Deep Neural Network for Multi-Pitch Estimation Using Weighted Cross Entropy Loss

Samuel Stone
SRC, Inc.
North Syracuse, NY
sstone@srcinc.com

Evan Spector
SRC, Inc.
North Syracuse, NY
espector@srcinc.com

***Abstract*— Multi-Pitch Estimation, the estimation of multiple overlapping or polyphonic harmonic fundamental frequencies, has a wide range of applications, including automatic music transcription, power systems, and radar signal processing. Multiple fundamental frequencies represent a challenge due to the added complexity of the overlapping signals. This paper presents a Deep Learning approach to estimating multiple fundamental frequencies. The network is trained in a supervised fashion to generate a pseudospectrum representing the fundamental frequencies. Training data is represented by a sparse binary vector the size of the pseudospectrum, indicating the location of fundamental frequencies. A weighted binary cross-entropy loss function is used to correct for class imbalance caused by the sparsity of the signal space relative to the full spectrum. We show comparable performance to existing techniques while requiring fewer operations and samples due to a simpler frequency-domain-only architecture.**

Keywords*— Frequency Estimation, Machine Learning, Harmonic Analysis

I. Introduction

Harmonics are signals which are multiples of a fundamental frequency. Depending on the source of the signal, a harmonic signal can have a finite number of peaks, such as in Helicopter Rotor Modulation [6]. The equation for a harmonic signal with N harmonics and M fundamental frequencies is shown below.

$$x(t) = \sum_{i=1}^{M} \sum_{n=1}^{N} \cos\left(2 * pi * f_i t\, n\right) \quad (1)$$

Harmonic signals come from a variety of sources, such as non-linearities in electronics and music. Differing from music signal transcription applications, we strive to estimate fundamental frequencies accurately and minimize false alarms rather than maximize a perception-based metric such as pitch salience. Conventional methods are subject to errors such as frequency halving/doubling [2] or rely on domain-specific knowledge of the signal [5]. Other methods relying on detections work well for high signal-to-noise ratio but perform poorly when detections are missed and rely on heuristics to discern between higher harmonics or identify spurious signals [3]. To mitigate these, we apply a machine learning algorithm to determine the fundamental frequencies.

The neural network architecture chosen is based on processing only the frequency domain signal with a single FFT. Our approach differs from recent work uses versions of the constant Q transform (CQT) that generate either a two-dimensional time-frequency representation of the signal [5] or three-dimensional time-frequency-octave representation [4]. This has a couple of advantages in applying Deep Learning to multi-pitch estimation. Including the time component means that frequency resolution and coherent processing gain are lost compared to a frequency-only representation of the same number of samples. This also eases processing, as the neural network can be constructed using fewer feature channels, and fewer samples must be pre-processed.

II. Approach

Time domain samples of a multi-pitch signal are converted to the frequency domain using the Fast Fourier Transform. The resulting FFT spectrum of the signal is converted to the log frequency domain through the constant Q transform [1].

On the log-frequency scale, harmonics are equally spaced regardless of fundamental frequency. A Deep Neural Network (DNN) Architecture using convolutional blocks is used for model, based on the work of Bittner, McFee, and Bello [5].

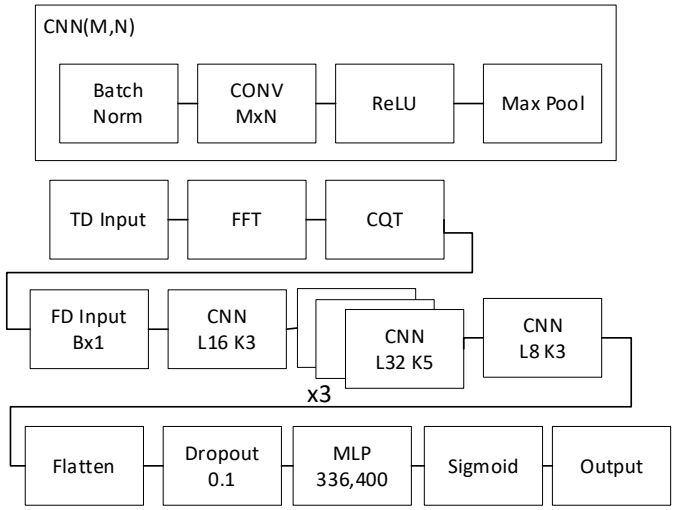

Fig. 1. Deep Network Architecture and preprocessing blocks including Convolutional Neural Network blocks with various number of layers and kernal sizes

Following pre-processing using FFT and Constant Q Transform, the signal is passed through this Deep Neural Network. This network uses a sequence of convolutional blocks

978-1-6654-3630-4/21/$31.00 ©2021 IEEE

with varying number of layers and kernel size. A sigmoid function scales the signal to a range (0,1) at the output.

The network treats comparison between the input and output as a binary classification problem, where the classes are either true for the presence of a signal at the frequency bin or false in bins where there is no signal. With ordinary binary cross-entropy, the reward from true negatives overwhelms the reward for true positives, causing the network to estimate low power in all frequency bins regardless of the input signal. In our approach, objective is determined by comparing the output pseudospectrum with the binary frequency label using the weighted binary cross-entropy loss function in Equation 2 with prediction $\hat{p}$, label p, and weights β_1, β_2 [7].

$$WBCE(p,\hat{p}) = -\beta_1\, p \log(\hat{p}) - \beta_2(1-p)\log(1-\hat{p}) \quad (2)$$

The weight is treated as a hyperparameter that increases the impact of true detection on the reward. The balance between the reward for true positives and true negatives is adjusted by changing this value. Because true positives represent the minority class in the sparse signal vector, ratio of the weight $\frac{\beta_1}{\beta_2} > 1$ is used. We scale the weight based on the expected number of signals and the number of frequency bins.

III. Training Data

A common problem in training supervised machine learning models for multi-pitch estimation is the lack of accurately labeled training data. For this work, training and testing data is generated through simulation. A pseudorandom number generator is used to select signal and noise parameters from a range of realistic inputs. Signal parameters chosen are the fundamental frequencies, number of harmonics fundamental frequency, and the signal to noise ratio of the harmonics. The range of parameters used in training and testing is shown in the figure below.

TABLE I. Simulated Signal Parameters Range

Parameter	***Minimum***	***Maximum***
Number of Fundamental Frequencies	1	6
Fundamental Frequency	30Hz	400Hz
Number of Harmonics of each F0	5	20
Signal to Noise Ratio	10dB	40dB

This range of parameters is selected to represent real-world signal parameters accurately.

IV. Results

The figures below illustrate the CQT of a simulated signal and an example output of the network with this sample input. In Figure 2, the density of a signal with five harmonics is shown.

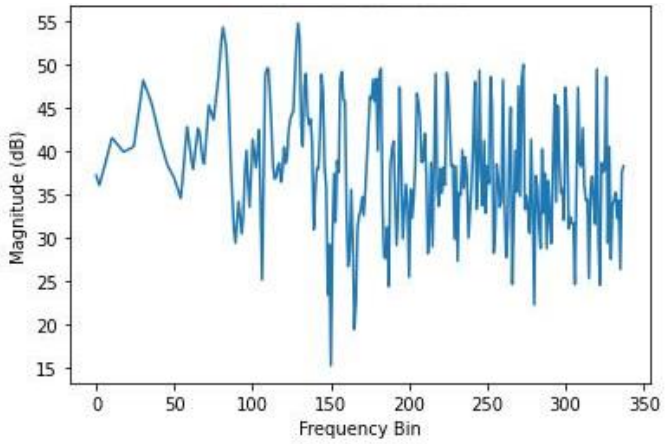

Fig. 2. Constant Q Transform of an input multi-pitch signal

This signal would typically represent a significant challenge for estimators, especially with the minimal spacing between some adjacent frequencies.

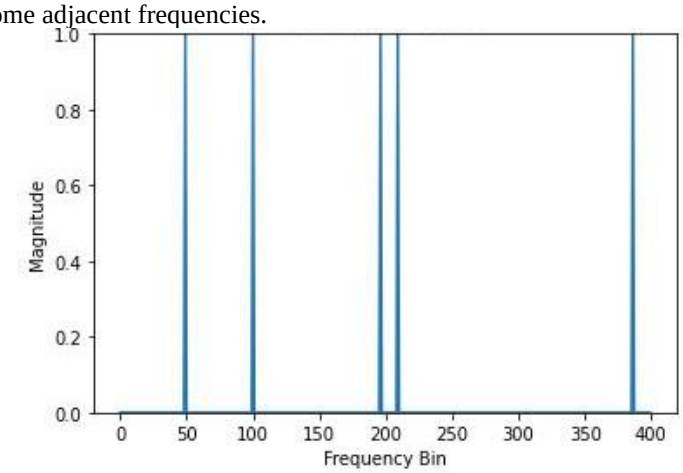

Fig. 3. Truth label boolean mask of an example input signal

The generated signal truth is a vector of Boolean values for each frequency bin, where bins with a signal present are true. The Deep Neural Network output is shown below.

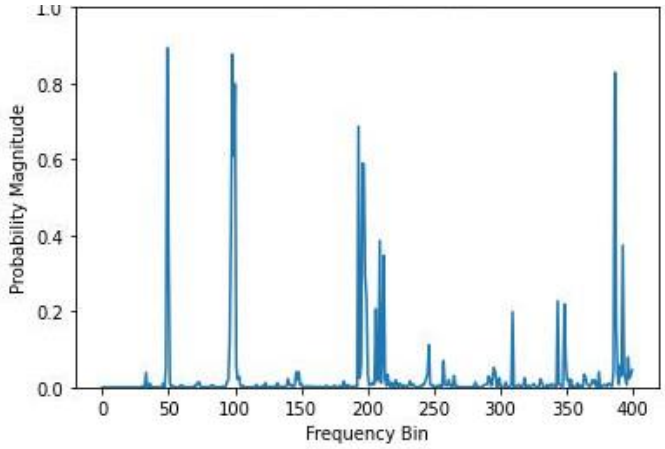

Fig. 4. Deep Neural Network Pseudospectrum of an example signal

This example figure shows that the network can generate a pseudospectrum that accurately represents the input fundamental frequencies.

The receiver operator characteristic was constructed to evaluate model performance across a range of conditions. For a single trained model, we vary the threshold for detection from 0 to 1 against the model and compare detected frequencies to truth. True detection is evaluated based on whether detected frequencies were found within 1Hz of the truth. False-positive peaks are other peaks. Performance was evaluated by computing the mean performance of the model output over batches of simulated data for each of the steps in the detector.

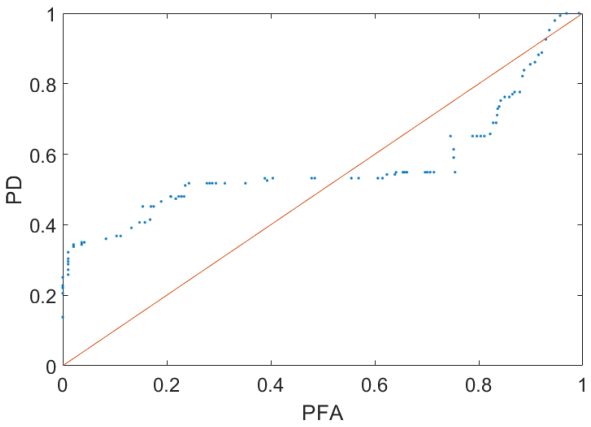

Fig. 5. Receiver-Operator Characteristic

With a probability of false alarm set to 0.05, the model can achieve a probability of detection of around 35 percent. This result is comparable to the accuracy of the single-task model trained in reference [5] in a significantly smaller number of operations due to the frequency-only representation of the signal. The receiver operator characteristic is biased at higher false alarm rates by the unequal distributions between the probability of false alarm and the probability of detection.

V. Conclusions and Future Work

Deep Neural networks show promise as a technique to estimate multiple fundamental frequencies. This paper also demonstrates the ability of a DNN to generate a pseudospectrum for frequency estimate. Weighted binary cross-entropy loss was found to be essential to the generation of this pseudospectrum, as, without it, the reward from correctly predicting empty frequency bins overshadows the reward for correctly estimating frequencies. However, there is still room for improvement in this technique, as the weighting for negative bins applies to both false negatives and true negatives, increasing the probability of false alarm. Weighted binary cross-entropy could be improved by separate weighting for true negatives and false positives. In addition, the current architecture makes use of a relatively small kernel size in the first layer, which may impact the networks' ability to predict harmonics. Other techniques for fundamental frequency estimation would consider more frequency bins at a time [2]. This technique or a version of it could potentially be used to directly replace traditional digital signal processing or super-resolution techniques in a variety of applications such as angle of arrival estimation or multi-path error mitigation in radar signal processing.

References

[1] J. C. Brown and M. S. Puckette, “An efficient algorithm for the calculation of a constant Q transform,” The Journal of the Acoustical Society of America, vol. 92, no. 5, pp. 2698–2701, Nov. 1992, doi: 10.1121/1.404385.

[2] M. Christensen and A. Jakobsson, Multi-Pitch Estimation. 2016. Accessed: Jun. 24, 2021. [Online]. Available: http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781598298390

[3] Zhiyao Duan, B. Pardo, and Changshui Zhang, “Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions,” IEEE Trans. Audio Speech Lang. Process., vol. 18, no. 8, pp. 2121–2133, Nov. 2010, doi: 10.1109/TASL.2010.2042119.

[4] M. Taenzer, S. I. Mimilakis, and J. Abeßer, “Informing Piano Multi-Pitch Estimation with Inferred Local Polyphony Based on Convolutional Neural Networks,” Electronics, vol. 10, no. 7, p. 851, Apr. 2021, doi: 10.3390/electronics10070851.

[5] R. M. Bittner, B. McFee, and J. P. Bello, “Multitask Learning for Fundamental Frequency Estimation in Music,” arXiv:1809.00381 [cs, eess, stat], Sep. 2018, Accessed: Sep. 24, 2021. [Online]. Available: http://arxiv.org/abs/1809.00381

[6] S. Stone, Davis, Eric, and Nashed, Kerolos, “UAS Rotor Length and Multiple Rotor RPM Estimation using SRC Inc.’s Precision Fire Control Radar,” unpublished.

[7] M. R. Rezaei-Dastjerdehei, A. Mijani, and E. Fatemizadeh, “Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function,” in 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), Tehran, Iran, Nov. 2020, pp. 333–338. doi: 10.1109/ICBME51989.2020.9319440.