

Oregon State University

Transposition of Simple Waveforms

from Raw Audio with Deep Learning

Patrick J Donnelly and Parker Carlson

12 April 2023

soundbendor lab

Introduction

The **task** of **automatic transposition** attempts to remap frequencies of an audio signal to different frequencies, while preserving time.

Introduction

The **task** of **automatic transposition** attempts to remap frequencies of an audio signal to different frequencies, while preserving time.

This **problem** is difficult because it requires time-consuming operations, such as **spectral manipulation** or **resampling and resynthesis**.

Introduction

The **task** of **automatic transposition** attempts to remap frequencies of an audio signal to different frequencies, while preserving time.

This **problem** is difficult because it requires time-consuming operations, such as **spectral manipulation** or **resampling and resynthesis**.

Our **goal** is to **transpose raw audio** directly from the signal without Fourier analysis or feature extraction.

Related Work

Time-domain Approaches

resample and time modification

- Synchronised Overlap-Add
  - `SOLA` (1985) [1]
- TD Pitch Synchronous Overlap-Add
  - `TD-PSOLA` (1990) [2]
- Waveform Similarity Overlap-Add
  - `WSOLA` (1993) [3]
- Adaptive Overlap-add
  - `AOLA` (2000) [4]
- Canonical Time Warping
  - `CTW` (2009) [5]

Related Work

Time-domain Approaches

resample and time modification

- Synchronised Overlap-Add
  - `SOLA` (1985) [1]
- TD Pitch Synchronous Overlap-Add
  - `TD-PSOLA` (1990) [2]
- Waveform Similarity Overlap-Add
  - `WSOLA` (1993) [3]
- Adaptive Overlap-add
  - `AOLA` (2000) [4]
- Canonical Time Warping
  - `CTW` (2009) [5]

Phase Vocoders

analyze and resynthesize

- **modulation vocoder** [6]
  - selective transposition
- **non-linear vocoder** [7]
  - intonation adjustments

Related Work

Time-domain Approaches

resample and time modification

- Synchronised Overlap-Add
  - `SOLA` (1985) [1]
- TD Pitch Synchronous Overlap-Add
  - `TD-PSOLA` (1990) [2]
- Waveform Similarity Overlap-Add
  - `WSOLA` (1993) [3]
- Adaptive Overlap-add
  - `AOLA` (2000) [4]
- Canonical Time Warping
  - `CTW` (2009) [5]

Phase Vocoders

analyze and resynthesize

- **modulation vocoder** [6]
  - selective transposition
- **non-linear vocoder** [7]
  - intonation adjustments

Pitch Correction

- Autotune[a]
- VariAudio[b]
- Melodyne[c]

---

[a] http://www.antarestech.com
[b] https://www.steinberg.net
[c] https://www.celemony.com

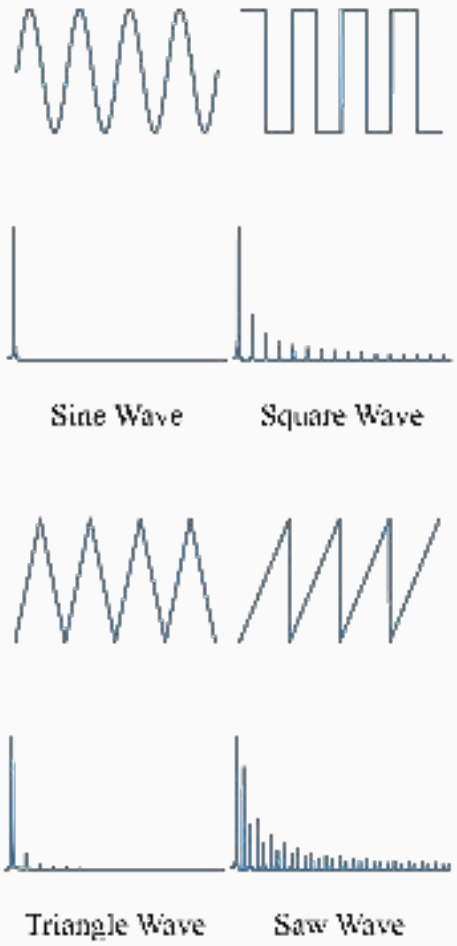
Dataset

Waveform Generation

- one second waveforms
- sr=16 kHz, mono WAV file
  - frequencies up to 8 kHz
- $n = 31,840$ examples

Input and Output

- 12 datasets
  - 1 to 12 semitones
- Input: $[20, 4000]$ Hz
- Output: $[20 \cdot d, 4000 \cdot d]$ Hz
  - where $d = 2^{\frac{s}{12}}$
  - for s step size



Time domain (top) and frequency domain (bottom) for each waveform shape.

Experimental Design

Model

- Bidirectional LSTM
  - three stacked layers
  - 1024 hidden nodes
  - **recurrent dropout** of 0.4
  - batch normalization

- Training sequences
  - chunks of 1024 samples
    - full cycle of 20 Hz
  - zero-pad last chunk
  - **batch size** of 64
  - **learning rate** of 0.0001
  - 250 epochs

Experimental Design

Model

- Bidirectional LSTM
  - three stacked layers
  - 1024 hidden nodes
  - **recurrent dropout** of 0.4
  - batch normalization
- Training sequences
  - chunks of 1024 samples
    - full cycle of 20 Hz
  - zero-pad last chunk
  - **batch size** of 64
  - **learning rate** of 0.0001
  - 250 epochs

Train & Test

- Examples every 1.0 Hz
  - **Train**: *freq*
  - **Test**: $freq + 0.5$
- Model by semitone ($n = 12$)

Example

Train: F_0 of 440.0, 441.0, 442.0, ...
Test: F_0 of 440.5, 441.5, 442.5, ...

Evaluation

- Librosa[a], Rubberband[b]
- error (time, frequency), noise

[a] https://librosa.org/
[b] https://breakfastquay.com/rubberband/

Results: Mean Squared Error

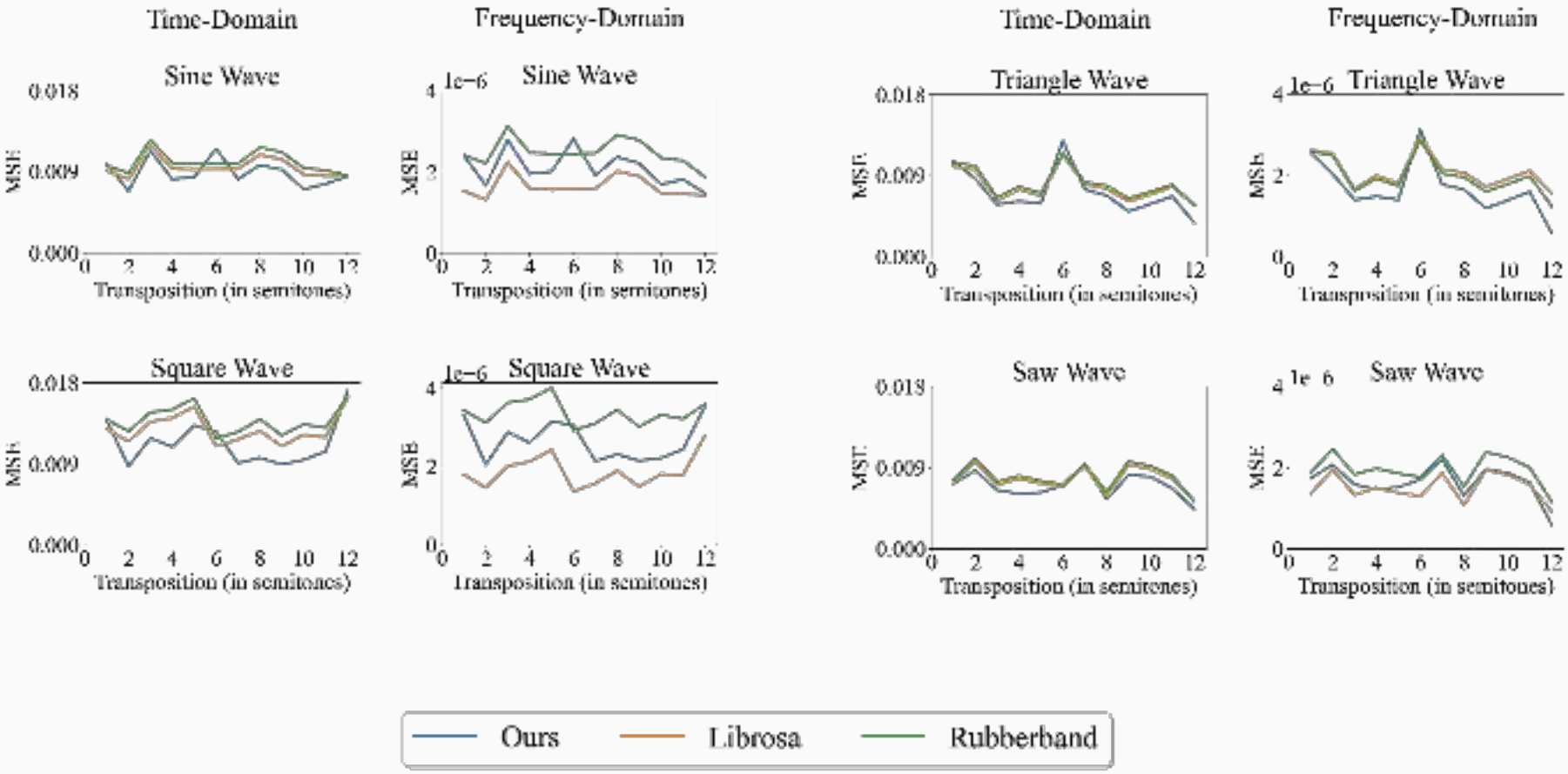
Time domain

Model	Sine	Square	Triangle	Saw	Overall
Our Approach	**0.0089**	**0.0113**	**0.0071**	**0.0070**	**0.0086**
Librosa	0.0094	0.0129	0.0078	0.0077	**0.0094**
Rubberband	0.0100	0.0138	0.0081	0.0080	**0.0100**

Frequency domain

Model	Sine	Square	Triangle	Saw	Overall
Our Approach	0.00000209	0.00000265	**0.00000168**	**0.00000163**	**0.00000201**
Librosa	**0.00000163**	**0.00000188**	0.00000209	0.00000150	**0.00000177**
Rubberband	0.00000247	0.00000338	0.00000199	0.00000195	**0.00000245**

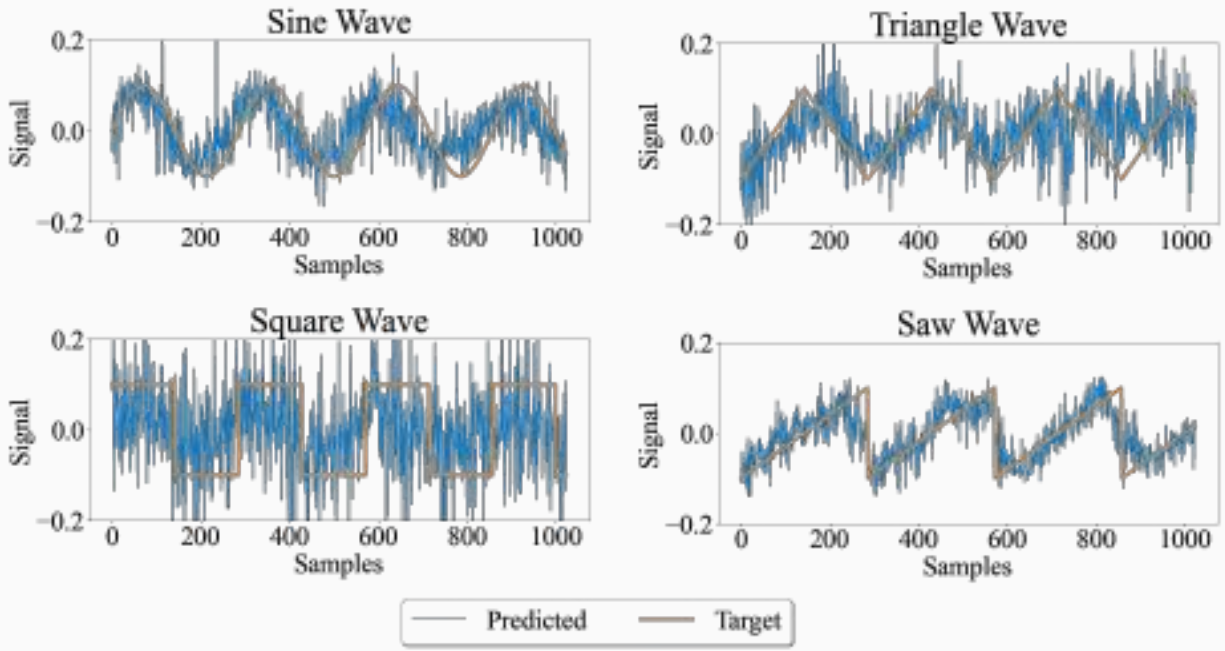
Results: Mean Squared Error



Error by transposition size in semitones and for each waveform type.

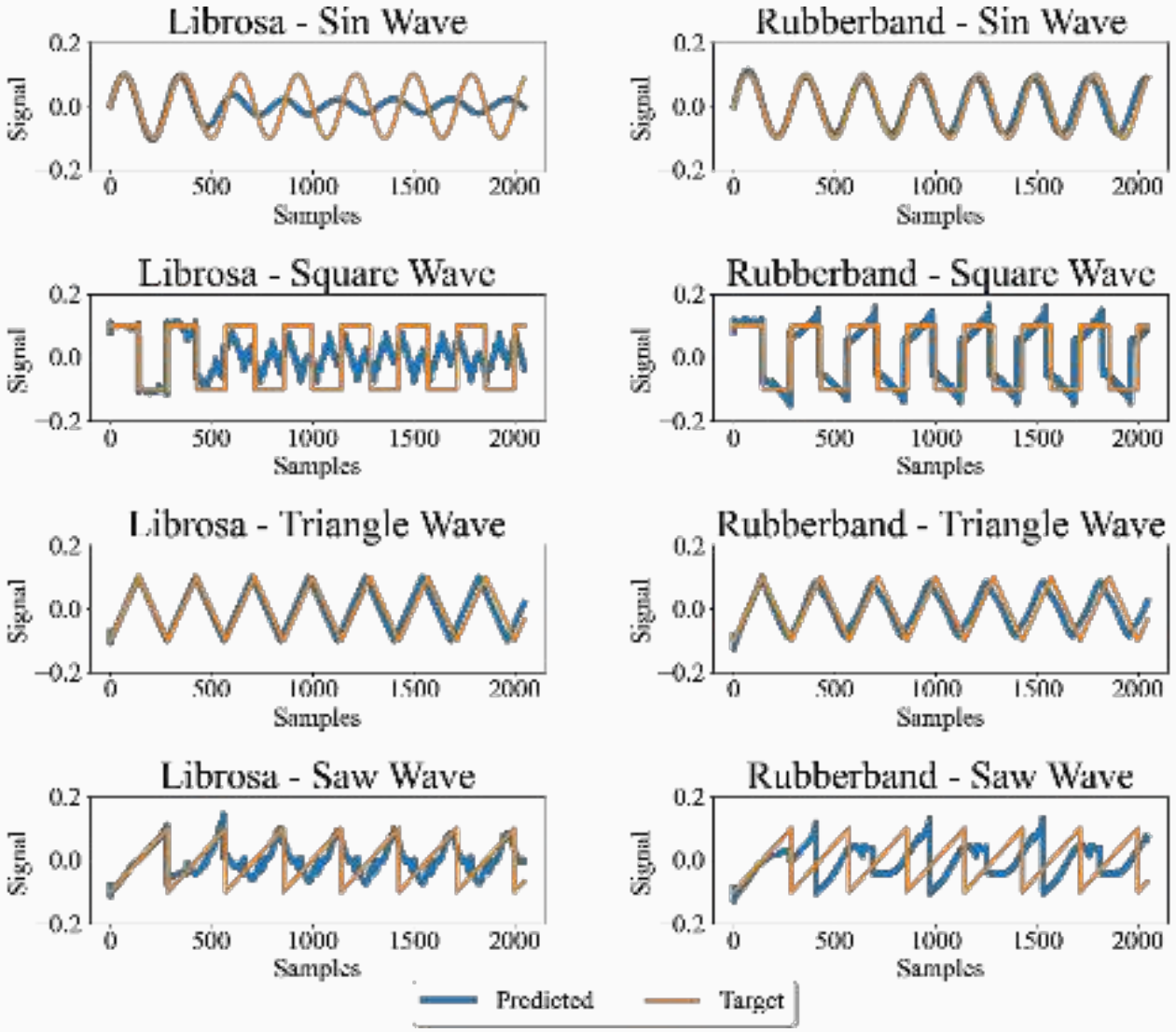
Results: Signal to Noise Ratio

Model	Sine	Square	Triangle	Saw	Overall
Our Approach	2.65	0.95	2.49	0.23	**1.58**
Librosa	**44.55**	**11.53**	**29.23**	**8.65**	**23.49**
Rubberband	41.05	8.93	27.41	6.31	**20.93**



Example of one octave ($s = 12$) transposition to 57 Hz.

Comparison: Librosa and Rubberband



Transposition Examples

Interval	Actual	Prediction	Librosa	Rubberband
m3	Play	Play	Play	Play
P5	Play	Play	Play	Play
P8	Play	Play	Play	Play

Limitations and Future Work

Limitations

Output is noisy

- jagged waveform shape
- imperfect approximations

Future Work

1. Deeper architectures
   - deeper recurrent nets
   - dilated convolution [8]
   - differential DSP [9]
2. Post-processing
   - denoising algorithms
   - ML approaches

Limitations and Future Work

Limitations

Output is noisy

- jagged waveform shape
- imperfect approximations

Simple waveforms

- four simple waveforms
- no standard evaluation metrics for this task

Future Work

1. Deeper architectures
   - deeper recurrent nets
   - dilated convolution [8]
   - differential DSP [9]
2. Post-processing
   - denoising algorithms
   - ML approaches

Real-world waveforms

- large datasets, NSynth [10]
- real-world examples
  - speech, polyphonic music

Conclusion

Summary

We trained deep recurrent neural networks to transpose raw audio *without explicit spectral analysis or manipulation*.

- approximate Fourier processes with deep learning
- produce accurate transposition
- scale to any step size and complex timbres
- **BUT** output is too noisy for use

Conclusion

Summary

We trained deep recurrent neural networks to transpose raw audio *without explicit spectral analysis or manipulation*.

- approximate Fourier processes with deep learning
- produce accurate transposition
- scale to any step size and complex timbres
- **BUT** output is too noisy for use

Applications

Intelligent systems that can perform transposition in near real-time.

- pitch correction
- music learning, analysis
- hearing aids

References i

[1] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 10, 1985, pp. 493–496. DOI: 10.1109/ICASSP.1985.1168381.

[2] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990. DOI: 10.1016/0167-6393(90)90021-Z.

[3] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 2, 1993, pp. 554–557. DOI: 10.1109/ICASSP.1993.319366.

[4] B. Lawlor and A. D. Fagan, "A novel efficient algorithm for music transposition," *Organised Sound*, vol. 4, no. 3, pp. 161–167, 2000. DOI: 10.1017/S135577180000306X.

[5] F. Zhou and F. d. l. Torre, "Canonical time warping for alignment of human behavior," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, ser. NIPS'09, Vancouver, British Columbia, Canada: Curran Associates Inc., 2009, pp. 2286–2294, ISBN: 9781615679119. DOI: 10.5555/2984093.2984349.

[6] S. Disch and B. Edler, "Frequency selective pitch transposition of audio signals," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2011, pp. 29–32. DOI: 10.1109/ICASSP.2011.5946320.

[7] S. Rosenzweig, S. Schwär, J. Driedger, and M. Müller, "Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings," in *2021 24th International Conference on Digital Audio Effects (DAFx)*, IEEE, 2021, pp. 121–128. DOI: 10.23919/DAFx51585.2021.9768268.

[8] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016. DOI: 10.48550/arXiv.1609.03499.

[9] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr.

[10] J. Engel, C. Resnick, A. Roberts, *et al.*, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning*, PMLR, 2017, pp. 1068–1077. DOI: 10.48550/arXiv.1704.01279.