MIX2MORPH: LEARNING SOUND MORPHING FROM NOISY MIXES

Annie Chu[1,2], Hugo Flores García[1,2], Oriol Nieto[1], Justin Salamon[1], Bryan Pardo[2], Prem Seetharaman[1]*

[1]Adobe Research ; [2]Northwestern University
*work conducted during internship

GOAL: Asymmetric Sound Morphs (aka Sound Infusions)

I want a SFX that has the... behavior of **PHONE RINGING** with timbre like **PIGEON COOING**

primary source *secondary source*

Sound Decomposition: Two Axes

phone ringing like phone	phone ringing like a pigeon cooing
pigeon cooing like a pigeon	pigeon cooing like a phone ringing

temporal *timbral*

How? Finetune DiT on Noisy Mixes

1 Make Surrogate Morphs (train what?)

- Sound A + Sound B = Mix — *2 audio sources*
- Augment the Mix — *(See `Our Method`)*

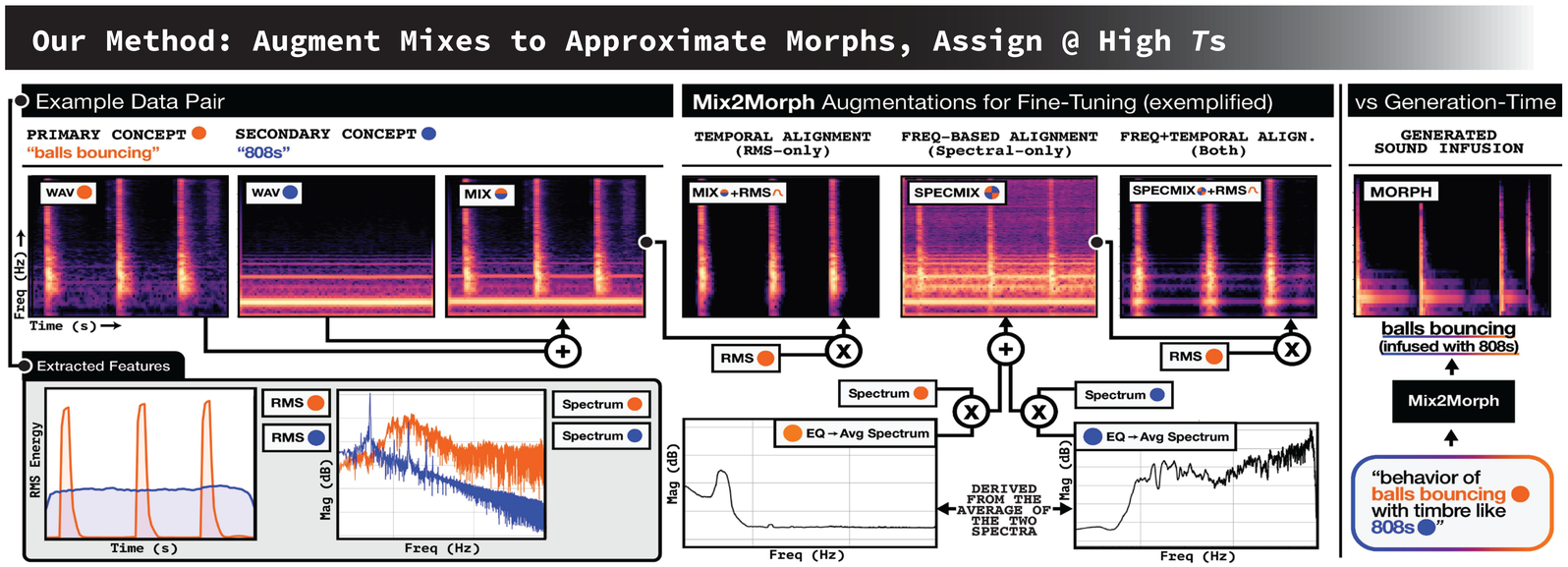
2 Assign to Noisy Timesteps (train when?)

- Assign to high noise timestep range
- Finetune pre-trained text-to-audio DiT

Augmentations turn plain additive mixes into more morph-like training samples

Denoising Dynamics: early timesteps (high noise) encode concepts >> details

Our Method: Augment Mixes to Approximate Morphs, Assign @ High *Ts*



Evaluating Sound Infusions: Morph Metrics

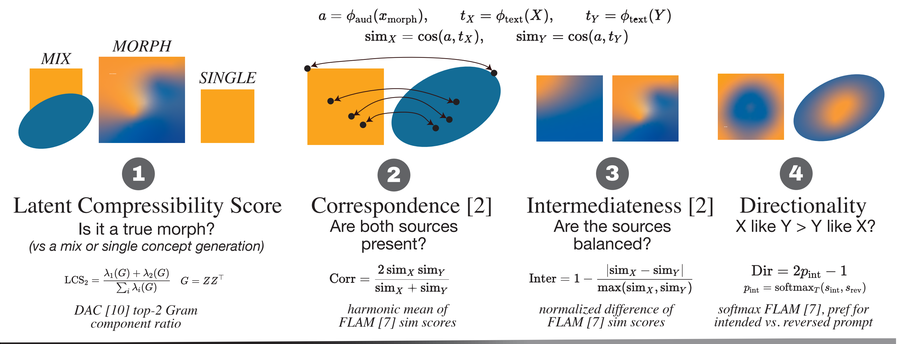


Table 1

Model	*t* start	*t* end	RMS	Spectral	Both	None	LCS ↑	Corr. ↑	Inter.. ↑	Direct. ↑	FAD ↓
base	-	-	-	-	-	-	0.136	0.678	0.611	0.525	1.219
+Timestep Allocation	0	1	✓	✗	✗	✗	0.128	0.699	0.646	0.173	1.230
	0.25	1	✓	✗	✗	✗	0.143	0.705	0.658	0.278	1.235
	★ 0.5	1	✓	✗	✗	✗	0.141	0.721	**0.672**	0.296	1.221
	0.75	1	✓	✗	✗	✗	0.134	0.717	0.653	0.364	1.225
+Augmentation Mode	0.5	1	✓	✗	✓	✗	0.135	0.700	0.623	0.363	1.226
	★ 0.5	1	✓	✓	✓	✗	**0.150**	**0.725**	0.648	**0.436**	**1.220**
	0.5	1	✓	✓	✓	✓	0.143	0.712	0.650	0.349	1.222
Simple Mixing	-	-	-	-	-	-	0.132	**0.758**	**0.690**	-9.25e-13	1.293
LGrS	-	-	-	-	-	-	0.173	0.539	0.638	-0.119	1.290
MorphFader	-	-	-	-	-	-	0.085	0.418	0.421	-9.72e-13	1.430
SoundMorpher	-	-	-	-	-	-	**0.242**	0.591	0.641	-9.64e-13	1.380
Mix2Morph	★ 0.5	1	✓	✓	✓	✗	0.150	0.725	0.648	**0.436**	**1.220**

Mix2Morph ablations

Best Mix2Morph Model
3-way augmentation at [0.5,1]

vs Baselines

Mix2Morph >>
More perceptually coherent midpoint sound infusions

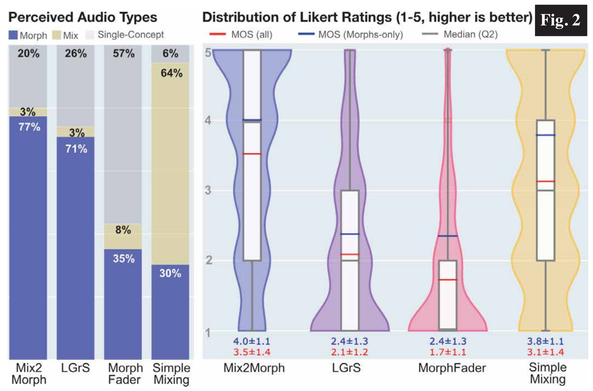


Fig. 2

Results: Listening Study (N=25) and Objective Metrics

References

[1] D. Maul, "From Smart Phones to Carrier Pigeons (when more tech is the last thing you need!)," Derek Maul: Words & Photographs for the Journey, Jan. 19, 2022. [Online]. Available: https://derekmaul.blog/2022/01/19/from-smart-phones-to-carrier-pigeons-when-more-tech-is-the-last-thing-you-need/.

[2] M. F. Caetano and N. Osaka, "A formal evaluation framework for sound morphing," in Non-Cochlear Sound: Proc. 38th Int. Computer Music Conf. (ICMC), Ljubljana, Slovenia, Sept. 9–14, 2012. Michigan Publishing, 2012.

[3] L. M. Heller and L. Wolf, "When hybrid sound effects are better than real recordings," Proc. Meetings Acoust., vol. 46, no. 1, Art. no. 050002, 2022, doi: 10.1121/2.0001581.

[4] A. J. Oxenham, "How we hear: The perception and neural coding of sound," Annu. Rev. Psychol., vol. 69, no. 1, pp. 27–50, Jan. 2018, doi: 10.1146/annurev-psych-122216-011635.

[5] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," Nature, vol. 416, no. 6876, pp. 87–90, Mar. 2002, doi: 10.1038/416087a.

[6] P. Kamath, C. Gupta, and S. Nanayakkara, "MorphFader: Enabling fine-grained controllable morphing with text-to-audio models," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890164.

[7] Y. Wu, C. Tsirigotis, K. Chen, C.-Z. A. Huang, A. Courville, O. Nieto, P. Seetharaman, and J. Salamon, "FLAM: Frame-wise language-audio modeling," in Proc. 42nd Int. Conf. Mach. Learn. (ICML), Proc. Mach. Learn. Res., vol. 267, pp. 67719–67740, 2025.

[8] N. Tokui and T. Baker, "Latent granular resynthesis using neural audio codecs," in Extended Abstracts for the Late-Breaking Demo Session of the 26th Int. Society for Music Information Retrieval Conf. (ISMIR), Daejeon, South Korea, 2025, doi: 10.48550/arXiv.2507.19202.

[9] X. Niu, J. Zhang, and C. P. Martin, "SoundMorpher: Perceptually-uniform sound morphing with diffusion model," arXiv:2410.02144, 2024, doi: 10.48550/arXiv.2410.02144.

[10] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," Advances in Neural Information Processing Systems, vol. 36, pp. 27980–27993, 2023.