Google Research

Scaling Polyphonic Transcription with Mixtures of Monophonic Transcriptions

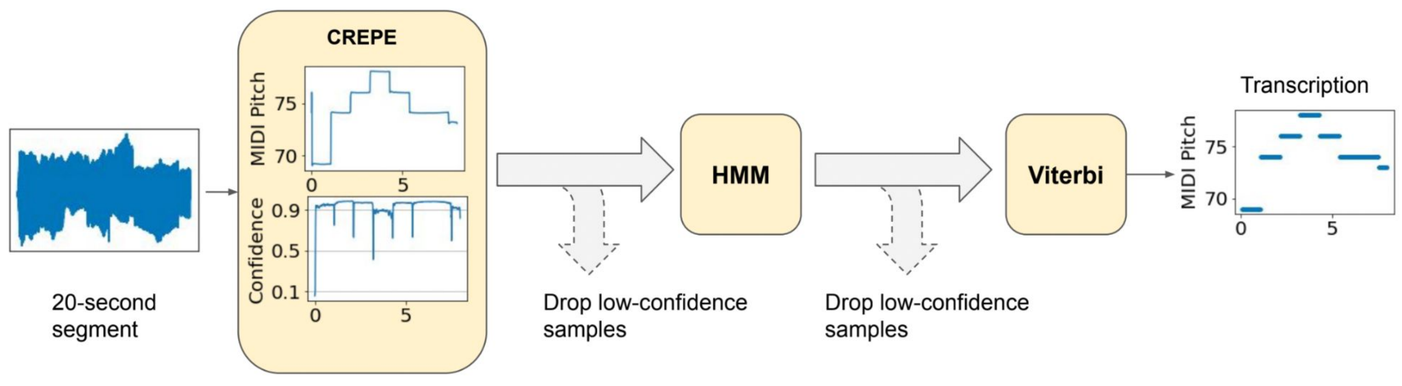
Ian Simon, Josh Gardner, Curtis Hawthorne, Ethan Manilow, Jesse Engel
{iansimon, jpgard, fjord, emanilow, jesseengel}@google.com

Overview

Existing datasets for automatic music transcription lack diverse audio. We train a model on a greater variety of audio using the following process:

1. Find monophonic music in the wild.
2. Transcribe it with CREPE and note detection heuristics.
3. Mix together monophonic examples at random, along with their labels.
4. Pretrain a model on these mixes.
5. Finetune on standard datasets.

Monophonic Detection & Transcription



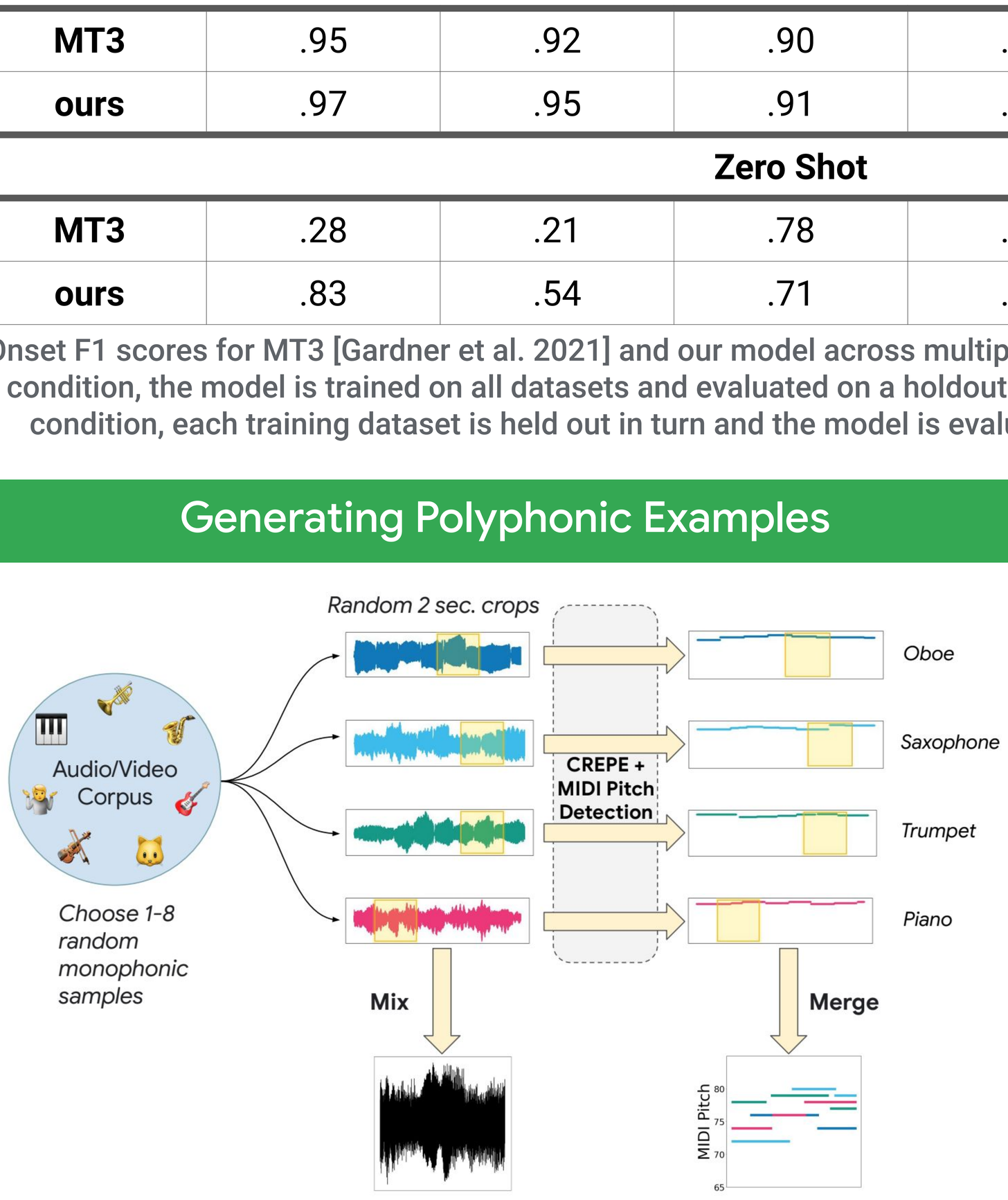
github.com/magenta/mt3

Results (Onset F1)

	MAESTRO	Cerberus4	GuitarSet	MusicNet	Slakh2100	URMP
Standard						
MT3	.95	.92	.90	.50	.76	.77
ours	.97	.95	.91	.56	.83	.90
Zero Shot						
MT3	.28	.21	.78	.18	.14	.23
ours	.83	.54	.71	.48	.45	.54

Onset F1 scores for MT3 [Gardner et al. 2021] and our model across multiple transcription datasets. In the “standard” condition, the model is trained on all datasets and evaluated on a holdout set from each dataset. In the “zero shot” condition, each training dataset is held out in turn and the model is evaluated on the corresponding holdout set.

Generating Polyphonic Examples



Conclusion

Existing music transcription datasets are small and do not contain diverse audio compared to real recordings.

We take advantage of the ease of monophonic transcription to produce a training dataset containing random mixes of 5000 hours of monophonic examples gathered from the wild.

Pretraining a model on this dataset and then finetuning on standard datasets greatly improves our transcription performance, especially in the zero-shot setting.

Try it out at: `github.com/magenta/mt3`