

Sponsoring Committee: Professor Juan Pablo Bello, Chairperson
Professor Robert Rowe
Doctor Eric J. Humphrey

AUTOMATIC MUSIC TRANSCRIPTION IN THE DEEP LEARNING ERA:
PERSPECTIVES ON GENERATIVE NEURAL NETWORKS

Jong Wook Kim

Program in Music Technology
Department of Music and Performing Arts Professions

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
Steinhardt School of Culture, Education, and Human Development
New York University
2019

Copyright © 2019 Jong Wook Kim

ABSTRACT

The problem of automatic music transcription (AMT) is considered by many researchers as the holy grail of the field, because of the notorious complexity and difficulty of the problem. Meanwhile, the current decade has seen an unprecedented surge of deep learning where neural network methods have achieved tremendous success in many machine learning tasks including AMT. The success of deep learning is largely enabled by the ever-increasing amount of available data and the innovation of GPU hardware, allowing a deep learning model to enjoy the increased capacity to process such scale of data. While having more data and higher capacity translates better performance in general, there still remains the question of how to design an AMT model that can effectively incorporate the inductive bias for the task and best utilize the increased capacity.

This thesis hypothesizes that an effective way to address this question is through the use of generative neural networks. Starting with a simplified setup of monophonic transcription, we learn the effectiveness of convolutional representation and the roles of dataset choices in data-driven models for music analysis. In the subsequent chapters, we examine the applications of deep generative models in music analysis and synthesis tasks, by introducing a WaveNet-based music synthesis model that learns a multi-dimensional timbre representation and a music language model applied in an adversarial manner to improve a piano transcription model. Finally, we combine the analysis and synthesis methods to develop a multi-instrument polyphonic music transcription system. From these observations, we conclude that deep generative models can be used to improve AMT in many ways, and they will be a crucial component for further advancing AMT.

To Nayoung.

ACKNOWLEDGEMENTS

I am truly grateful to many wonderful people who believed in me along this long journey. This dissertation would not have been able to come into existence without their support and guidance.

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Juan Pablo Bello. I am incredibly lucky to have you as my doctoral advisor; thank you for being a dependable teacher, an incredible researcher, and a welcoming friend to me. To my committee members and readers — Prof. Robert Rowe, Dr. Eric Humphrey, Prof. Johanna Devaney, and Prof. Brian McFee — I deeply appreciate taking your time to read through my awkward sentences and giving insights to make them better. And to all professors with whom I have had the pleasure of working with: thank you for your classes, chats, and smiles.

I have been fortunate enough to become friends with almost all of MARL's PhD students and postdocs, and I am thankful to every one of them. To the MARL-doctors — Taemin, Areti, Jon, Aron, Braxton, Finn, Eric, Uri, Rachel, and Finn — thanks for showing the ways that I can follow, including but not limited to the coffee shops and bars. To the MARL-doctor-to-be's — Andrea, Andrew, Marta, Peter, Yu, Ho-Hsiang, Willie, Dirk, Tom, Jason, and Chris — it was so much fun sharing office and hanging out with you, and I will miss the occasional beer sessions. To the MARL postdocs — Brian, Justin, Mark, Charlie, Ron, Vincent, Claire, Magdalena, Hitomi — thank you for the insights during the lab meetings and collaborations, and also for sometimes bearing with my unscholarliness.

And to my Korean friends: thanks for being on KakaoTalk whenever I had silly memes to share with, but more importantly for being always welcoming and cheering for me every time I visited Korea, especially that time when you guys gladly spent a whole day at my wedding.

I am honored to have been a recipient of Samsung Scholarship, which is another factor that made all this possible; thank you for your support, networking, and gifts from Leeum. I would also like to thank everyone I worked with during my brief industry experiences — NCSOFT, Kakao, Pandora, and Spotify — for allowing me to learn and achieve what I could not do in schools, and of course for all the free foods and swags.

To my parents who have always believed me and prayed for me, I can't express enough gratitude for your limitless love and unwavering support. Your passion in education made me grow from an aspiring teenager to a respectable scientist. Now that there are no more degrees for you to worry about, please take it easy and enjoy your 60s!

Finally, to my wife Nayoung, you are the foremost reason why I am writing these words now. I cannot believe how lucky I am to have met you and plowed through this journey mixed with joys and tears with you. Thank you for being my best friend, a fellow researcher, a delightful travel partner, a world-class cook, a witty comedian, and a lovely cheerleader who makes me a better person every day.

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER	
I INTRODUCTION	1
1. Statement of Problem	1
2. Research Questions	4
3. Limitations	5
3.1 Scope of Music	5
3.2 Symbolic Processing of Notes	6
3.3 On the Need for Perceptual Studies	8
4. Need for Study	9
4.1 Applications of Automatic Music Transcription	10
4.2 Generative Modeling for Fully Capturing Semantics	10
4.3 On the Broader Context of Machine Listening in AI Research	13
4.4 Organization of The Thesis	14
II MUSIC INFORMATION RETRIEVAL FOR TRANSCRIPTION	15
1. Introduction	16
2. Monophonic Pitch Estimation	20
3. Multiple Fundamental Frequency Estimation	21
4. Source Separation and Music Translation	25
5. Machine Learning Models for Music Synthesis	27
6. Music Language Models for Symbolic Music Generation	28
7. Summary	30

III	DEEP LEARNING	31
1.	Neural Network Architectures	32
2.	Performance Optimization Techniques	35
3.	Toward Deep Generative Models	38
3.1	Traditional Generative Models	39
3.2	Early Deep Generative Models and Autoregressive Models	39
3.3	Variational Autoencoders	41
4.	Generative Adversarial Networks	43
4.1	Evolution of the GAN Architecture	44
4.2	The GAN Zoo	44
4.3	Conditional Generation and Other Applications	46
4.4	Evaluation of Generated Samples	47
4.5	Theories on GAN Convergence	48
5.	Summary	49
IV	CREPE: DEEP MONOPHONIC PITCH ESTIMATION	50
1.	Introduction	50
2.	Architecture	52
3.	Experiments	55
3.1	Datasets	55
3.2	Methodology	56
3.3	Results	57
4.	Open-Sourcing CREPE	63
4.1	Python Package and Command-Line Interface	64
4.2	Real-Time Web Demo	66
4.3	Argmax-Local Weighted Averaging	67
4.4	Data-Driven Models and Real-World Applications	68
5.	Conclusions	68
V	LEARNING TIMBRE SPACE FOR MUSIC SYNTHESIS	70
1.	Introduction	71
2.	Background	72
2.1	Timbre Control in Musical Synthesis	72
2.2	Timbre Morphing	72
2.3	Timbre Spaces and Embeddings	73
2.4	Neural Audio Synthesis using WaveNet	73
3.	Method	74
3.1	Timbre Conditioning using FiLM Layers	75
3.2	Model Details	76
4.	Experiments	78
4.1	Datasets	78
4.2	Ablation Study on Model Design	79
4.3	Synthesis Quality	81
4.4	The Timbre Embedding Space	82
5.	Conclusions and Future Directions	84

VI	ADVERSARIAL LEARNING FOR PIANO TRANSCRIPTION	87
1.	Introduction	88
2.	Background	90
2.1	Automatic Transcription of Polyphonic Music	90
2.2	Generative Adversarial Networks and $p_i \times 2p_i \times$	92
3.	Method	94
3.1	Musically Inspired Adversarial Discriminator	95
3.2	TTUR and <i>mixup</i> to Stabilize GAN Training	96
4.	Experimental Setup	98
4.1	Model Architecture	98
4.2	Hyperparameters	99
4.3	Dataset	100
4.4	Evaluation Metrics	100
5.	Results	102
5.1	Comparison with the Baseline Metrics	102
5.2	Visualization of Frame Activations	104
5.3	Training Dynamics and The Generalization Gap	105
6.	Conclusions	105
VII	SYNTHESIZER-AIDED MULTI-INSTRUMENT TRANSCRIPTION	108
1.	Introduction	109
2.	Related Work	111
2.1	Multi-Instrument Music Transcription	111
2.2	Deep Clustering	112
2.3	Generative Modeling for Transcription	113
3.	Method	114
3.1	Synthesizer Model	114
3.2	Training Transcriber with Appended Synthesizer	115
4.	Experimental Setup	117
5.	Results	118
5.1	Synthesizer Output	118
5.2	Transcription Accuracy	120
5.3	Multi-Instrument Transcription	123
5.4	MusicNet Inspector	124
6.	Future Work	125
7.	Conclusions	126
VIII	CONCLUSIONS AND FINAL REMARKS	128
1.	Summary and Takeaways	128
2.	Future Research Directions	130
	BIBLIOGRAPHY	134