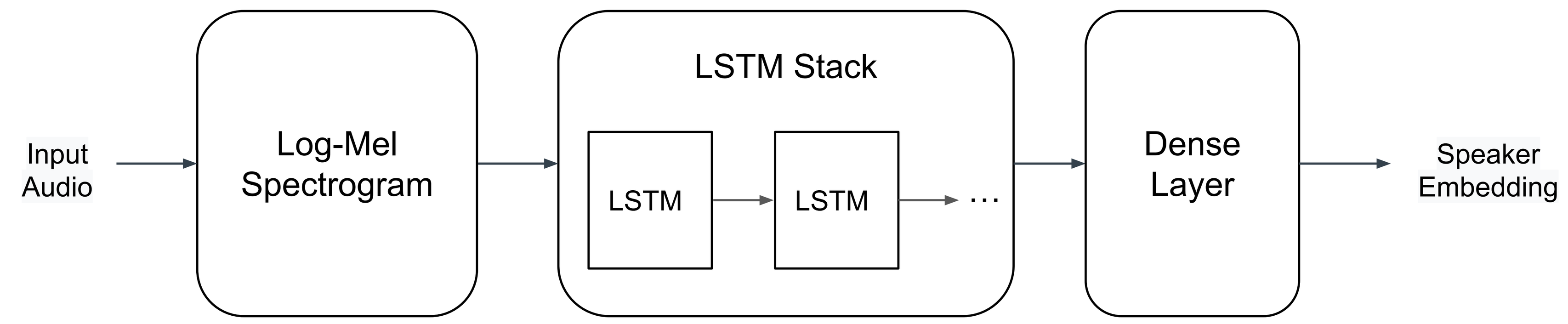
Zero-shot singing voice conversion

Shahan Nercessian
iZotope, Inc.
shahan@izotope.com

Overview

- Singing voice conversion (SVC) is the transformation of a sung vocal performance from one vocalist's identity to another.
- It is a challenging problem because models need to be able to disentangle vocalist identity from acoustic features, while preserving pitch and phonetic content in the output.
- Recently, speaker embedding networks were found to be successful for enabling zero-shot voice conversion of speech [1], whereby the system can model and adapt to new unseen voices on the fly.
- In this paper, we adapt zero-shot voice conversion methodologies for SVC using the WORLD [2] vocoder.

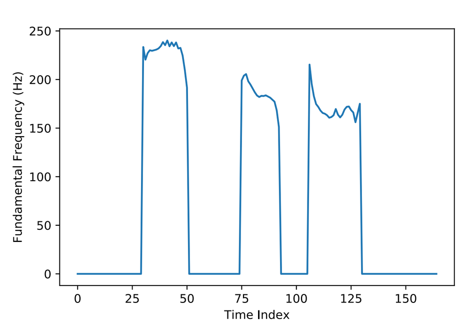
Speaker embedding network



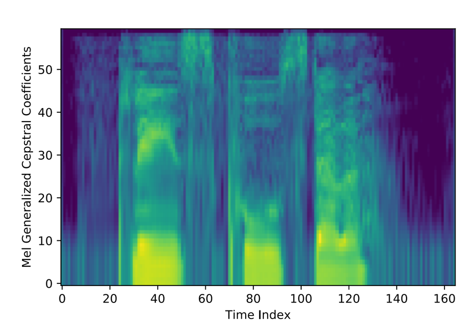
- Speaker embedding network maps input log-Mel spectrograms into 256-dimensional embeddings using a stack of LSTMs.
- This network is trained over large speech corpora to minimize the Generalized End-to-End Loss [3].
- Embeddings from vocal performances from the same performer form clusters in the embedding space.
- We can replace one-hot encodings of voices with speaker embeddings.
  - During inference, new voices can be characterized by their speaker embeddings.
  - So long as the model has been exposed to several voices during training, the model will be able to adapt to a new voice by feeding the new speaker embedding as input to the model.
- We use the pretrained speaker embedding network found at https://github.com/CorentinJ/Real-Time-Voice-Cloning.
- The speaker embedding network is frozen during training of the SVC network.

WORLD vocoder

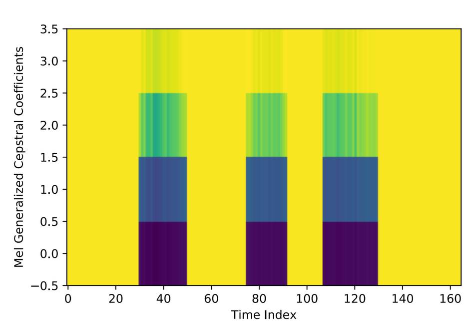
- Estimates 3 sets of parameters for (speech) synthesis
  - Fundamental frequency (f_0)
  - Periodic spectral envelope
  - Aperiodic spectral envelope, defined as a ratio relative to the periodic spectral envelope
- Vocoder representation can be compressed using Mel Generalized Cepstral Coefficients [4]
- Advantages
  - Boasts real-time synthesis
  - Synthesized output guaranteed to match detected pitch
- Main disadvantage is lack of expressivity relative to a neural vocoder



Fundamental Frequency

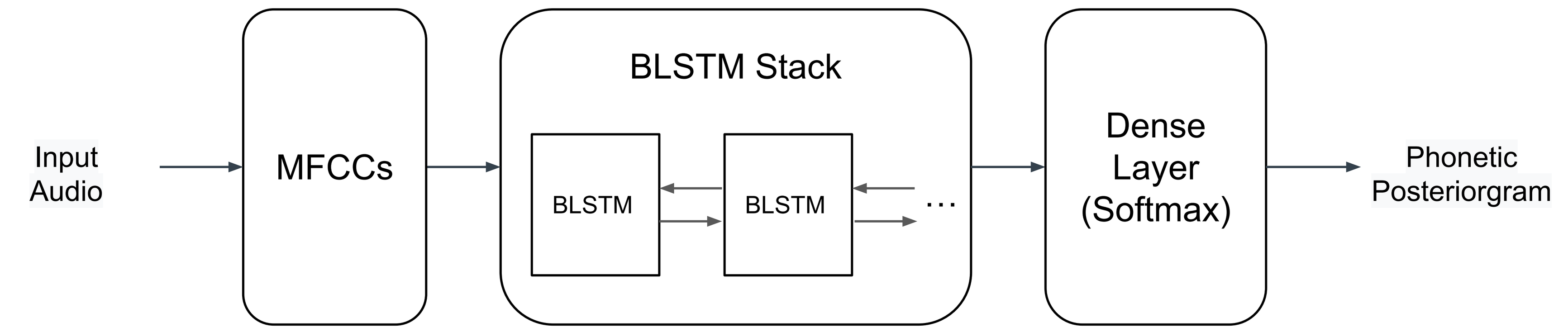


Periodic Spectral Envelope



Aperiodic Spectral Envelope

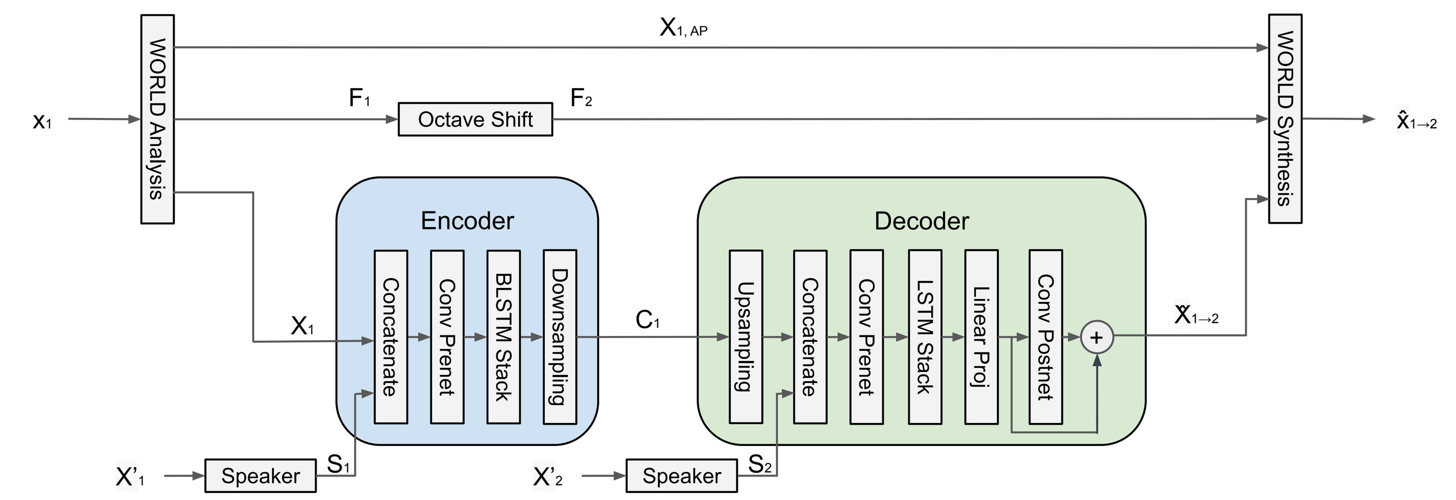
Linguistic content and loudness



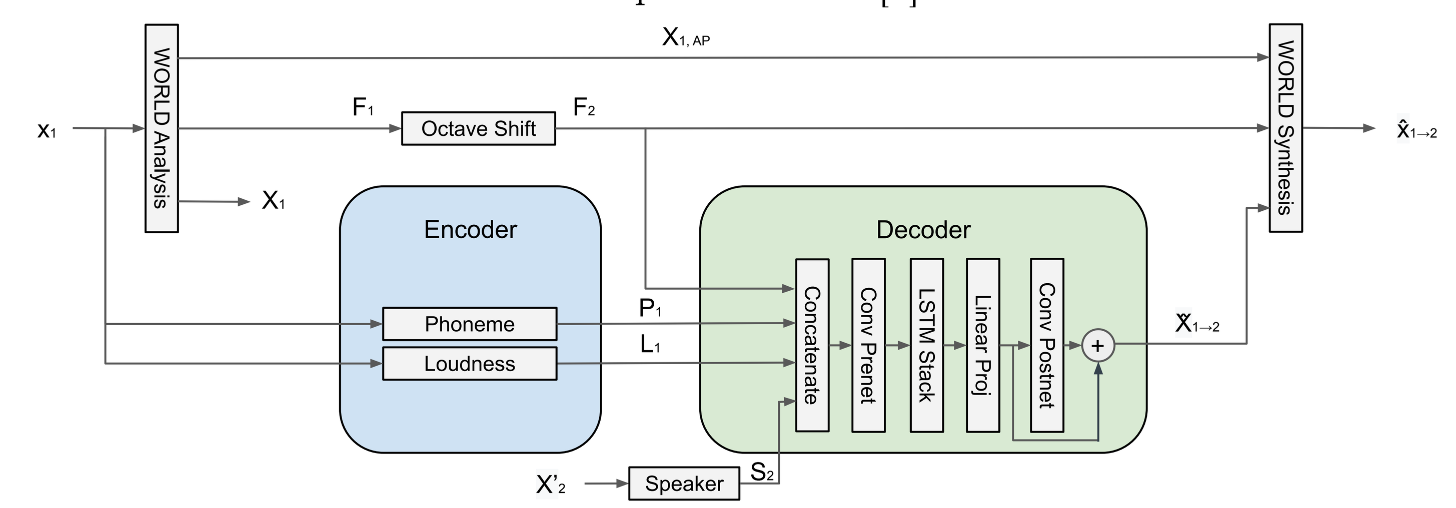
Phoneme classification network

- As an alternative to an unsupervised latent representation, linguistic content and loudness can be represented explicitly.
- Linguistic content modeled with phonetic posteriorgrams
  - BLSTM classification network trained on a multi-speaker dataset (TIMIT [5]) operating on MFCCs.
  - Phoneme classification accuracy does not need to be terribly accurate: 65% that we achieve is sufficient to act as a reasonable speaker-independent representation.
  - This model is frozen during training of the SVC network.
- Dynamics modeled using a deterministic A-weighted loudness computation [6].

Neural network architectures



Adapted AutoVC [1]



Fixed encoder (FE) model

- Adapted AutoVC learns a speaker-invariant latent representation during SVC model training.
- FE model encodes audio using loudness computation, WORLD-derived f_0, and pretrained phoneme classifier. Only the decoder is trained during SVC model training.
- Conversion models trained as conditional autoencoders to reconstruct periodic spectral envelopes.
- Octave shifts computed automatically to accommodate register differences between source and target.

Universal Background Model (UBM)

- The use of a pretrained speaker embedding network trained once in a supervised fashion allows for unsupervised zero-shot SVC system training.
- This not only eases data engineering, but allows us to train an initial network on large amounts of speech, and perform fine tuning using limited singing voice datasets.
- Network trained on large speech corpora acts as a "UBM" akin to those used in speaker recognition [7].

Experimental results

- Models are trained using VCTK [8] and an internally sourced unlabeled singing voice dataset, which we simply call the SVC dataset.
- To illustrate the effectiveness of our methods and insights, we consider 4 training configurations, trained identically with ADAM optimizer, learning rate of 10^{-3}, and batch size of 2.
  - VCTK using one-hot encoding of speakers (500K steps).
  - VCTK using speaker embeddings (500K steps).
  - SVC using speaker embeddings (500K steps).
  - VCTK UBM (350K steps) and model fine tuning with SVC (150K steps) using speaker embeddings.
- We evaluate model performance quantitatively by reporting validation loss, and qualitatively with subjective testing, reporting mean opinion scores (MOS).
- Overall, the best approach when evaluated on singing voice is the proposed UBM/SVC adaptation strategy.

Reconstruction loss on VCTK (left) and SVC (right) test sets

Training Configuration	AutoVC	FE
VCTK (one-hot)	0.1837	**0.1882**
VCTK (zero-shot)	**0.1634**	0.1891
SVC (zero-shot)	0.2930	0.3590
VCTK→SVC (zero-shot)	0.2557	0.3232

Training Configuration	AutoVC	FE
VCTK (one-hot)	N/A	N/A
VCTK (zero-shot)	0.3007	0.4314
SVC (zero-shot)	0.1650	0.1959
VCTK→SVC (zero-shot)	**0.1439**	**0.1850**

MOS on singing voice with FE model, target voices from the VCTK (left) and SVC (right) test sets

Training Configuration	Quality	Similarity
VCTK (one-hot)	2.377	2.828
VCTK (zero-shot)	2.447	**3.051**
SVC (zero-shot)	2.289	2.549
VCTK→SVC (zero-shot)	**2.476**	2.664

Training Configuration	Quality	Similarity
VCTK (one-hot)	N/A	N/A
VCTK (zero-shot)	2.154	2.610
SVC (zero-shot)	2.477	2.772
VCTK→SVC (zero-shot)	**2.674**	**2.937**

Conclusions

- Speaker embedding networks can indeed be extended to enable zero-shot SVC.
- An advantage of our SVC system is that it can be trained on unlabeled data.
- This enables pretraining of a UBM on speech, followed by adaptation to singing voice, which yields improved performance.
- Future work will consider end-to-end training using a differentiable parametric vocoder.

References

[1] K. Qian et al., "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in Proc. of the International Conference on Machine Learning, 2019.
[2] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE Transactions on Information and Systems, vol. E99-D, no. 7, pp. 1877–1884, 2016.
[3] L. Wan et al.,"Generalized end-to-end loss for speaker verification," in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 4879–4883.
[4] K. Tokuda et al., "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in Proc. of the International Conference on Spoken Language Processing, 1994.
[5] J. S. Garapolo et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. guistic Data Consortium, 1993.
[6] L. Hantrakul et al., "Fast and flexible neural audio synthesis," in Proc. of the International Society for Music Information Retrieval Conference, 2019.
[7] T. Hasan and J. H. L. Hansen, "A study on universal background model training in speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 1890–1899, 2011.
[8] C. Veaux et al., CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. Edinburgh: The Centre for Speech Technology Research (CSTR), University of Edinburgh, 2016.

Audio demo

Please visit our audio demo at https://sites.google.com/izotope.com/ismir2020-audio-demo.