，間引率 M を 3 としている．

チレート処理 (3.2 節) を概説する．その後，それらを組み合わせたサブバンドリフタ学習法 (3.3 節) を提案する．

3.1　フィルタ打ち切りを考慮したリフタ学習

フィルタ打ち切りを考慮したリフタ学習法では，フィルタの打ち切りを微分可能な形で学習過程に組み込み，DNN のパラメータのみならずヒルベルト変換のためのリフタ係数をも更新する．

3.1.1　学習時

学習時の処理を Fig. 2 に示す．まず，2.1 節と同様に，DNN によって差分フィルタの低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{D})}$ を推定する．$\boldsymbol{C}_t^{(\mathrm{D})}$ の高次の項を 0 埋めし，学習によって更新するリフタ係数 $\boldsymbol{u}$ を掛ける．さらに逆フーリエ変換して exp を取ることにより，差分フィルタの複素スペクトル系列 $\boldsymbol{F}_t^{(\mathrm{D})}$ を得る．これを逆フーリエ変換して時間領域の差分フィルタ $\boldsymbol{f}_t^{(\mathrm{D})}$ とし，式 (4) ように窓関数 $\boldsymbol{w}$ を適用することによって打ち切りを行う．

$$\boldsymbol{f}_t^{(l)} = \boldsymbol{f}_t^{(\mathrm{D})} \cdot \boldsymbol{w}, \tag{4}$$

$$\boldsymbol{w} = \left[\overset{0\text{th}}{1}, \cdots, \overset{(l-1)\text{th}}{1}, \overset{l\text{th}}{0}, \cdots, \overset{(N-1)\text{th}}{0}\right]^\top \tag{5}$$

時間領域で打ち切られたフィルタ $\boldsymbol{f}_t^{(l)}$ を再度フーリエ変換し，タップ長 l の差分フィルタの複素スペクトル系列 $\boldsymbol{F}_t^{(l)}$ を得る．変換音声の複素スペクトル系列 $\hat{\boldsymbol{F}}_t^{(\mathrm{Y})}$ は，$\boldsymbol{F}^{(\mathrm{X})}$ と $\boldsymbol{F}_t^{(l)}$ の要素積を取ることによって得られる．さらに，$\hat{\boldsymbol{F}}_t^{(\mathrm{Y})}$ から変換音声の実ケプストラム $\hat{\boldsymbol{C}}_t^{(\mathrm{Y})}$ を抽出する．学習時に用いる損失関数は式 (1) と同じだが，DNN のパラメータだけでなく，リフタ係数をも学習する．学習の全過程において微分可能であり，誤差逆伝播法によりパラメータ更新を行うことができる [9]．

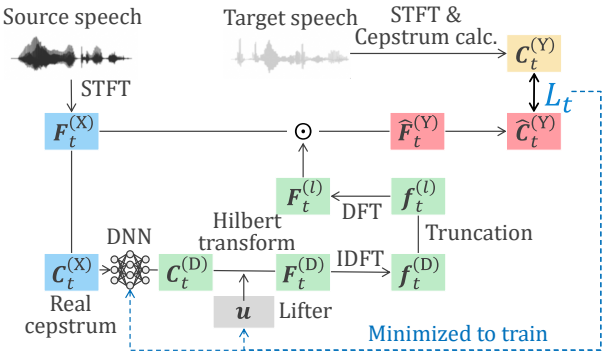

Fig. 2　リフタ学習法での学習過程．

ここで，リフタ係数はフィルタの位相を決定するためのパラメータであり，リフタ係数を学習によって更新することにより，打ち切りによるフィルタ形状の変化の影響を補償することが期待される．

3.1.2　変換時

変換時には，学習済みの DNN とリフタ係数で $\boldsymbol{F}_t^{(\mathrm{D})}$ を推定する．これをフーリエ変換して時間領域のフィルタ $\boldsymbol{f}_t^{(\mathrm{D})}$ とし，タップ長 l で打ち切ることによって $\boldsymbol{f}_t^{(l)}$ を得る．これを変換元話者の自然音声に直接適用することによって変換音声を得る．

3.2　サブバンドマルチレート処理

サブバンドマルチレート処理を用いた分析・合成の処理について概説する．Fig. 3 に一連の処理を示す．

3.2.1　分析時

変換元話者の自然音声を N 個のサブバンド信号に帯域分割し，$W_N^{-t(n-1/2)}$ で変調してベースバンドに周波数シフトする．

$$x(t) = x(t) W_N^{-t(n-1/2)} \tag{6}$$

ただし，$n = 1, 2, \cdots, N$ であり，$W_N = \exp(j2\pi/2N)$ とする．次に，全ての帯域で共通なローパスフィルタ $f(t)$ を適用することにより $[-\pi/2N, \pi/2N]$ に帯域制限する．

$$x_{n,pp}(t) = f(t) * x_n(t) \tag{7}$$

ここで，$*$ は畳み込みの演算子である．$x_{n,pp}(t)$ は複素数値として得られるため，実数値として扱うために Single Sideband (SSB) 変調法を導入する．実数値信号 $x_{n,\mathrm{SSB}}(t)$ は以下のようにして得られる．

$$x_{n,\mathrm{SSB}}(t) = x_{n,pp}(t) W_N^{t/2} + x_{n,pp}^*(t) W_N^{-t/2} \tag{8}$$

ただし，$\cdot^*$ は複素共役を表す．ここで得られた $x_{n,\mathrm{SSB}}(t)$ を間引率 M で間引くことにより，n 番目のサブバンド信号 $x_n(k)$ が得られる．

$$x_n(k) = x_{n,\mathrm{SSB}}(kM) \tag{9}$$

3.2.2　合成時

変換時には，$x_n(k)$ のうち低い周波数領域に対応する成分のみにフィルタを適用し，高い周波数領域の成分には何も処理を行わない．このようにして変換

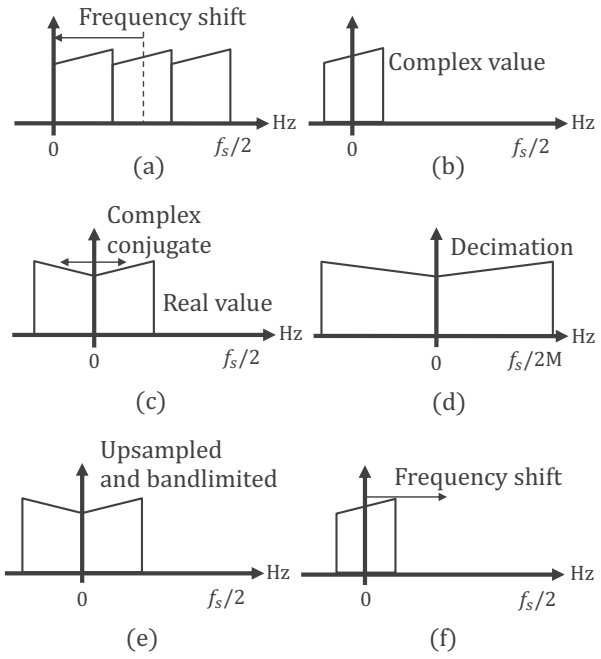

Fig. 3　サブバンドマルチレート処理．(a) から (d) は分析時の処理を，(d) から (f) は合成時の処理を表す．

されたサブバンド信号を $\hat{x}_t(t)$ とする．これらを用いてフルバンド信号を合成するために，$\hat{x}_n(t)$ を M でアップサンプリングする．

$$\hat{x}_{n,\mathrm{SSB}}(t) = \begin{cases} \hat{x}_n(t/M) & (t = 0, M, 2M, \cdots) \\ 0 & (\text{otherwise}) \end{cases} \tag{10}$$

ここでエイリアシングを避けるために，$\hat{x}_{n,\mathrm{SSB}}(t)$ をベースバンドに周波数シフトし，ローパスフィルタ $g(t)$ によって帯域制限する．

$$\hat{x}_{n,pp}(t) = g(t) * \left(\hat{x}_{n,\mathrm{SSB}}(t) W_N^{-t/2}\right) \tag{11}$$

フルバンド変換音声 $\hat{x}(t)$ は，最終的に以下のように合成することができる．

$$\hat{x}(t) = \sum_{n=1}^{N} \{\hat{x}_{n,pp}(t) W_N^{t(n-1/2)} + \hat{x}^*_{n,pp}(t) W_N^{-t(n-1/2)}\} \tag{12}$$

3.3　フィルタ打ち切りを考慮したサブバンドリフタ学習

Fig. 1 に，サブバンドリフタ学習での処理フローを示す．サブバンドマルチレート処理によって分析した信号に対してリフタ学習による学習・変換を行う．

3.3.1　学習時

まず，サブバンドマルチレート処理の分析により，複数のサブバンド信号を取り出す．具体例として，サンプリング周波数を 48 kHz，$N = M = 3$ とすると，0–8 kHz，8–16 kHz，16–24 kHz の 3 つのサブバンド信号が取り出される．このうち，低域の信号のみを用いて，3.1.1 節のリフタ学習法を適用する．

3.3.2　変換時

変換時も，学習時と同様に複数の帯域それぞれからサブバンド信号を取り出す．このうち，学習された差分フィルタを用いて低域のサブバンド信号のみを変換し，それ以外の高域のサブバンド信号については変換を行わない．その後，3.1.2 節に示す合成処理によって最終的なフルバンド変換音声を得る．

4　実験的評価

4.1　実験条件

男性話者から男性話者 (m2m)，女性話者から女性話者 (f2f) の 2 種類の変換について実験を行った．男性の変換元話者・変換先話者にはいずれも JVS コーパス [10] の男性話者を用いた．女性の変換元話者には JSUT コーパス [11] の女性話者，変換先話者には声優統計コーパス [12] の女性話者を用いた．それぞれの話者データについて 100 発話 (約 12 分) を使用し，80 文を training データ，10 文を validation データ，10 文を test データとした．

評価には 48 kHz サンプリング音声を用いた．48 kHz サンプリング音声に短時間フーリエ変換を行う際は，窓長を 25 ms，フレームシフトを 5 ms，FFT 長を 2048 点，低次ケプストラムの次元を 120 とした．0–8 kHz の音声に短時間フーリエ変換を行う際は，窓長とフレームシフトには 48 kHz の場合と同じものを用い，FFT 長を 512 点，低次ケプストラムの次元を 30 とした．前処理として，training データと validation データの無音区間を除去し，変換元話者の音声と変換先話者の音声のデータ長を dynamic time warping により揃えた．提案法でサブバンドマルチレート処理を行う際は，$N = M = 3$ とし，0–8 kHz に対してフィルタ打ち切りを考慮したリフタ学習を適用した．

実験に用いた DNN アーキテクチャは，隠れ層 2 層の Feedforward Neural Network とした．従来法で 48 kHz サンプリングの音声を変換する場合は，隠れユニット数はそれぞれ 840，300 とした．提案法で 0–8 kHz の帯域の音声を変換する場合は，隠れユニット数はそれぞれ 280，100 とした．隠れ層の活性化関数として，sigmoid 関数，tanh 関数からなる Gated Linear Unit [13] を持ち，各々の活性化関数に通す前に Batch Normalization [14] を行った．また，最適化手法には Adam [15] を用いた．学習時に変換元話者と変換先話者のケプストラムを平均 0・分散 1 に正規化した．バッチサイズとエポック数はそれぞれ 1000，100 とし，リフタ学習法の DNN パラメータは，最小位相フィルタに基づく差分スペクトル法の学習後の値で初期化し，その際のリフタ係数は最小位相化のためのリフタ係数の値で初期化した．48 kHz サンプリングの音声に従来法を適用する場合の学習率は 0.0001 とした．また，0-8 kHz の音声にリフタ学習法を適用する際，最小位相フィルタに基づく差分スペクトル法とリフタ学習法の学習率はそれぞれ 0.0001，0.000005 とした．

Table 1　リフタ学習法とサブバンド処理を組み合わせて用いた場合と，従来法を用いた場合のプリファレンススコア **(48 kHz)**

(a) **話者類似性**

Proposed	Score	p-value	Conventional
l : 32 (m2m)	0.537 vs. 0.463	7.3×10^{-2}	l : 2048 (m2m)
l : 32 (f2f)	0.516 vs. 0.484	2.5×10^{-1}	l : 2048 (f2f)
l : 48 (m2m)	0.493 vs. 0.507	7.4×10^{-1}	l : 2048 (m2m)
l : 48 (f2f)	0.475 vs. 0.525	8.3×10^{-2}	l : 2048 (f2f)
l : 64 (m2m)	0.520 vs. 0.480	3.3×10^{-1}	l : 2048 (m2m)
l : 64 (f2f)	0.532 vs. 0.468	1.1×10^{-1}	l : 2048 (f2f)

(b) **音質**

Proposed	Score	p-value	Conventional
l : 32 (m2m)	**0.840** vs. 0.160	$< 10^{-10}$	l : 2048 (m2m)
l : 32 (f2f)	**0.810** vs. 0.190	$< 10^{-10}$	l : 2048 (f2f)
l : 48 (m2m)	**0.828** vs. 0.172	$< 10^{-10}$	l : 2048 (m2m)
l : 48 (f2f)	**0.593** vs. 0.407	4.2×10^{-6}	l : 2048 (f2f)
l : 64 (m2m)	**0.830** vs. 0.170	$< 10^{-10}$	l : 2048 (m2m)
l : 64 (f2f)	**0.700** vs. 0.300	$< 10^{-10}$	l : 2048 (f2f)

4.2　主観評価実験

クラウドソーシングを用いて音質に関する AB テストおよび話者類似性に関する XAB テストを行った．各条件につき，30 人の聴取者が 10 文の音声サンプルを評価した．XAB テストの参照音声 X には変換先話者の自然音声を用いた．

比較手法は，最小位相フィルタを用いる従来法と，サブバンドリフタ学習を用いる提案法とした．従来法のタップ長は $l = 2048$，提案法のタップ長は $l = 32, 48, 64$ のいずれかとした．Table 1 に評価結果を示す．話者類似性については，従来法と提案法の間には有意な差は確認できなかった．音質については，従来法よりも著しく高い評価結果を示した．以上より，提案法を用いることによって，計算量を削減でき，さらに品質も大幅に向上することが確認できる．

5　おわりに

本稿では，差分スペクトル法に基づく広帯域声質変換に対し，サブバンドリフタ学習を提案した．48 kHz サンプリングの音声を用いて実験的評価により，提案法は計算量を削減しながら変換音声の品質を有意に改善できることを示した．

謝辞:　本研究開発は総務省 SCOPE(受付番号 182103104) の委託を受けたものです．

参考文献

[1] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012, pp. 94–97.

[2] R. Arakawa, S. Takamichi, and H. Saruwatari, "Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," in *Proc. SSW10*, Vienna, Austria, Sep. 2019, pp. 93–98.

[3] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, May. 2018.

[4] 佐伯高明, 齋藤佑樹, 高道慎之介, and 猿渡洋, "差分スペクトル法に基づく DNN 声質変換の計算量削減に向けたフィルタ推定," in **音講論 （秋）**, no. 2-4-1, 滋賀, Sep. 2019.

[5] R. Crochiere and L. Rabiner, *Multirate digital signal processing*. Englewood Cliffs, N.J. : Prentice-Hall, 1983.

[6] H. Suda, G. Kotani, S. Takamichi, and D. Saito, "A revisit to feature handling for high-quality voice conversion," in *Proc. APSIPA ASC*, Hawaii, U.S.A., Nov. 2018, pp. 816–822.

[7] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, San Francisco, U.S.A., Mar. 1992, pp. 137–140.

[8] S.-C. Pei and H.-S. Lin, "Minimum-phase FIR filter design using real cepstrum," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 10, pp. 1113–1117, 2006.

[9] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[10] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free japanese multi-speaker voice corpus," *arXiv*, vol. abs/1908.06248, 2019.

[11] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," vol. abs/1711.00354, 2017.

[12] y_benjo and MagnesiumRibbon, "Voice-actress corpus," http://voice-statistics.github.io/.

[13] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv*, vol. abs/1612.08083, 2016.

[14] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[15] D. Kingma and B. Jimmy, "Adam: a method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.