Music & Voice translation

Julia Gusak
Skoltech

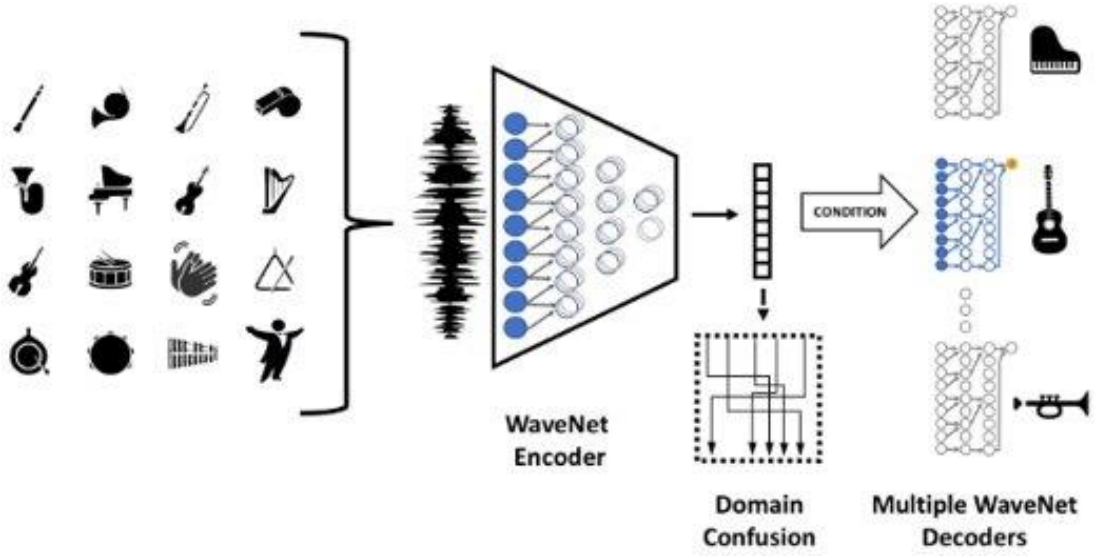
y.gusak@skoltech.ru

A Universal Music Translation Network (Mor et al., 2018)

Proposed method

- Translate music across musical instruments, genres, styles
  - Can convert from musical domains that were not heard during training to any of the domain encountered
- Based on **multi-domain wavenet autoencoder** (several decoders, a shared encoder and a disentangled latent space)
- Autoencoder is trained **end-to-end on waveforms**
  - **Domain confusion network** is used to train single encoder (to make sure the domain spesific info is not included)
  - **Input distortion by random local pitch modulation** (to make sure encoder dosn't memorize input signal)
- **Unsupervised** (no matched samples between domains or musical transcriptions)
  - During training, the network is trained as a denoising autoencoder, which recovers the undistorted version of the original input.

https://www.youtube.com/watch?v=vdxCqNWTpUs (0:46-2:10s, 3:42 – 5:00s)



A Universal Music Translation Network (Mor et al., 2018)

Domain transfer

- Images and text (many papers)
  - unsupervised translation between domains A and B (without being shown any matching pairs);
  - employs GANs;
  - circularity constraint (reconstraction of original image is obtained by mapping from A to B and back to A).
- Music (Mork et al., 2018)
  - Output is generated by an autoregressive model
  - **Training takes place using the ground truth output of the previous time steps** ("teacher forcing"), instead of the predicted ones
  - A **complete autoregressive inference is only done during test time**
  - Circulanty constraint is unrealistic**.** (Because output during training does not represent the future test time output)

A Universal Music Translation Network (Mor et al., 2018)

Domain transfer: Cross domain translation

- StarGAN (Choi et al, 2017)
  - **single generator** that receives as input the source image and the specification of the target domain,
  - **produces "fake" image from the target domain**
- (Mor et al, 2018)
  - **Multiple decoders, one per domain**
  - Experiments with conditioning a single decoder on the selection of the output domain failed

A Universal Music Translation Network (Mor et al., 2018)

Domain transfer: Shared latent space

- CoGAN (Liu et al, 2016)
  - **the earlier generator layers are shared,** while the top layers are domain-specific
  - given a sample $x \in A$, a latent vector z_x is fitted to minimize the distance between the image generated by the first generator $G_A(z_x)$ and the input image x. Then, the analogous image in B is given be $G_B(z_x)$
- UNIT(Liu et al., 2017)
  - encoder-decoder pair per each domain
  - the layers that are distant from the image **(the top layers of the encoder and the bottom layers of the decoder) are the ones shared**
  - Variational autoencoder to structure to the latent space
- (Mor et al., 2018)
  - **Single encoder**
  - No VAE loss term on the latent space, **domain confusion loss** is used instead

A Universal Music Translation Network (Mor et al., 2018)

Audio synthesis

- **WaveNet** (Van den OOrd et al., 2016) is an autoregressive model that predicts the probability distribution of the next sample, given the previous samples and an input conditioning signal.
- (Rethage et al., 2017) used **WaveNet for denoising waveforms** by predicting the middle ground-truth sample from its noisy input signal
- DeepVoice3 (Ping et al.), Tacotron2 (Shen et al., 2018) used **WaveNet conditioned on linguistic and acoustic features** to obtain **state of the art synthesis performance**.
- (Van den OOrd et al., 2017) proposed **voice conversion using VAE (decoder based on WaveNet)** that produces a quantized latent space that is conditioned on the speaker identity.
- (Mor et al., 2018) **WaveNet autoencoder as in (Engel et al., 2017), but with multiple decoders and an additional auxiliary network used for disentanglement**

A Universal Music Translation Network (Mor et al., 2018)

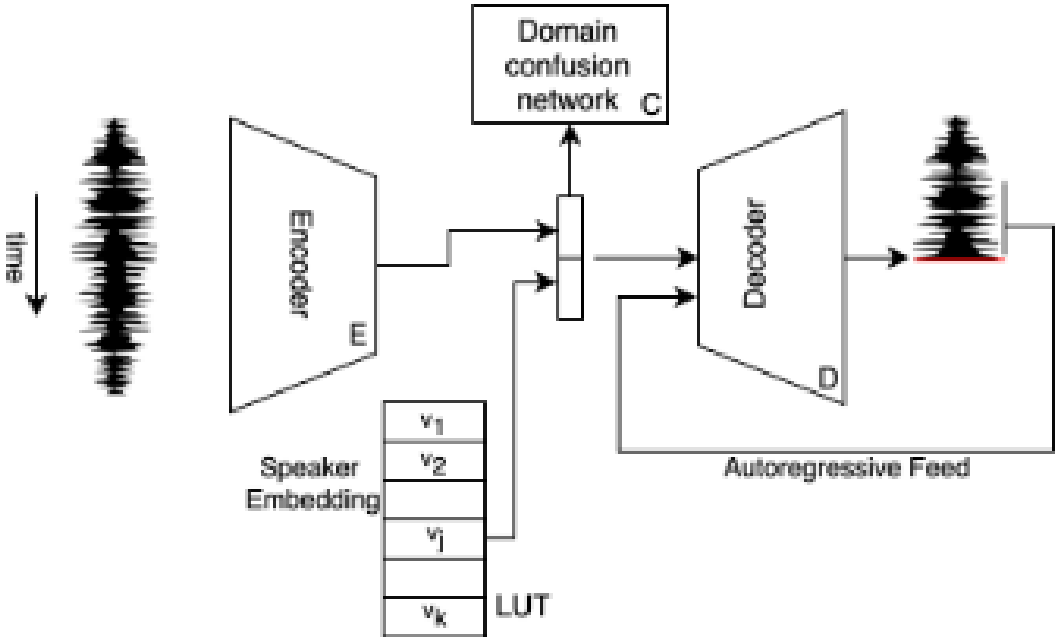
Resume

- Train multiple autoencoder pathways, one decoder ***Dj(…)*** per musical domain ***j***, such that encoders ***E(…)*** are shared.
- During training, a softmax-based reconstruction loss is applied to each domain ***j*** separately
- The input data ***s*** is randomly augmented (random seed ***r)*** prior to applying the encoder (tool against data memorizing), ***O(s, r)***
- A domain confusion loss ***C(…),*** (Ganin et al., 2016), is applied to the latent space to ensure that the encoding is not domain-specific

$$\sum_j \sum_{s^j} \mathop{\mathbb{E}}_r \mathcal{L}(D^j(E(O(s^j, r))), s^j) - \lambda \mathcal{L}(C(E(O(s^j, r))), j)$$

Unsupervised Singing Voice Conversion (Nachmani & Wolf, 2019)

- Proposed network directly **converts the audio of one singer (5-30 mins) to the voice of another**,
- **NOT conditioned** on the text or on the notes.
- **Unsupervised training** (no lyrics, no phonetic features, no notes, no matching samples between singers,... no supervision in any form).
- **Single CNN encoder** for all singers, **Single WaveNet decoder,** and **Classifier** (enforces the latent representations to be **singer-agnostic)**
- Each singer is represented by one embedding vector, which the decoder is conditioned on
- New data augmentation scheme
- New traing losses and protocols based on backtranslation (backtranslation + mixup)



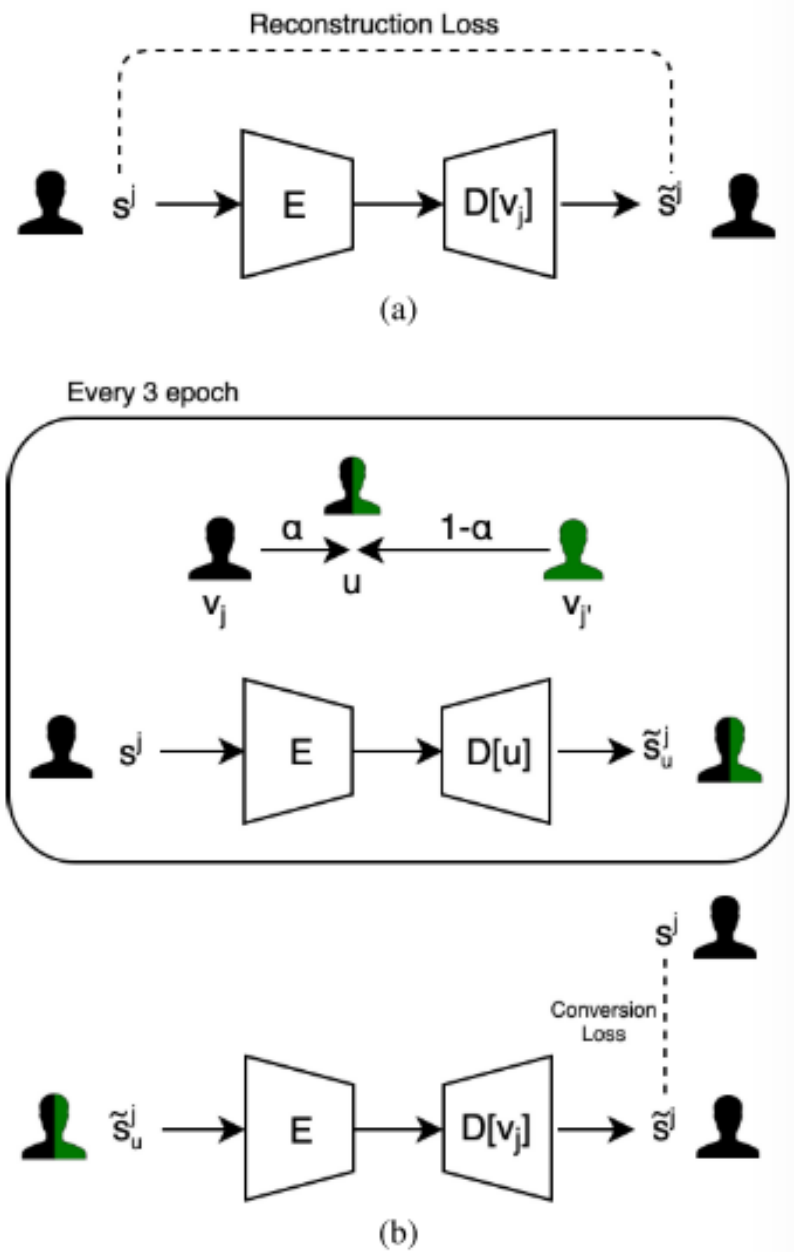
Unsupervised Singing Voice Conversion (Nachmani & Wolf, 2019)

Two phases of training

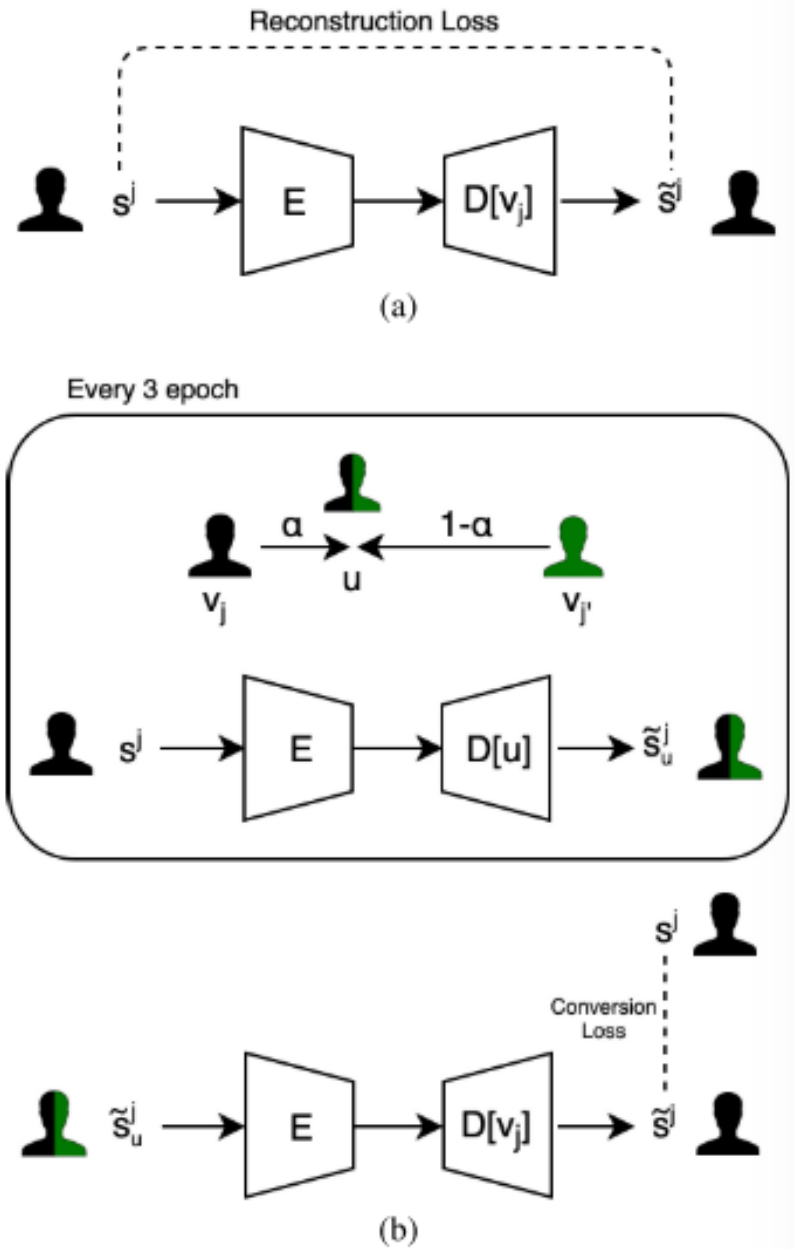
- The confusion network ***C*** minimizes the classification loss, ***j=1..k*** autoencoders are trained with the loss

$$\sum_j \sum_{s^j} \mathcal{L}(D[v_j](E(s^j)), s^j) - \lambda \sum_j \sum_{s^j} \mathcal{L}(C(E(s^j)), j)$$

- The network that is trained with such loss is able to reconstruct the original signal, and its encoder produces an embedding that is (somewhat) singer-agnostic. However, it is not trained directly to perform a singer translation.



Unsupervised Singing Voice Conversion (Nachmani & Wolf, 2019)



Two phases of training

- In the secon phase, backtranslation is applied in order to create parallel samples and train the network on these samples. This is done in combination with the mixup technique in order to generate “in-between” singers that are easier to fit.
- Once the second phase of training starts, we create new mixup samples every three epochs and use them in order to add the following loss to the training of D and E
- $$\sum_{s^j_u} \mathcal{L}(D[v_j](E(s^j_u)), s^j) \quad u = \alpha v_j + (1-\alpha)v_{j'} \quad s^j_u = D[u](E(s^j)$$

Unsupervised Singing Voice Conversion (Nachmani & Wolf, 2019)

Backtranslation

In NLP*:*

A sample ***a*** in language ***A***, which does not have a matching translation in the target language ***B***, is automatically translated by the current AMT(automatic translation systems) system to a sample ***b*** in that language. One then considers the training pair ***(b,a)*** for translating from the language ***B*** back to language ***A***. Since

References

Covered in presentation

- Mor, N., Wolf, L., Polyak, A., & Taigman, Y. (2018). A universal music translation network. *arXiv preprint arXiv:1805.07848*.
- Nachmani, E., & Wolf, L. (2019). Unsupervised Singing Voice Conversion. *arXiv preprint arXiv:1904.06590*.

Worth read (most of papers are authored by Lior Wolf https://www.cs.tau.ac.il/~wolf/)

- Polyak, A., Wolf, L., & Taigman, Y. (2019). TTS Skins: Speaker Conversion via ASR. *arXiv preprint arXiv:1904.08983*.
  - **Conversion between speaker voices without relying on text** (wav-to-wav network)
  - Encoder-decoder architecture; encoder is pretrained for the task of Automatic Speech Recognition, multi-speaker waveform decoder is trained to reconstruct the original signal in an autoregressive manner
  - This approach separates target voice generation from TTS module, enables client-side personalized TTS in a privacy-aware manner

- Nachmani, E., & Wolf, L. (2019, May). Unsupervised Polyglot Text-to-speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7055-7059). IEEE.
  - **Able to transfer voice, which was presented as a sample in a source language, into one of several target languages**;
  - No parallel data for training;
  - conversion based on learning polyglot network, which has multiple per-language sub-networks and adding loss terms that preserves the speaker's identity in multiple languages

References

- Nachmani, E., Polyak, A., Taigman, Y., & Wolf, L. (2018). Fitting new speakers based on a short untranscribed sample. *arXiv preprint arXiv:1802.06984*.
  - **A method that is designed to capture a new speaker from a short untranscribed audio sample.**
  - This is done by employing an additional network that given an audio sample, places the speaker in the em- bedding space. This network is trained as part of the speech synthesis system using various consistency losses
- Michelashvili, M., & Wolf, L. (2019). Audio Denoising with Deep Network Priors. *arXiv preprint arXiv:1904.07612*.
  - **Deep Image Prior is not suitable for audio denoising as it is!**
  - Given a noisy signal, the method trains a deep neural network to fit this signal, WaveUnet is used (CNN encoder-decoder architecture with skip-connections between the two subnetworks)
  - **Estimate a-priori SNR of the clean signal, apply classical denoising method using this estimation**
- Michelashvili, M., Benaim, S., & Wolf, L. (2019, May). Semi-supervised Monaural Singing Voice Separation with a Masking Network Trained on Synthetic Mixtures. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 291-295). IEEE.
  - The problem of semi-supervised singing voice separation, in which the training data contains a set of samples of mixed music (singing and instrumental) and an unmatched set of instrumental music
- Taigman, Y., Wolf, L., Polyak, A., & Nachmani, E. (2017). Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*.
  - **New neural text to speech (TTS) method that is able to transform text to speech in voices that are sampled in the wild**
  - Solution is able to deal with unconstrained voice samples and without requiring aligned phonemes or linguistic features.
  - The network architecture is simpler than those in the existing literature and is based on a novel shifting buffer working memory.