Dolby

Research at Dolby: my personal journey

Jordi Pons (@jordiponsdotme – www.jordipons.me)

Dolby

Source Separation

Music, speech, universal

Speech Enhancement / Synthesis

Packet loss concealment, quality metrics, generative

Dolby

Source Separation
Music, speech, universal

Speech Enhancement / Synthesis
Packet loss concealment, quality metrics, generative

Intelligent Music Production
Neural audio effects, automatic mixing

Dolby

Source Separation

Music, speech, universal

Speech Enhancement / Synthesis

Packet loss concealment, quality metrics, generative

Intelligent Music Production

Neural audio effects, automatic mixing

UPSAMPLING ARTIFACTS IN NEURAL AUDIO SYNTHESIS

Jordi Pons, Santiago Pascual, Giulio Cengarle, Joan Serrà

Dolby Laboratories

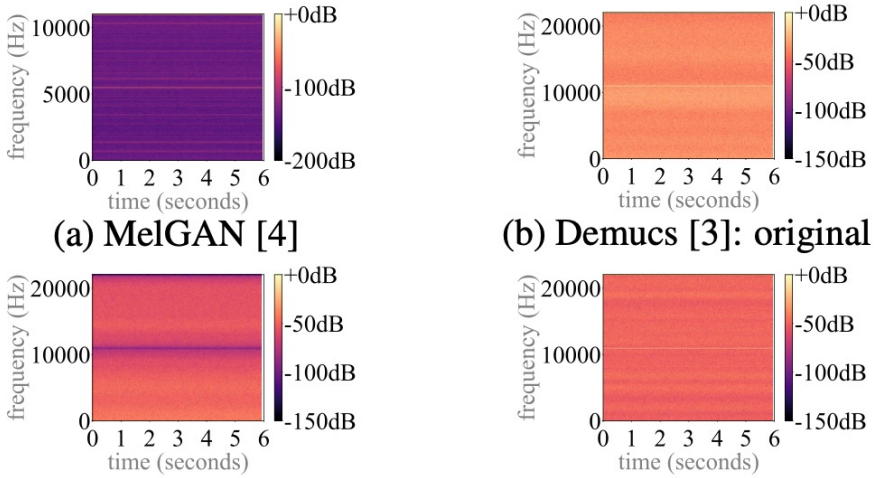
ABSTRACT

A number of recent advances in neural audio synthesis rely on upsampling layers, which can introduce undesired artifacts. In computer vision, upsampling artifacts have been studied and are known as checkerboard artifacts (due to their characteristic visual pattern). However, their effect has been overlooked so far in audio processing. Here, we address this gap by studying this problem from the audio signal processing perspective. We first show that the main sources of upsampling artifacts are: (i) the tonal and filtering artifacts introduced by problematic upsampling operators, and (ii) the spectral replicas that emerge while upsampling. We then compare different upsampling layers, showing that nearest neighbor upsamplers can be an alternative to the problematic (but state-of-the-art) transposed and subpixel convolutions which are prone to introduce tonal artifacts.

***Index Terms* —** upsampling, neural networks, synthesis, audio.

1. INTRODUCTION

Feed-forward neural audio synthesizers [1–4] were recently proposed as an alternative to Wavenet [5], which is computationally demanding and slow due to its dense and auto-regressive nature [6].



(a) MelGAN [4] (b) Demucs [3]: original

(c) Demucs: nearest neighbor (d) Demucs: subpixel CNN

Fig. 1. **Upsampling artifacts after initialization**: tonal artifacts (horizontal lines: a,b,d) and filtering artifacts (horizontal valley: c). Input: white noise. MelGAN operates at 22kHz, Demucs at 44kHz.

2. TRANSPOSED CONVOLUTIONS

Transposed CNNs are widely used for audio synthesis [1, 3, 9] and can introduce tonal artifacts due to [15]: (i) their weights' initialization, (ii) overlap issues, and (iii) the loss function. Issues (i) and (ii) are related to the model's initialization and construction, respectively,

Dolby

Source Separation

Music, speech, universal

Speech Enhancement / Synthesis

Packet loss concealment, quality metrics, generative

Intelligent Music Production

Neural audio effects, automatic mixing

AN EMPIRICAL STUDY OF CONV-TASNET

Berkan Kadıoğlu [‡ ⋆] *Michael Horgan* [† ⋆] *Xiaoyu Liu* [† ⋆] *Jordi Pons* [†] *Dan Darcy* [†] *Vivek Kumar* [†]

[‡] Electrical and Computer Engineering Department, Northeastern University
[†] Dolby Laboratories

ABSTRACT

Conv-TasNet is a recently proposed waveform-based deep neural network that achieves state-of-the-art performance in speech source separation. Its architecture consists of a learnable encoder/decoder and a separator that operates on top of this learned space. Various improvements have been proposed to Conv-TasNet. However, they mostly focus on the separator, leaving its encoder/decoder as a (shallow) linear operator. In this paper, we conduct an empirical study of Conv-TasNet and propose an enhancement to the encoder/decoder that is based on a (deep) non-linear variant of it. In addition, we experiment with the larger and more diverse LibriTTS dataset and investigate the generalization capabilities of the studied models when trained on a much larger dataset. We propose cross-dataset evaluation that includes assessing separations from the WSJ0-2mix, LibriTTS and VCTK databases. Our results show that enhancements to the encoder/decoder can improve average SI-SNR performance by more than 1 dB. Furthermore, we offer insights into the generalization capabilities of Conv-TasNet and the potential value of improvements to the encoder/decoder.

and in [15] a clustering mechanism is integrated into the separator. Interestingly, only a few works touch on the encoder/decoder of Conv-TasNet. In a multi-channel setting [16,17] a second encoder is used to learn phase differences between channels, and in [15] a magnitude STFT is appended to the learned encoder transform. As seen, most previous works use a (shallow) linear encoder/decoder. To the best of our knowledge, only [18] used a deep encoder/decoder for a Conv-TasNet inspired model for speech enhancement, which has not been extended to or fully tested for speech source separation.

In this work, we conduct an empirical study of Conv-TasNet, which is formally introduced in Section 2. Our contributions focus on two areas: architectural improvements to the encoder/decoder, and a study of the generalization capabilities of the developed models. In Section 3, we introduce the deep encoder/decoder we propose and we discuss several variants of this structure. In Section 4.1, we evaluate the studied models against the WSJ0-2mix database to gain insights on the performance of each variant. In Section 4.2, we explore the impact of using a larger, more diverse training set and we study the generalization capabilities of the trained models via employing a cross-dataset evaluation. In Section 4.3, we compare the

ON PERMUTATION INVARIANT TRAINING FOR SPEECH SOURCE SEPARATION

Xiaoyu Liu *Jordi Pons*

Dolby Laboratories

ABSTRACT

We study permutation invariant training (PIT), which targets at the permutation ambiguity problem for speaker independent source separation models. We extend two state-of-the-art PIT strategies. First, we look at the two-stage speaker separation and tracking algorithm based on frame level PIT (tPIT) and clustering, which was originally proposed for the STFT domain, and we adapt it to work with waveforms and over a learned latent space. Further, we propose an efficient clustering loss scalable to waveform models. Second, we extend a recently proposed auxiliary speaker-ID loss with a deep feature loss based on "problem agnostic speech features", to reduce the local permutation errors made by the utterance level PIT (uPIT). Our results show that the proposed extensions help reducing permutation ambiguity. However, we also note that the studied STFT-based models are more effective at reducing permutation errors than waveform-based models, a perspective overlooked in recent studies.

Index Terms— Speech source separation, permutation invariant training, waveform-based models, spectrogram-based models.

1. INTRODUCTION

Deep CASA, an spectrogram-based model, to Conv-TasNet, which uses very short waveform frames (such as 2 ms). We find that tPIT based on such short waveform frames can be challenging. Therefore, we propose performing tPIT in a pre-trained latent space—which allows for a more meaningful feature space for tPIT than the short waveform frames. Further, when training the clustering model, Deep CASA employs a memory and computationally expensive pairwise similarity loss that does not scale for waveform inputs. We propose a loss that reduces the complexity from quadratic to linear, making the training of the clustering model feasible for waveform models.

In section 3, we also extend the uPIT+speaker-ID loss with PASE, a problem agnostic speech encoder [14–16]. PASE is pre-trained in a self-supervised fashion with a collection of objectives much broader than speaker-ID, to extract general-purpose speech embeddings from waveforms. In addition, we also look at conditioning Conv-TasNet with PASE embeddings in a cascaded system consisting of two steps: (i) uPIT+PASE speaker separation, and (ii) conditioning Conv-TasNet with PASE embeddings computed from the speakers separated in step (i).

Conv-TasNet and Deep CASA can both be interpreted as architectures with an encoder/decoder and a separator, where the en-

Dolby

Source Separation

Music, speech, universal

Speech Enhancement / Synthesis

Packet loss concealment, quality metrics, generative

Intelligent Music Production

Neural audio effects, automatic mixing

MULTICHANNEL-BASED LEARNING FOR AUDIO OBJECT EXTRACTION

Daniel Arteaga *Jordi Pons*

Dolby Laboratories

ABSTRACT

The current paradigm for creating and deploying immersive audio content is based on audio objects, which are composed of an audio track and position metadata. While rendering an object-based production into a multichannel mix is straightforward, the reverse process involves sound source separation and estimating the spatial trajectories of the extracted sources. Besides, cinematic object-based productions are often composed by dozens of simultaneous audio objects, which poses a scalability challenge for audio object extraction. Here, we propose a novel deep learning approach to object extraction that learns from the multichannel renders of object-based productions, instead of directly learning from the audio objects themselves. This approach allows tackling the object scalability challenge and also offers the possibility to formulate the problem in a supervised or an unsupervised fashion. Since, to our knowledge, no other works have previously addressed this topic, we first define the task and propose an evaluation methodology, and then discuss under what circumstances our methods outperform the proposed baselines.

Index Terms— object-based audio, source separation

Fig. 1. Multichannel-based learning: supervised and unsupervised. For supervised learning, a set of multichannel mixes, rendered from a reference object-based production, are compared. In contrast, unsupervised learning only relies on the 5.1 mixes and does not require a reference object-based production.

source separation literature, also arises here. The output to ground truth pairs required for supervised learning can not be arbitrarily assigned due to the source- (or speaker-) independent nature of the task. Note that in cinematic audio productions, the number of po-

Dolby

Source Separation
Music, speech, universal

Speech Enhancement / Synthesis
Packet loss concealment, quality metrics, generative

_

Intelligent Music Production
Neural audio effects, automatic mixing

ADVERSARIAL AUTO-ENCODING FOR PACKET LOSS CONCEALMENT

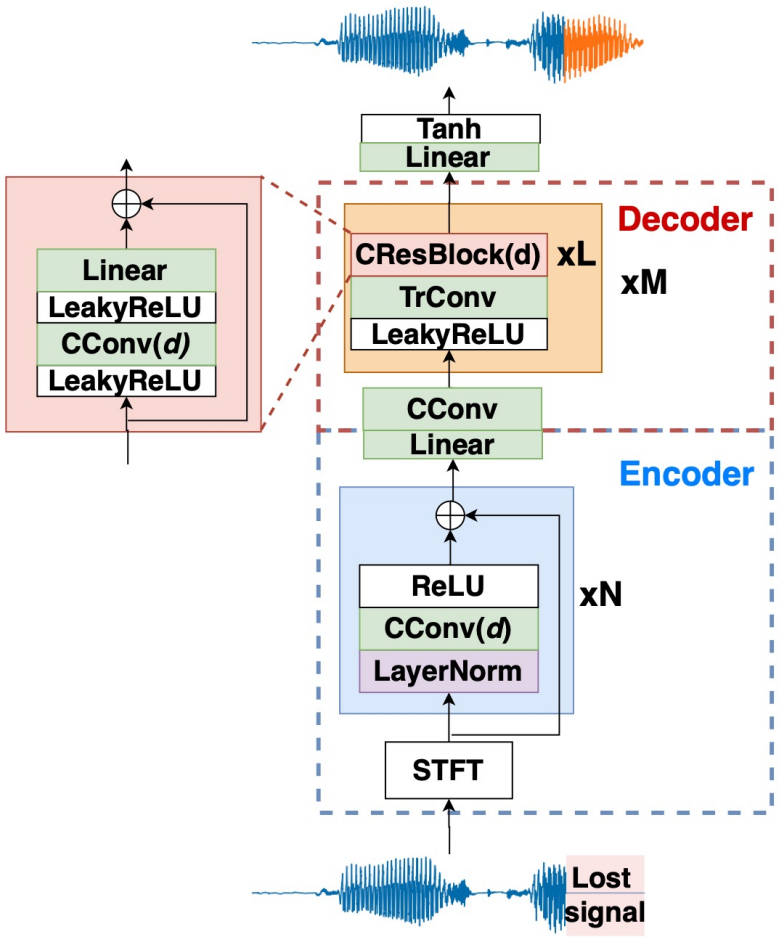
Santiago Pascual, Joan Serrà, Jordi Pons

Dolby Laboratories
santiago.pascual@dolby.com

ABSTRACT

Communication technologies like voice over IP operate under constrained real-time conditions, with voice packets being subject to delays and losses from the network. In such cases, the packet loss concealment (PLC) algorithm reconstructs missing frames until a new real packet is received. Recently, autoregressive deep neural networks have been shown to surpass the quality of signal processing methods for PLC, specially for long-term predictions beyond 60 ms. In this work, we propose a non-autoregressive adversarial auto-encoder, named PLAAE, to perform real-time PLC in the waveform domain. PLAAE has a causal convolutional structure, and it learns in an auto-encoder fashion to reconstruct signals with gaps, with the help of an adversarial loss. During inference, it is able to predict smooth and coherent continuations of such gaps in a single feed-forward step, as opposed to autoregressive models. Our evaluation highlights the superiority of PLAAE over two classic PLCs and two deep autoregressive models in terms of spectral and intonation reconstruction, perceptual quality, and intelligibility.

Index Terms— packet loss concealment, adversarial networks, auto-encoder, speech enhancement



Dolby

Source Separation
Music, speech, universal

Speech Enhancement / Synthesis
Packet loss concealment, quality metrics, generative

Intelligent Music Production
Neural audio effects, automatic mixing

SESQA: SEMI-SUPERVISED LEARNING FOR SPEECH QUALITY ASSESSMENT

Joan Serrà, Jordi Pons, Santiago Pascual

Dolby Laboratories

ABSTRACT

Automatic speech quality assessment is an important, transversal task whose progress is hampered by the scarcity of human annotations, poor generalization to unseen recording conditions, and a lack of flexibility of existing approaches. In this work, we tackle these problems with a semi-supervised learning approach, combining available annotations with programmatically generated data, and using 3 different optimization criteria together with 5 complementary auxiliary tasks. Our results show that such a semi-supervised approach can cut the error of existing methods by more than 36%, while providing additional benefits in terms of reusable features or auxiliary outputs. Improvement is further corroborated with an out-of-sample test showing promising generalization capabilities.

Index Terms— Speech quality, semi-supervised learning, multi-objective, neural encoders, raw audio.

Learning-based systems, on the other hand, are usually easy to repurpose to other tasks and degradations, but require considerable amounts of human annotated data. Both rule- and learning-based systems might additionally suffer from lack of generalization, and thus perform poorly on out-of-sample but still on-focus data.

Semi-supervised learning is a possible strategy to deal with lack of annotations and poor generalization [17]. By leveraging both labeled and unlabeled data, it can bring substantial performance and generalization improvements, specially when annotations are scarce. Semi-supervised learning is behind many recent advancements in machine learning [18], and has been successfully applied to image quality assessment [19, 20]. However, surprisingly, we find no purely semi-supervised approaches for audio or speech quality assessment. To the best of our knowledge, the few works that exploit unlabeled data for this task only make indirect use of it: they either exploit the output of other existing (rule-based)

Dolby

Source Separation
Music, speech, universal

Speech Enhancement / Synthesis
Packet loss concealment, quality metrics, generative

Intelligent Music Production
Neural audio effects, automatic mixing

ON TUNING CONSISTENT ANNEALED SAMPLING FOR DENOISING SCORE MATCHING

Joan Serrà, Santiago Pascual, Jordi Pons

Dolby Laboratories

ABSTRACT

Score-based generative models provide state-of-the-art quality for image and audio synthesis. Sampling from these models is performed iteratively, typically employing a discretized series of noise levels and a predefined scheme. In this note, we first overview three common sampling schemes for models trained with denoising score matching. Next, we focus on one of them, consistent annealed sampling, and study its hyper-parameter boundaries. We then highlight a possible formulation of such hyper-parameter that explicitly considers those boundaries and facilitates tuning when using few or a variable number of steps. Finally, we highlight some connections of the formulation with other sampling schemes.

***Keywords*—** Score-based generative models, denoising score matching, sampling, Langevin.

INTRODUCTION

Score matching [1] has become a successful approach to

In recent work, Jolicoeur-Martineau et al. [10] point out some inconsistencies for ALS in the context of score-based generative models. In particular, given that $N \ll \infty$ and

Dolby

Source Separation
Music, speech, universal

Speech Enhancement / Synthesis
Packet loss concealment, quality metrics, generative

_

Intelligent Music Production
Neural audio effects, automatic mixing

AUTOMATIC MULTITRACK MIXING WITH A DIFFERENTIABLE MIXING CONSOLE OF NEURAL AUDIO EFFECTS

Christian J. Steinmetz [1,2,⋆] *Jordi Pons* [1] *Santiago Pascual* [1] *Joan Serrà* [1]

[1] Dolby Laboratories
[2] Music Technology Group, Universitat Pompeu Fabra

ABSTRACT

Applications of deep learning to automatic multitrack mixing are largely unexplored. This is partly due to the limited available data, coupled with the fact that such data is relatively unstructured and variable. To address these challenges, we propose a domain-inspired model with a strong inductive bias for the mixing task. We achieve this with the application of pre-trained sub-networks and weight sharing, as well as with a sum/difference stereo loss function. The proposed model can be trained with a limited number of examples, is permutation invariant with respect to the input ordering, and places no limit on the number of input sources. Furthermore, it produces human-readable mixing parameters, allowing users to manually adjust or refine the generated mix. Results from a perceptual evaluation involving audio engineers indicate that our approach generates mixes that outperform baseline approaches. To the best of our knowledge, this work demonstrates the first approach in learning multitrack mixing conventions from real-world data at the waveform level, without knowledge of the underlying mixing parameters.

IMP systems generally implement either rule-based or classical machine learning approaches [4]. Rule-based approaches rely upon establishing a set of rules and logic surrounding best practices [5–7]. While they generate convincing results for some cases, they do not provide a level of expressivity that matches human audio engineers [8]. In comparison, classical machine learning approaches allow for greater model flexibility, but have typically suffered from the lack of parametric mixing data (i.e., the exact settings of each processor in the mix). For this reason, they have been of low-complexity, limiting their practical application [9–11]. While both approaches have seen some success in addressing particular aspects of the mixing process, they ultimately have failed to capture the entire process and generalize at the scale of real-world projects.

The previous shortcomings, along with the promise of deep learning methods in multiple audio signal processing tasks, motivate the application of those within IMP. Nevertheless, there are a number of unique challenges in the application of deep learning methodologies to automatic mixing that have yet to be addressed:

Dolby

Interested in interning at Dolby?
Contact us.. ☺

Source Separation
Music, speech, universal

Speech Enhancement / Synthesis
Packet loss concealment, quality metrics, generative

_

Intelligent Music Production
Neural audio effects, automatic mixing

Dolby

Research at Dolby: my personal journey

_

Jordi Pons (@jordiponsdotme – www.jordipons.me)