Vol.2017-MUS-116 No.9
2017/8/25

楽譜と演奏履歴を用いた深層自己回帰過程に基づく演奏タイミング予測

前澤 陽[1,a)]

概要：本稿では，既知の楽曲を演奏している人間の演奏データにおける，発音タイミングの予測手法を提案する．従来の発音タイミング予測では，楽曲毎に個別の予測モデルを学習していたため，未知の楽曲に対する予測が行えなかった．そこで，演奏者は楽譜から得られる文脈と演奏タイミング履歴の両者に依存しながら次の発音タイミングを決定しているという仮説に基づき，これらを用いた発音タイミング予測手法を提案する．楽譜や演奏履歴情報からタイミング予測に有用な特徴量設計を行うのは困難であるため，これらを入力とし線形予測係数を出力する深層ニューラルネットワーク (DNN) を用いて，タイミングを自己回帰過程としてモデル化・予測する．評価実験の結果，テンポを移動平均としてモデル化する場合と比べ，予測誤差が約 23%削減されることが示された．

Expressive Timing Prediction Based on Deep Autoregressive Model Using Score and Performance Data

Akira Maezawa[1,a)]

1. はじめに

複数の人間が合奏するとき，演奏者は互いの演奏を聞きあうことで，互いのタイミングを予測しながら適切に応答する．このようなタイミングの読み合いは，自動伴奏などのインタラクティブな音楽システムにおいて計算機が実現すべき重要な要素である．そこで，既知の楽曲に対して未知の演奏データを逐次与えたときに，現時点以降の発音タイミングをオンラインで予測する手法が必要になる．

従来の演奏タイミング予測・生成手法では，初めて演奏する楽曲に対するタイミングの予測が困難であった．例えば，自動伴奏システムでは，リハーサルを通じて楽曲のタイミング予測モデルを学習する [1]．しかし，学習は各楽曲に対して独立に行うため，ある楽曲で学習された予測モデルを，未知の楽曲に対して適用させることができないという問題があった．一方，演奏表情付け手法では，楽譜情報を与えると，初めて演奏される曲に対しても，人間らしい発音タイミングを生成できる [3]．このようなことが実現できるのは，共起されやすい楽譜と演奏の特徴が存在するからである．このように，楽譜データから想起される演奏タイミングの傾向を活用することで，演奏されたことのない楽曲からでも，リーズナブルなタイミング情報が生成できる．しかし，演奏表情付けはあくまでタイミングの生成を行うものである．そのため，部分的に与えられている演奏データから，与えられていないデータのタイミングを予測することは問題の対象外である．

そこで，本稿では，初めて演奏される楽曲に対しても頑健に動作する，演奏タイミングのオンライン予測手法を提案する．本手法では，楽譜データと，その楽曲を演奏している演奏データの発音時刻を逐次入力したときに，その次の発音時刻を予測する．初めて演奏されるデータに対してタイミングを予測するため，演奏表情付けに倣い，楽譜データから想起される演奏タイミングの傾向と，演奏者が実際に演奏しているタイミングの履歴情報を活用することを考える．

[1] ヤマハ株式会社
Yamaha Corporation, Iwata, Shizuoka 438–0942, Japan
[a)] akira.maezawa@music.yamaha.com

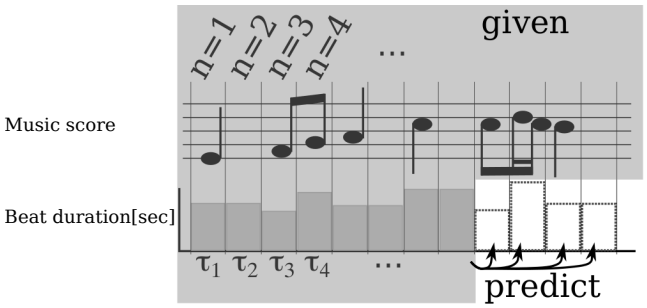

図 1 本手法の問題設定．八分音符の長さを各八分音符区切りの位置に対して予測する．

2. 関連研究

自動伴奏システムでは，演奏者の発音タイミングを適切に予測することが求められる．そのため，リハーサルを通じてテンポ軌跡を予測することや [1, 4] タイミング予測モデルのパラメータを学習すること [5] が重要である．学習データを用意するためには煩雑な通し練習が必要になる．したがって，弾いたことのない曲に対して，少ない回数の通し練習で，高精度の予測を行えることが重要になる．既知の楽曲に対して，楽譜情報を活用することで，少ない回数のリハーサルでタイミング予測の精緻化が可能になることは報告されているが [6]，未演奏の楽曲に対するタイミング予測については議論されていなかった．

未知の楽曲に対して人間らしい発音タイミングを生成するタスクとして演奏表情付けがある [3]．演奏表情付けでは，演奏に有用な特徴量から演奏に必要なパラメータを推定する [7]．特徴量設計は困難であるため，DNN を用いてデータドリブンに特徴抽出機を学習することの有用性が確認されている [8]．演奏表情付けでは人間らしい演奏データを生成できるものの，部分的に与えられた演奏データからその先の演奏をオンラインで予測することは対象外である．本手法では，演奏表情付けのアイディアを演奏のオンライン予測に活用することを考える．

3. 提案手法

本稿では，図 1 に示すように，八分音符を一拍と定義し，楽曲開始から n 拍が経過したときの拍長を τ_n とする．本手法はタイミング予測を，楽譜データと n 拍目までの拍長履歴 $\{\tau_{n'}\}_{n'=1}^{n}$ が与えられたときに τ_{n+1} から任意の P に対して τ_{n+P} の値を予測するマルチステップ予測問題として定式化する．なお，本稿では単一の演奏者のタイミングを予測することを考え，合奏における奏者間の相互作用は考慮しない．

拍長を予測するためには二つの情報を用いる．まず，演奏者が現時点まで演奏した実際の拍長の履歴 $\{\tau_{n'}\}_{n'<n}$（以下「演奏情報」と呼ぶ）を用いる．なぜならば，演奏における発音時刻は直前の演奏情報に強く依存するためである．次に，予測したい楽譜上の周辺位置から得られる楽譜情報を用いる．なぜならば，演奏における発音情報は演奏情報だけでなく，楽譜に表記されている音型や和声進行等といった情報にも強く依存すると考えられるためである．このように楽譜のみから得られる情報を「楽譜情報」と呼ぶ．

これらを踏まえ，拍長 τ_{n+p} が演奏情報と楽譜情報のみに依存すると仮定し，τ_{n+p} を次のような I 次の自己回帰過程としてモデル化する:

$$\tau_n|\theta,\phi_n,\{\tau_{n'}\}_{n'=1}^{n-1} \sim \mathcal{L}\left(\sum_{i=1}^{I} a_{p,i}(\theta,\phi_n,\{\tau_{n'}\}_{n'<n})\tau_{n-i},\lambda\right). \quad (1)$$

ここで $\mathcal{L}(\mu,\lambda)$ は位置パラメータ μ，尺度パラメータ λ に従うラプラス分布である．ϕ_n は拍 n の周辺から得られる楽曲に関する情報であり，$a_{p,i}(\theta,\phi_n,\{\tau_n\})$ は演奏情報 $\{\tau_{n'}\}_{n'<n}$，楽譜情報 ϕ_n，そしてパラメータ θ により定義される関数である．この関数を以後「予測係数関数」と呼ぶ．

このように自己回帰過程としてモデル化することにより，拍長に対する微小なスケールの違いに対して，一貫した挙動が確保できる，そのため，学習データに現れないテンポ値に対しても，リーズナブルな予測が実現できることが期待される．また，連続値を離散化するアプローチ [9] と違い離散化の基準を決める必要がない．

予測係数関数は拍長のダイナミクスを表現する．ダイナミクスの傾向は楽譜情報と演奏情報に応じて動的に変化するため，変動要因が多い．とはいえ，似た演奏や似た曲は似た拍長のダイナミクスを持ちやすいことを考えると，予測係数関数は森羅万象の演奏及び楽譜を記述するのに必要な空間と比べ，低次元な空間により表せられると考えられる．そこで予測係数関数を，楽譜情報 ϕ_n や演奏情報 $\{\tau_{n'}\}_{n'<n}$ から得られる，少数の低次元特徴量から生成されたものとしてモデル化する．低次元の特徴量や予測係数関数の形は，自明ではなく，手動の設計が困難である．そこで，DNN を用いて，データドリブンに特徴量抽出機と予測係数関数を学習することを考える．

3.1 DNN による予測係数関数

予測係数関数 $a_{p,i}(\cdot)$ は，周辺の楽譜情報と演奏情報 $\tau_{n'<n}$ に依存する．そこで，楽譜情報 ϕ_n から特徴量 $u_n \in \mathbb{R}^U$，演奏情報 $\tau_{n'<n}$ から特徴量 $v_n \in \mathbb{R}^V$ を抽出し，$a_{p,i} = f_{p,i}(u_n, v_n, \theta)$ という形で表す．以後 u_n, v_n をそれぞれ楽譜特徴量，演奏特徴量と呼ぶ．$f_{p,i}$ は二層の全結合層から構成されるニューラルネットワークで，活性化関数は leaky ReLU を用いる．中間層の素子数は 300，出力層の素子数は $P \times I$ である．また，各層にはバッチ正規化 [10] を適用する．ネットワークの詳細を図 2 に示す．

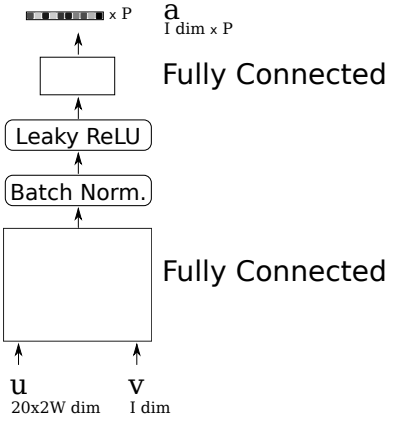

図 2　予測係数関数．楽譜特徴量 u と演奏特徴量 v を入力として，P 拍先までの I 次元の予測係数を出力する DNN である．

3.2　特徴量のモデル化

楽譜特徴量 v_n の算出には，楽譜から得られる次のデータを用いる:

(1) $Q_n^{(1)}$: 拍 n に対する小節内の位置を 16 分音符単位で表したものを，One-hot encoding で表した二値変数

(2) $Q_n^{(2)}$: 小節の拍子を One-hot encoding で表した二値変数

(3) $Q_n^{(3)} \in \{0,1\}^{127\times 12}$: 32 部音符でクオンタイズされ，拍 n から前後八分音符 2 つ分の範囲で算出されたピアノロール．音高は MIDI ノート番号で示している．

このようなデータに対して，図 3 に示すような DNN を適用することで，楽譜特徴量を得る．以下ではその詳細を述べる．

まず，$Q^{(1,2)}$ を入力とした全結合層を通して表現 $Q^{(1,c)}$ を得る．次に，$Q^{(3)}$ からは三層の畳込みニューラルネット (Convolutional Neural Network; CNN) を適用し，全結合層を適用することで低次元表現 $Q^{(3,c)}$ を得る．活性化関数は leaky ReLU であり，各レイヤーではバッチ正規化と最大プーリングを行っている．CNN はテンポ軌跡の変動を適切に説明するよう学習されるため，テンポ変化を説明する上で有用な楽譜の局所的な「型」が学習されると考えられる．特に最下層の畳み込み層ではカーネルサイズを (12 半音 ×32 部音符 2 つ) とすることで，楽曲に頻出するヴォイシングを捉えるような学習を促す．

最終的な特徴量の算出においては，現在の拍位置とその周辺の情報を統合する．そこで，現在の周辺 W 拍で得られた $Q_n^{(1,c)}$，$Q_n^{(3,c)}$ を統合したデータ $\{Q_{n'}^{(1,c)}, Q_{n'}^{(3,c)}\}_{n'=n-W}^{n+W}$ を楽譜特徴量 u_n として用いる．図 4 にピアノ曲の楽譜と対応する楽譜特徴量を図示する．この図から，似たピアノロールから得られる楽譜特徴量が似ていることが分かる．

演奏特徴量 v_n には直近 I 拍に対する拍長の履歴を，平均 0，分散 1 に正規化したものを用いる．これにより，平均テンポに対して不変であるような，テンポ軌跡のダイナミクスに関する情報が得られると考えられる．

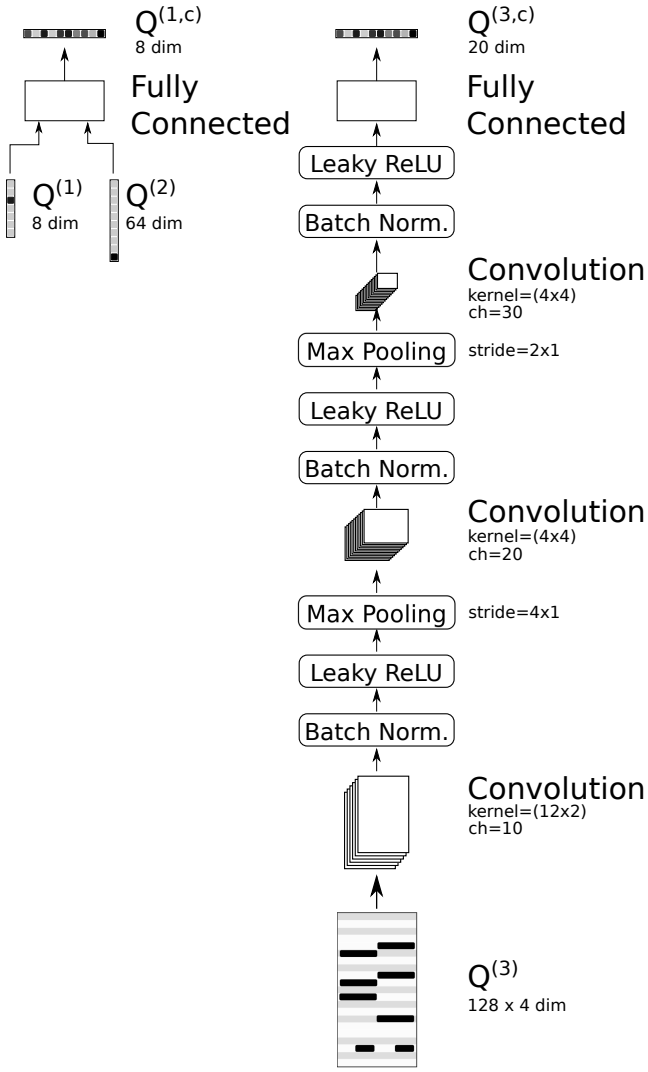

図 3　演奏特徴量算出のための DNN.

3.3　学習

提案手法では，楽譜特徴量を算出するための DNN のパラメータと，予測係数関数 $f(u_n, v_n, \theta)$ のパラメータを学習する必要がある．学習では，学習データと正解となる拍長系列に対して算出される尤度を最大化する．すなわち，予測誤差の累積 L1 ノルムを最小化するよう DNN のパラメータを最適化する．

学習する際には，楽曲の移調に対してタイミングの予測が不変であることから，学習データに対してランダムに-5 から+5 半音の間で移調したものを各エポックにおける学習データとして与える．これにより，学習データに含まれるキーの分布に対する依存が緩和されることが期待される．

3.4　演奏表情付けへの応用に関して

本手法では，楽譜と適当な拍長履歴の初期値を与えたときにタイミング系列を生成することも可能である．ただし，得られる自己回帰モデルパラメータが安定である保証はなく，テンポ軌跡が発散する可能性がある．そこで，実

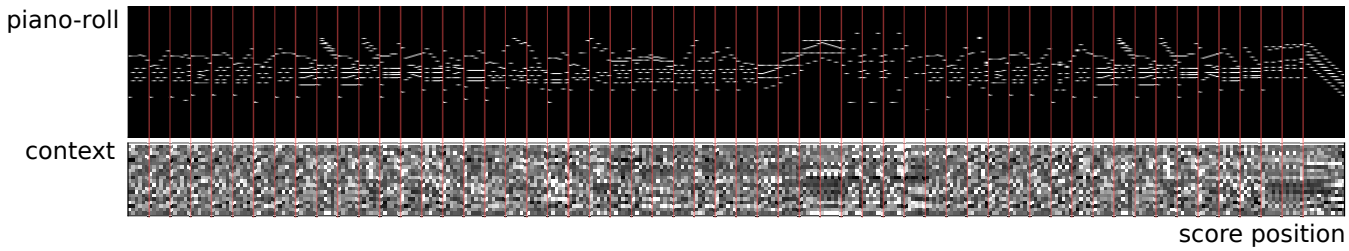

図 4　楽曲のピアノロール（上）とその演奏特徴量（下）.

際には次のような平滑化を行うことが好ましい:

$$\tau_n = (1-\alpha)m + \alpha \sum_{i=1}^{I} a_{1,i}(\theta, \phi_n, \{\tau_{n'}\}_{n'<n})\tau_{n-i}. \quad (2)$$

ここで m は目標となる拍長であり， $\alpha \in [0,1]$ は τ_n を m に近づける程度を示すパラメータである.

4. 評価実験

本手法の有効性を評価した．まず，演奏情報と楽譜情報を併用することの有用性を評価し，次に楽譜特徴量のモデル化に対する妥当性について評価を行った.

4.1 データセット

ピアノソロ曲 52 種類を MIDI 出力機能のあるグランドピアノで演奏したデータを用意した．このうち学習に 41 曲，評価に 11 曲を用いた．楽曲はショパン，リスト，ラフマニノフといった，テンポの抑揚が大きい後期ロマン派の楽曲に加え，ベートーヴェンやモーツアルトといった，テンポの抑揚が比較的穏やかである古典派楽曲が選ばれた．各楽曲に対して，少なくとも 2 名以上の演奏者がデータを収録し，合計 250 個の標準 MIDI ファイルを用意した.

それぞれの演奏データはフリーテンポの演奏であるため，各楽曲の楽譜データに対してアライメントを算出することでテンポトラックを算出した．また，得られたテンポトラックに対して手動による検品を行った．特に，ピアノ演奏では同時に演奏するよう和音がかならずしも同時に発音されないため，検品者が発音時刻と感じる場所に楽譜上の発音時刻が発生するよう修正を施した.

また，勾配法による最適化には ADAM [12] を用い，超パラメータは [12] のものを用いた．また $P = 8$，$I = 24$，$W = 24$ とした.

4.2 実験 1: 演奏・楽譜特徴量を使うことの有用性評価

まず演奏情報と楽譜情報を併用することの有用性を評価した.

4.2.1 実験条件

前述のデータセットを用いてマルチステップ予測に対する誤差を評価した．比較手法として次の手法を用いた:

(1) MA: 拍長の移動平均を予測値とする.
(2) AR: 拍長に対する AR(I) 過程を各予測ステップ数に対して学習し，予測に用いる.
(3) Score Only: 提案手法において演奏特徴量を入力しないもの.

表 1　マルチステップ予測誤差.

予測ステップ数	MA	AR	Score Only	Proposed
1-step	155 ms	121 ms	112 ms	113 ms
2-step	287 ms	234 ms	224 ms	225 ms
4-step	512 ms	444 ms	423 ms	406 ms
8-step	900 ms	789 ms	756 ms	691 ms

4.2.2 実験結果

実験結果を表 1 に示す．Score Only の予測誤差 が MA と AR どちらと比べても低いことが分かる．このことから，譜面情報の使用が発音タイミング予測において有効であることが分かる．また，Score Only と提案手法を比較すると，演奏情報を用いることで 4-step (二分音符先の予測) や 8-step（全音符先の予測）といった，比較的長期間における予測精度が改善することが分かる．誤差中央値を見ると，4-step では Score Only が 72 ms，提案手法が 69 ms であり，また 8-step は Score Only で 137 ms，提案手法で 139 ms である．予測誤差の中央値は大差がないことから，演奏特徴量を用いることにより，外れ値となるような著しい予測ミスが減少していることが分かる.

4.3 実験 2: 楽譜特徴量算出方法の評価

次に楽譜特徴量のモデル化に対する妥当性を評価した.

4.3.1 実験条件

以下に示す条件で楽譜特徴量を算出し，マルチステップ予測誤差を評価した:

(1) 全結合層 1 層（FC）: ピアノロールに対して全結合層を適用することで楽譜特徴量を得る．特徴抽出を行う溜めにピアノロールに対して全結合層を適用する意味では，ダイナミクスの演奏表情付け手法 [11] に近い.
(2) 1 層の畳込み（Conv-1）: 畳み込み層と最大プーリングを経てから，全結合層を適用することで楽譜特徴量を得る.
(3) 2 層の畳込み（Conv-2）: 2 層の畳み込み層を経てから全結合層を適用することで楽譜特徴量を得る.

最適化アルゴリズムの超パラメータは全ての条件で同一のものを用いた.

表 2　楽譜特徴量の算出方法に対するマルチステップ予測誤差.

予測ステップ数	FC	Conv-1	Conv-2	Proposed
1-step	122 ms	119 ms	116 ms	113 ms
2-step	240 ms	226 ms	227 ms	225 ms
4-step	430 ms	411 ms	406 ms	406 ms
8-step	724 ms	696 ms	694 ms	691 ms

4.3.2　実験結果

表 2 に結果を示す．FC と Conv-1 を比べると，畳みこみ層を入れることにより精度が上がることが分かる．これには二つの理由が考えられる．まず，Conv-1 は二層のネットワークであるため，FC よりも柔軟な関数が表現できるためと考えられる．また，Conv-1 では音高軸に対する最大プーリングを行っているため，特定の音列が発生する音域は捉えるがそのオフセットに関しては比較的柔軟になったためと考えられる．

また，Conv-1，Conv-2 と提案手法を比較すると，層が増えると精度が上がることも分かる．ただし，Conv-1 と提案手法の性能差は FC と Conv-1 と比べるとさほど著しくない．このことから，本手法では畳込み層と最大プーリングの導入が，良い楽譜特徴量の獲得においてより有効であることが示唆される．

5.　おわりに

本稿では，楽譜情報と演奏情報の履歴から発音タイミングをオンラインで予測する手法を提案した．タイミング予測において重要な楽譜特徴量を抽出するために，DNN を用いた楽譜情報からの特徴抽出を用いた．また，演奏情報から算出された特徴量を併用した DNN を構築することで自己回帰過程のパラメータを算出する DNN を構築した．評価実験の結果楽譜特徴量と演奏特徴量を導入することでオンラインタイミング予測の精度が向上することが示され，楽譜特徴量の算出において CNN を使うことの有効性が示唆された．

今後の課題としては次のようなものが挙げられる．応用面では，実時間合奏システムへの統合や，拍長以外の音楽要素の予測への応用が挙げられる．また，合奏との統合においてはリハーサルを通じた特定演奏者及び特定楽曲への転移学習や，人間と合奏システムの間に発生する相互作用のモデル化なども今後の課題として挙げられる．理論面では，安定性が保証された回帰係数関数の構築は演奏の生成において重要であり，また，自己回帰過程をガウス過程回帰として拡張することは，非均等に区切られた区間に対する予測を行う上で重要な課題である．

参考文献

[1] Roger B Dannenberg. An on-line algorithm for real-time accompaniment. In *Proc. ICMC*, pages 193–198, 1984.
[2] Akira Maezawa and Kazuhiko Yamamoto. Muens: A multimodal human-machine music ensemble for live concert performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017.*, pages 4290–4301, 2017.
[3] Gerhard Widmer, Sebastian Flossmann, and Maarten Grachten. Yqx plays chopin. *AI Magazine*, 30(3):35, 2009.
[4] Christopher Raphael. A Bayesian network for real-time musical accompaniment. In *Proc. NIPS*, pages 1433–1439, 2001.
[5] Shizuka Wada, Yasuo Horiuchi, and Shingo Kuroiwa. Temo prediction model for accompaniment system. In *Proc. ICMC*, pages 1298–1303, 2014.
[6] Guangyu Xia, Yun Wang, Roger B. Dannenberg, and Geoffrey Gordon. Spectral learning for expressive interactive ensemble music performance. In *Proc. ISMIR*, pages 816–822, 2015.
[7] Kenta Okumura, Shinji Sako, and Tadashi Kitamura. Laminae: A stochastic modeling-based autonomous performance rendering system that elucidates performer characteristics. In *Proc. ICMC*, 2014.
[8] F. Krebs and M. Grachten. Combining score and filter based models to predict tempo fluctuations in expressive music performances. In *Proc. SMC*, Copenhagen, Denmark, 2012.
[9] Aron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *Arxiv*, 2016.
[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.
[11] Sam Van Herwaarden, Maarten Grachten, and W Bas De Haas. Predicting expressive dynamics in piano performances using neural networks. In *Proc. ISMIR*, pages 45–52, 2014.
[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.