

Combolutional Neural Networks

Cameron Churchwell, Minje Kim, Paris Smaragdis

University of Illinois at Urbana-Champaign, Siebel School of Computing and Data Science, Urbana IL, USA, 61801

Abstract—Selecting appropriate inductive biases is an essential step in the design of machine learning models, especially when working with audio, where even short clips may contain millions of samples. To this end, we propose the combolutional layer: a learned-delay IIR comb filter and fused envelope detector, which extracts harmonic features in the time domain. We demonstrate the efficacy of the combolutional layer on three information retrieval tasks, evaluate its computational cost relative to other audio frontends, and provide efficient implementations for training. We find that the combolutional layer is an effective replacement for convolutional layers in audio tasks where precise harmonic analysis is important, e.g., piano transcription, speaker classification, and key detection. Additionally, the combolutional layer has several other key benefits over existing frontends, namely: low parameter count, efficient CPU inference, strictly real-valued computations, and improved interpretability.

1. INTRODUCTION

Feature engineering in machine learning has largely been replaced by model architectures with inductive biases which encode knowledge of trends in the target distribution. For example, convolutional networks classify handwritten digits with equal accuracy to fully-connected networks while requiring significantly fewer parameters, which also reduces training complexity [1]. By selecting an appropriate inductive bias, we bypass the need to hand-craft features and allow the model to learn more tailored features as part of the model training: CNNs encode dependencies of adjacent samples [2]; RNNs encode backward dependencies in time [3]; LSTMs encode longer-term backward dependencies [4]; Transformers encode dependencies across entire sequences [5]. However, the improvements in efficiency enabled by these models do not erase the challenges inherent in modeling long sequences of audio samples.

Many models avoid these challenges by relying on dimensionality reduction through time-frequency domain transforms and hand-crafted features to attain state-of-the-art performance [6]–[8]. The mel spectrogram is an especially popular choice for speech-related tasks as it utilizes frequency scaling informed by human auditory perception [9], [10]. While this is an effective approach for many tasks, the exact choice of features used can lead to loss of task-critical information, which, in turn, can have significant impacts on model performance [11].

To avoid task-feature mismatch issues with hand-crafted features, a more recent approach is to learn features from the raw inputs. Models such as Conv-TasNet for source separation [12], Wav2vec for self-supervised feature learning [13], and WaveNet for neural vocoding [14], operate on waveforms directly using convolutional layers, capitalizing on the relatedness of neighboring samples to learn relationships which might not be captured by DSP-style features. The choice of CNNs to model audio is strongly supported by signal processing theory, as convolutional layers are equivalent to learned finite impulse response (FIR) filters [15]. Notably, the filters learned by these models are unconstrained apart from by their kernel size, in the same way that a fully-connected network is unconstrained compared to a CNN. Meanwhile, convolutional layers are not always the most suitable operation for signal processing, often requiring a large model to properly handle the learning task involving raw waveform signals [14].

The practice of parameterizing DSP elements for use in neural networks is commonly referred to as differentiable digital signal processing (DDSP) [15]. The most well known DDSP application is to use neural networks to predict synthesizer parameters to reconstruct realistic audio of musical instruments [16]. Other works have demonstrated the potential for machine learning models to learn DSP filters, either to digitalize existing analog filters [17], [18] or to learn input features for a downstream task, such as SincNet [19], where bandpass filters with parameterized cutoff frequencies are incorporated as part of the neural network. This acts as an additional degree of weight-sharing which further reduces parameter count and improves interpretability. Based on this line of work, we propose a simple and efficient neural network layer which parameterizes comb filters, and use it for applications where an efficient analysis of periodic sound components is crucial.

In this work we propose the use of feedback comb filters for audio modeling applications as a replacement for unconstrained convolutional layers. The feedback (IIR) comb filter passes the original signal through and sums it with a delayed and scaled copy of its own output. For a delay of K samples and a feedback gain of $0 < \alpha < 1$, the filter is written as:

$$y[n] = x[n] + \alpha y[n - K] \quad (1)$$

Combined with an envelope detector, the feedback comb filter may be used as an efficient, learned harmonic feature extractor we call the *combolutional layer*. The combolutional layer offers several key benefits over other audio model frontends: each output channel of the combolutional layer requires only a single parameter; each output sample requires only a single multiply-accumulate (MAC) operation at inference time, making combolutional layers especially attractive for on-device applications.

Our primary contributions are as follows:

- We introduce the combolutional layer, a novel parameteric, thus trainable, function that replicates the behavior of a comb filter. Along with other neural network layers, a combolutional network (CombNet) that employs a combolutional layer works as a learnable harmonic feature extractor.
- We show that the combolutional layer is more efficient both in terms of parameter count and MACs than an equivalent convolutional layer.
- We provide efficient Triton implementations of the non-recursive proxy of the combolutional layer to be used during training. In this way, a combolutional layer can also leverage GPU computing.
- We empirically show that CombNets compete with other ConvNet architectures and more efficient ones, such as SincNet, on three harmonically sensitive audio tasks, showcasing efficient inference-time properties.

2. COMB FILTERS AS FEATURE EXTRACTORS

2.1. Definition of Comb Filters

The magnitude response of the feedback (IIR) comb filter with a gain of $0 < \alpha < 1$, sampling rate f_s , and delay of K samples, is given

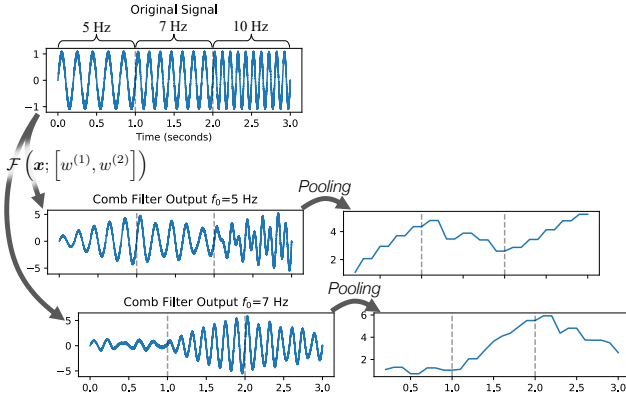


Fig. 1: A combolational layer with two internal comb filters applied to an input signal comprised of three sequential tones of equal duration at 5 Hz, 7 Hz and 10 Hz.

by:

$$|H(f)| = \frac{1}{\sqrt{1 + \alpha^2 - 2\alpha \cos(2\pi f / f_s K)}} \quad (2)$$

If we choose $K = f_s / f_0$, we can rewrite this in terms of a tunable frequency parameter f_0 :

$$|H(f)| = \frac{1}{\sqrt{1 + \alpha^2 - 2\alpha \cos(2\pi f / f_0)}} \quad (3)$$

In simple terms, this filter amplifies signal components with frequencies near to integer multiples of f_0 while attenuating all other frequencies. Our goal in developing combolational layers is to replace the comb filter with a surrogate function which is continuous and differentiable in f_0 . As a result, we obtain a learnable filter with a single parameter, f_0 , that amplifies the frequencies in the vicinities of f_0 and all of its harmonics. As with most neural network layers, a combolational layer can learn multiple channels in parallel, each corresponding to a different comb filter with its own f_0 value. This learned comb filter bank can be combined with the typical activation functions, max pooling, and other additional operations as necessary.

A significant benefit of learning such a combolational layer is that it can learn a small and custom set of filters tailored for the task rather than using an over-complete dictionary of many filters. In addition, compared to a convolutional layer, the inherent periodicity of the filters can introduce an inductive bias as well as an additional efficiency to the learning task, mitigating the effort of learning such a structure within a convolutional layer. It should be noted that these advantages come with the cost of the potentially suboptimal assumption that the analysis of the input signal benefits from the periodic filters. Fig. 1 depicts the harmonic nature of the combolational layer feature representation. The filter with $f_0 = 5$ Hz detects activity for both the 5 Hz and 10 Hz sine waves at the beginning and end of the signal, but is not too effective in detecting the 7 Hz sine wave. Note that the time-lag shown in the right panel of 1 is exaggerated due to the relatively low frequencies and correspondingly large filter delays.

2.2. Definition of Combolational Layers

A combolational layer can be defined as a parametric function $\mathcal{F}(\cdot)$ that takes input signal $\mathbf{x} \in \mathbb{R}^T$ and calculates the filter responses \mathbf{y} :

$$\mathbf{Y} \leftarrow \mathcal{F}(\mathbf{x}; \mathbf{w}), \quad (4)$$

where $\mathbf{w} = [w^{(1)}, w^{(2)}, \dots, w^{(M)}] \in \mathbb{R}^M$ denotes the set of M trainable model parameters, representing the fundamental frequencies of M combolational kernels. Consequently, there are M output

channels as a result of M simultaneous filtering operations, i.e., $\mathbf{Y} \in \mathbb{R}^{M \times T'}$, where T' is the length of the response signal. We then take the absolute value and apply max pooling as shown in Fig. 1.

Since fundamental frequencies f_0 are defined as nonnegative continuous value in Hz, instead of optimizing the model for f_0 directly, we use a simple scaling function $s: \mathbb{R} \rightarrow [f_{\min}, f_{\max}]$ that converts the freely learned model parameter $w^{(m)}$ into its corresponding fundamental frequency value if necessary:

$$f_0^{(m)} = s(w^{(m)}) = f_{\min} \cdot \left(\frac{f_{\max}}{f_{\min}}\right)^{\sigma(w^{(m)})}, \quad (5)$$

where $\sigma(\cdot)$ is the sigmoid function to introduce nonnegativity into conversion, and f_{\min} and f_{\max} are pre-defined hyperparameters, i.e., the frequency range that the input signal can potentially cover. Note that the exponential scaling is important for training, as a small change $\Delta w^{(m)}$ will result in a roughly equivalent perceptual change $\Delta f_0^{(m)}$, invariant to the current value of $w^{(m)}$.

3. EFFICIENT IMPLEMENTATIONS

Eq. (1) shows the typical discrete computation of a feedback comb filter. While this implementation may be fine for inference time, the discrete nature is not compatible with automatic differentiation packages, nor is it efficient on GPU hardware, where inner product-based parallelization is crucial for efficiency.

3.1. Non-Recurrent Approximation

As the first step toward efficient GPU computation, during training, the comb filters are approximated by FIR surrogates via truncated impulse response estimation. The kernel for these filters is straightforward once α and K are defined:

$$h_K^{(m)}[n] = \begin{cases} \alpha^t & \text{if } n = tK, \quad t \in \mathbb{Z}^+ \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

In practice, once a combolational layer parameter $w^{(m)}$ is updated, it is converted to corresponding $f_0^{(m)}$ using eq. (5). Based on the known relationship $K = f_s / f_0$, K is inferred. α is a hyperparameter we set to 0.9 for all experiments.

At each forward pass, this kernel can be used in a standard convolution:

$$\mathbf{Y}_{m,:} \approx h_K^{(m)} * \mathbf{x}. \quad (7)$$

Now, the computation no longer depends on y , and so can be fully parallelized across timesteps, which is critical given the alternative is a recurrent operation with potentially millions of time steps.

3.2. Differentiable Proxy

Another important step is to relax the definition of comb filtering eq. (1) with a discrete delay K in a continuous version to enable gradient computation. This is because during feedforward, the conversion eq. (5) results in a continuous f_0 value, leading to a continuous delay value, while the modulo operation in eq. (6) is defined with an integer K . For clarity, we introduce \bar{K} as a continuous version of K learned during training.

To resolve this issue, we use interpolation between two filter responses based on two adjacent integer delays $\lfloor \bar{K} \rfloor$ and $\lceil \bar{K} \rceil$ to compute a continuous proxy of the comb filter with a continuous delay \bar{K} . We find experimentally that linear interpolation between the two is sufficient to compute approximate gradients, turning the filtering into a differentiable operation. The interpolation is defined as follows:

$$\mathbf{Y}_{m,:} \approx (1 - \beta(\bar{K})) h_{\lfloor \bar{K} \rfloor}^{(m)} * \mathbf{x} + \beta(\bar{K}) h_{\lceil \bar{K} \rceil}^{(m)} * \mathbf{x}, \quad (8)$$

where $\beta(\bar{K}) = \bar{K} - \lfloor \bar{K} \rfloor$ means the fractional part of \bar{K} , which is complementary to the interpolation ratio.

3.3. Sparsity

Since the kernel h_K is sparse, we can avoid computational redundancy by implementing the operation as a sum of scaled slices rather than as a typical convolution. To illustrate this point, the following is the discrete (non-differentiable) filter computed via this approach:

$$y[n] = x[n] + \sum_{t=1}^T \alpha^t x[n - tK]. \quad (9)$$

By applying the differentiable proxy (eq. (8)) to eq. (9), the full implementation of the comb filters used in training becomes:

$$y[n] = x[n] + \sum_{t=1}^T (1 - \beta(t\bar{K})) \alpha^t x[n - \lfloor t\bar{K} \rfloor] + \beta(t\bar{K}) \alpha^t x[n - \lceil t\bar{K} \rceil]. \quad (10)$$

We provide efficient Triton [20] kernels and torch forward/backward definitions for the differentiable comb filter as well as a fused combolution layer which is drastically more memory efficient.¹

3.4. Inference

During inference, gradient computation is unnecessary and can be removed. This also means that filter f_0 values can be discretized for a small performance penalty, and each output sample can be computed sequentially as in eq. (1). This is particularly ideal for models which must run in real time on-device, as many of these devices are already optimized for IIR filters.

4. EXPERIMENTS

We evaluate the performance of the comb filters on three tasks: monophonic note transcription, speaker classification [19], and acoustic musical key estimation. All models were trained on a single Nvidia L40S GPU using the Adam optimizer [21] and clipping gradients at a value of 0.5. For all comb filters, we use a fixed α value of 0.9 and a fixed T value of 10, i.e., we use 10 echoes in our truncated impulse response as written in eq. (9).

4.1. Note Transcription

As an illustrative example, we create a synthesized dataset of monophonic piano note sequences, each with a random number of notes between 3 and 10. The notes are sampled uniformly between C4 and B4, also with random durations and velocities². We train two architectures: ConvNet is composed of three 1D convolutional layers using ELU activations [22], while the proposed CombNet replaces ConvNet’s first layer with a combolutional layer. We train both of these architectures with 8, 16, 32, 64, and 128 channels in the first and second layer. The final layer always has 12 output channels, one for each of the 12 notes being estimated. All CombNet models use $f_{\min} = 200$ Hz, $f_{\max} = 500$ Hz for this experiment.

4.2. Speaker Classification

We base our speaker classification experiments closely off of the 462-speaker TIMIT [23] classification experiment outlined in [19]. We reimplement their SincNet and ConvNet models as baselines and compare them with a CombNet model, which is identical to SincNet except for the replacement of the first sinc layer with a combolutional layer. All models have a selected first layer, then 2 convolutional

¹<https://github.com/CameronChurchwell/combnet>

²Further details can be found in [combnet/data/synthesize/notes.py](https://github.com/CameronChurchwell/combnet/data/synthesize/notes.py)

Table 1: Note transcription results. MACs are computed per sample for the first layer only, as the remainder of the network takes multiple samples and process them as a batch, hence their MACs/sample are negligible.

	F1 \uparrow	MACs/sample \downarrow	Parameters \downarrow
CombNet ₁₂₈	0.95	167	18K
CombNet ₆₄	0.94	83	5K
CombNet ₃₂	0.93	42	1K
CombNet ₁₆	0.91	21	492
CombNet ₈	0.65	10	188
ConvNet ₁₂₈	0.95	7M	1M
ConvNet ₆₄	0.95	3M	569K
ConvNet ₃₂	0.92	1M	283K
ConvNet ₁₆	0.86	987K	141K
ConvNet ₈	0.75	493K	70K

layers followed by 3 fully-connected layers and a classifier head. We use the same normalization [24], [25], pooling, and activation [26] layers, but swap out the optimizer for Adam [21]. SincNet performs well on this task due to its well-aligned inductive bias which can learn the formant bands critical for speaker classification. We show that CombNet also has a powerful inductive bias and is much more efficient at inference time as can be seen in table 2. CombNet uses $f_{\min} = 50$ Hz, $f_{\max} = 8000$ Hz for this experiment.

4.3. Key Estimation

For musical key estimation, we use the GiantSteps-MTG dataset for training and validation, and the GiantSteps [27] dataset for testing. This task involves predicting one of 24 major and minor keys given a two-minute excerpt of a song. The baseline, CK [28] takes a 4097 bin magnitude spectrogram (i.e., the DFT size is 8192), filters it with 105 hand-crafted triangular filters into quartertone bins, and then applies a series of 2D convolutions followed by adaptive pooling and a classifier head. We also train three other baselines adapted from CK:

- 1) CK_{tri}: replaces the 105 triangular filters with a learned 4097×105 filter matrix.
- 2) CK_{learned}: Same as CK_{tri}, also learns a 4097×4097 replacement for the spectrogram computation.
- 3) CK_{chroma}: uses 12 chroma filters instead of 105 quartertone filters.

For our combolutional network, CombNet₆₄, we replace the spectrogram and filters with a single combolutional layer with 64 channels. Then, CombNet₆₄ mimics the hop size and DFT length of CK as its max pooling stride and window size, respectively. Given the harmonic-centric nature of the key detection problem, we expect that the 64 harmonic features learned by CombNet should perform comparably to the 105 filters used in CK. We use both accuracy and a weighted sum of prediction categories as proposed by the MIREX evaluation campaign³ and also used in [28]. Combnet uses $f_{\min} = 25.95$ Hz, $f_{\max} = 1046.5$ Hz to match the range of CK

5. RESULTS

5.1. Note Transcription

Table 1 shows the evaluation results for the note transcription task. It is clear from these results that CombNet is orders of magnitude more efficient than ConvNet in terms of both parameter count and inference-time multiply-accumulate operations. In particular, CombNet₁₂₈ and ConvNet₁₂₈ both achieve an F1 score of 0.95 but ConvNet₁₂₈ requires more than four orders of magnitude more MACs and has two orders of magnitude more parameters. In addition, Fig. 2 shows that across

³<http://www.music-ir.org/mirex>

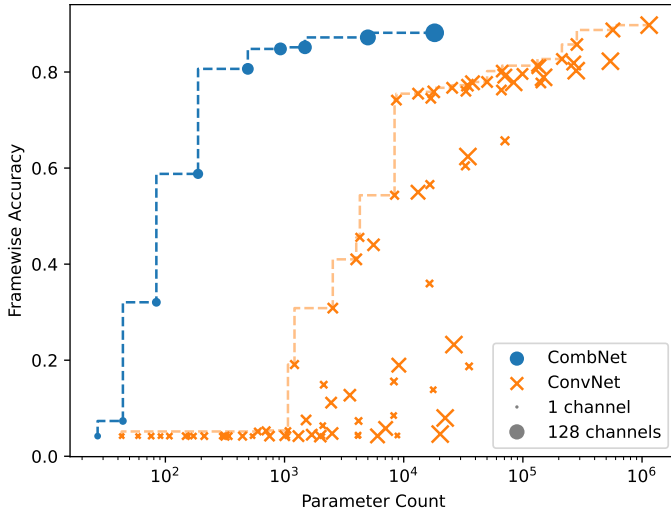


Fig. 2: Pareto fronts for CombNet and ConvNet evaluated on the piano transcription task.

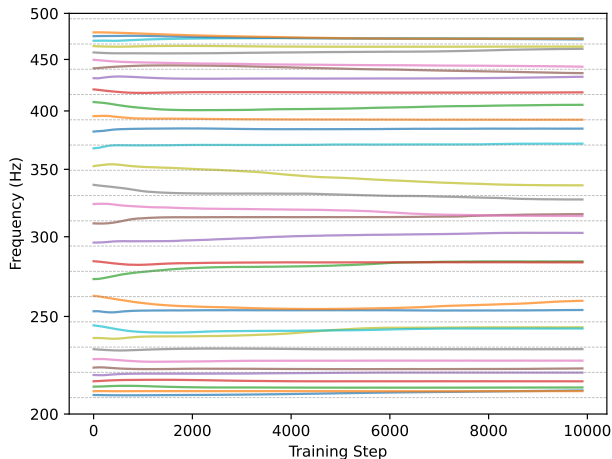


Fig. 3: Evolution of the combolational layer f_0 parameters of CombNet₃₂ during training on the note transcription task.

all of the models we trained, CombNet always achieved superior performance at any given model size.

Fig. 3 shows the process of learning f_0 parameters for the combolational layer when performing note transcription. We observe that multiple filter channels will often converge to the same f_0 value, perhaps indicating a particularly useful harmonic feature. CombNet also exhibits octave confusion as in other f_0 estimation systems. For example, two filters converge to ~ 243.5 Hz whose first harmonic at 487 Hz is close to B4 (493.88 Hz). Meanwhile, it could be said that CombNet is still efficient when octave confusion is not immediately problematic as in this toy transcription problem.

5.2. Speaker Classification

Table 2 shows the speaker classification evaluation results. CombNet₈₀ and SincNet₈₀ both outperform the ConvNet₈₀ baseline, with SincNet₈₀ performing better than CombNet₈₀ by 2.45 percentage points. However, SincNet₈₀’s slight increase in accuracy comes at the cost of 100 times as many inference MACs for the first layer alone. While each sinc layer only needs two parameters, one for each of the high and low cutoff frequencies of the learned bandpass filter, the computation still involves a convolution with a large FIR

Table 2: Speaker classification results from CombNet, SincNet, and ConvNet. SincNet and ConvNet results are only slightly different from the originally reported ones in [19] due to reimplementing. MACs are normalized by the number of input audio samples.

	Accuracy % \uparrow	MACs/sample \downarrow	
		First Layer	Total
CombNet ₈₀	96.18	80	17K
SincNet ₈₀ [19]	98.63	9K	25K
ConvNet ₈₀ [19]	94.52	18K	34K

Table 3: Key estimation results on the GiantSteps dataset.

	Accuracy % \uparrow	Weighted \uparrow	Parameters \downarrow
CombNet ₆₄	61.09	68.89	32K
CK	64.40	70.41	48K
CK _{tri}	63.25	69.40	478K
CK _{learned}	40.23	51.11	17M
CK _{chroma}	61.76	68.76	12K

filter kernel. The sinc kernels are symmetrical, and so halve the computational cost of an equivalent convolutional layer. However, we claim that CombNet’s IIR comb filter is equivalent to a length 1 kernel convolution, and is therefore more efficient. Speaker classification is a task which is frequently performed on-device, and therefore could benefit greatly from this drastic increase in efficiency, even if it comes with a slight performance penalty.

5.3. Key Estimation

The results for the key estimation task are shown in Table 3. CK, with its hand-crafted filterbanks, performs the best out of all the models trained. However, CombNet₆₄, CK_{tri}, and CK_{chroma} are all within 1.65 points weighted score of CK. We consider this to support the claim that the learned harmonic features of the combolational layer are equally as useful for machine learning tasks as more traditional DSP-style features. It is also worth noting that the reduction in channel count from 105 to the 64 in CombNet₆₄ reduces the parameter count by approximately one quarter, losing only to the highly-customized chroma feature-based CK_{chroma} in terms of parameter count.

The relatively strong performance of CK_{tri} and comparatively weak performance of CK_{learned} indicate that the choice of DSP features is not necessarily as important as the use of a time-frequency transform, as CK_{learned} seems to have failed to learn an equally useful representation despite its high parameter count. This reinforces the utility of the combolational layer as an efficient, trainable time-frequency transform.

6. CONCLUSION AND FUTURE WORK

In this work we presented combolational layers as efficient harmonic feature learners. When compared against convolutional frontends, the combolational layer requires significantly fewer parameters and is drastically cheaper to compute. We also explore efficient training implementations which enable gradient computation and avoid recurrence, which will be open-sourced upon acceptance. We demonstrated experimentally that combolational layers perform roughly equivalently to hand-crafted features, and outperformed even very large convolutional layers. We further showed that the combolational layer can be used on both speech and music, showcasing its broad applicability to harmonic-centric tasks. Our key estimation and speaker classification tasks were minimally different from the baseline sources, which indicates the combolational layer can be used as a drop-in replacement for other frontends. In terms of future work, we will further investigate the effect of the hyperparameter α and the potential merit of making it trainable.

REFERENCES

- [1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [2] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, vol. 1, no. 2, pp. 119–130, 1988.
- [3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
- [4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023.
- [7] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.
- [8] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.
- [9] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.
- [11] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Feature learning and deep architectures: new directions for music informatics," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 461–481, Dec. 2013.
- [12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.
- [14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW 9)*, 2016, p. 125.
- [15] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020.
- [16] Y. Wu, E. Manilow, Y. Deng, R. Swavely, K. Kastner, T. Cooijmans, A. Courville, C.-Z. A. Huang, and J. Engel, "MIDI-DDSP: Detailed control of musical performance via hierarchical modeling," in *International Conference on Learning Representations*, 2022.
- [17] B. Kuznetsov, J. Parker, and F. Esqueda, "Differentiable IIR filters for machine learning applications," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20)*, Vienna, Austria, 2020, pp. 297–303.
- [18] C.-Y. Yu, C. Mitcheltree, A. Carson, S. Bilbao, J. D. Reiss, and G. Fazezas, "Differentiable all-pole filters for time-varying audio systems," in *International Conference on Digital Audio Effects (DAFx)*, 2024.
- [19] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.
- [20] P. Tillet, H. T. Kung, and D. Cox, "Triton: an intermediate language and compiler for tiled neural network computations," in *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, ser. MAPL 2019, New York, NY, USA, 2019, p. 10–19.
- [21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." in *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.
- [22] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)." in *Proc. of the International Conference on Learning Representations (ICLR)*, 2016.
- [23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report*, 1993.
- [24] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.
- [25] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ser. ICML'15, 2015, p. 448–456.
- [26] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv:1505.00853*, 2015.
- [27] P. Knees, A. Faraldo, P. Herrera, R. Vogl, S. Bock, F. Horschlager, and M. L. Goff, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections." in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, October 2015.
- [28] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 966–970.