

TRIGGER SOUND SUPPRESSION FOR MISOPHONIA

Vaishnavi Vidyasagar, Jasmine Zhang,* Mahima Uliyar,*
Seunghyun Oh,* Emily Catherine Gates,¹ Mark Zachary Rosenthal,¹ Shyamnath Gollakota**

** University of Washington, ¹Duke Center for Misophonia and Emotion Regulation*

ABSTRACT

Misophonia, a disorder of decreased tolerance to specific sounds, affects 5–20% of the population, yet sufferers have no good options: therapy helps a minority, and earplugs or noise cancellation silence everything. We present a study for neural trigger sound suppression for misophonia, selectively removing trigger sounds. We curate a dataset covering the 10 most common trigger classes. Using streaming dual-path networks operating on 6 ms audio chunks, we explore both one-hot and multi-hot-conditioned models that suppress 1–3 triggers from the acoustic scene. We validate our model outputs in a listening study with 30 adults with clinically elevated misophonia impairment. Participants reported significantly lower distress and arousal, and improved valence, for suppressed audio.

1. INTRODUCTION

Misophonia is a disorder characterized by decreased tolerance to specific sounds [1, 2, 3]. Individuals experience intense emotional and physiological reactions, including anger, anxiety, disgust, and fight-or-flight responses, to specific “trigger” sounds. Common triggers are oral and nasal sounds produced by others, such as chewing, lip smacking, slurping, throat clearing, sniffing, and heavy breathing [4]. Because these sounds are pervasive, misophonia can cause social withdrawal, strained relationships, and impaired functioning [5].

Many individuals rely on coping strategies with their own costs: earplugs and noise-canceling headphones suppress *all* sound, isolating users from conversation and important auditory cues, while white noise masking [6] can cause listening fatigue. Ideally we need selective suppression: remove trigger sounds while not suppressing the rest of the acoustic scene.

Recent advances in neural target sound extraction (TSE) [7, 8] make such selective suppression technically plausible. However, prior systems are primarily evaluated for general-purpose sound classes like sirens, birds, speech, and music. Whether they can suppress misophonia triggers remains underexplored. Success must also be measured not in decibels, but in reduced patient distress. No prior work provides isolated recordings of common misophonia triggers at a scale suitable for training separation models, nor validates suppression with people who have misophonia.

In this paper, we explore deep learning-based trigger sound suppression for misophonia. Our contributions are:

1. A misophonia trigger dataset. We curate a dataset for ten common trigger classes, i.e., chewing, coughing, sneezing, throat clearing, lip smacking, drinking/slurping, heavy breathing, nose sniffing, gum popping, tapping.

2. A streaming, on-device suppression model. We train models conditioned on one-hot and multi-hot queries that suppress one to three triggers from the acoustic scene. We use a streaming variant of a dual-path neural network and train 0.496M-parameter models that operate with 10 ms algorithmic latency and run in real time on an Orange Pi device. We also ablate the STFT parameters to characterize latency-quality tradeoff down to a fully causal 4 ms configuration.

3. Validation with misophonic listeners. We evaluate model outputs in a listening study with $N = 30$ adults with clinically elevated misophonia impairment, using chewing, a commonly reported primary trigger. Suppressed clips significantly reduced subjective distress and arousal and improved valence relative to raw audio ($p < .001$).¹

2. RELATED WORK

Misophonia and its treatment. Misophonia is a disorder of decreased tolerance to specific sounds or their associated stimuli [9]. Representative population studies estimate a prevalence of roughly 5-20%, depending on the instrument and threshold [1, 2, 3], while a recent nationally representative U.S. study estimated 4.6% at clinical severity [5]. A randomized controlled trial of psychotherapy found clinical improvement in only 37% of patients [10], and no pharmacological treatment exists. Audiological strategies adapted from tinnitus retraining therapy, including ear-level noise generators and masking [6], offer only partial and non-selective relief. We explore a complementary approach that selectively removes trigger sounds from the acoustic scene in real time.

Target sound extraction. TSE extracts a target sound from a mixture using cues such as class labels [11, 12, 13, 8, 14], enrollment audio [7], or text [15]. More recent work enables streaming separation on low-power hardware [16, 17]. While

¹Code, models and datasets will be open-sourced.

Table 1: Number of audio clips per trigger class in our dataset.

Trigger	Chewing	Cough	Drinking	Gum pop	Heavy breath	Lip smack	Sneezing	Sniffing	Tapping	Throat clearing	Total
FSD50K	231	-	-	-	-	-	-	-	-	-	231
CoughVID	-	1,187	-	-	-	-	-	-	-	-	1,187
ESC-50	-	40	40	-	-	-	20	-	-	-	100
MATA	-	-	34	-	-	21	12	28	-	-	95
VocalSound	-	-	-	-	-	-	332	-	-	3,504	3,836
Freesound	-	-	-	4	-	-	38	8	149	-	199
Deeply	-	-	-	-	-	45	-	-	-	-	45
YouTube	1,191	-	1,530	455	1,175	885	-	1,219	1,208	-	7,663
All	1,422	1,227	1,604	459	1,175	951	402	1,255	1,357	3,504	13,356

these systems show that real-time, class-conditioned extraction is feasible, they focus on general environmental sounds and have not been evaluated on misophonia triggers.

Real-time hearable constraints and clinical validation.

Hearables require low latency, with delays above 20-30 ms being perceptible [14]. Prior work uses streaming models, low-latency STFTs [18], and on-device accelerators [19], but has not evaluated performance on misophonia trigger classes. Clinical studies typically assess audio quality, whereas misophonia requires measuring distress. We address both gaps with per-trigger separation metrics and distress ratings [4].

3. METHOD

Problem formulation. Let $\mathbf{x}[n] \in \mathbb{R}^2$ be a binaural scene of sources $\mathbf{s}_i[n]$ with labels $\ell_i \in \{0, \dots, C\}$, where $\ell_i \geq 1$ indexes a trigger class and $\ell_i = 0$ all non-trigger sources.

The user specifies which triggers bother them through a *multi-hot* query $\mathbf{q} \in \{0, 1\}^C$. Given $\mathbf{q}$, the scene decomposes into a trigger component and a residual,

$$\mathbf{t}_{\mathbf{q}}[n] = \sum_{i: \ell_i \in \mathcal{Q}} \mathbf{s}_i[n], \quad \mathbf{r}_{\mathbf{q}}[n] = \mathbf{x}[n] - \mathbf{t}_{\mathbf{q}}[n]. \quad (1)$$

The desired output is the channel-averaged residual signal, i.e., the full scene minus the queried triggers.

Misophonia trigger dataset. Existing corpora cover our triggers only partially. The AudioSet ontology, whose vocabulary FSD50K inherits, has no classes for lip smacking, gum popping, or drinking/slurping. Where classes exist, labels are weak (clip-level) and triggers typically co-occur with other sounds. We therefore manually curate a dataset covering the top ten clinically motivated classes across oral, nasal, and repetitive triggers. Candidate clips are drawn from FSD50K [20], ESC-50 [21], Freesound tags [22], CoughVID [23], MATA [24], VocalSound [25], Deeply Non-verbal Vocalization Dataset [26] and YouTube videos (to be released as metadata), then verified by two human annotators who (i) confirm the class and (ii) ensure the trigger is isolated without background noise. The resulting dataset contains 13,356 clean trigger clips across 10 classes.

For background sources, we combine non-trigger FSD50K and ESC-50 classes (environmental and domestic sounds), TAU Urban Acoustic Scenes 2024 Mobile Evaluation [27],

and YouTube videos curated to not have the trigger classes. We include speech and music because they overlap spectrally with oral triggers. The dataset contains 26,742 background clips: 5,142 real recordings and 21,600 TAU urban scene clips from environments such as restaurants, classrooms, and transportation. We filtered the available TAU clips for ones only 10 seconds long. For both trigger and background sources, we split 80/10/10% into train/val/test with no overlap.

Training mixtures are generated on the fly. Binaural mixtures are synthesized by convolving each source, with the head-related impulse responses (HRIRs) from the CIPIC HRTF database [28]. Each source is independently assigned a direction drawn uniformly from the 1,250 measured HRIR positions. Each 5 second example draws 1-3 trigger events from distinct classes and a background source. Each trigger clip is placed at a random position in (or randomly cropped to) the 5 s window, and scaled to a trigger-to-background ratio drawn from [0, 15] dB. If clips are under 5 seconds, the clip is zero-padded to 5s with a random amount of silence in front and the rest at the end. We focus on this regime, where triggers are at least as loud as the background, because clearly audible triggers are the most disruptive for the target population. Triggers quieter than the background are left to future work. Evaluation and test mixtures are generated deterministically from fixed per-index seeds, so they are exactly reproducible from the released code and data splits.

Network architecture. We use the streaming TF-GridNet [29] variant from [17]. It takes the two-channel (binaural) mixture $\mathbf{x} \in \mathbb{R}^{2 \times N}$ and directly outputs an estimate of $\mathbf{r}_{\mathbf{q}}[n]$.

The real and imaginary parts of the short-time Fourier transform (STFT) of both channels are stacked and projected to latent channels by a causal convolution with layer normalization. A stack of TF-GridNet blocks follows, each with a residual spectral stage (a bidirectional LSTM along frequency) and a residual temporal stage (a unidirectional LSTM along time for every frequency bin). A causal transposed convolution then maps the latent channels to two output channels, the real and imaginary parts of the spectrogram.

Each STFT frame consists of the current chunk of L_C samples as well as L_B lookback and L_F lookahead samples. Following [18, 19], the synthesis window is zero over the first L_B samples and solved for reconstruction over the remaining $L_C + L_F$, so the algorithmic latency is $(L_C + L_F)/f_s$. We use $L_C/L_B/L_F = 6/6/4$ ms (256 samples at $f_s = 16$ kHz).

Table 2: Performance of a single-trigger multi-class model trained across all 10 trigger classes, evaluated with single-trigger mixtures and a single output channel.

Trigger Class	SI-SNRi (dB)	SNRi (dB)
Chewing	11.88 $\pm$ 4.36	13.18 $\pm$ 4.26
Cough	15.79 $\pm$ 5.47	15.91 $\pm$ 5.30
Drinking	10.73 $\pm$ 3.52	11.74 $\pm$ 3.48
Gum popping	13.29 $\pm$ 4.07	13.85 $\pm$ 3.96
Heavy Breathing	12.48 $\pm$ 3.79	13.31 $\pm$ 3.81
Lip smacking	14.35 $\pm$ 5.27	15.07 $\pm$ 5.05
Sneezing	16.47 $\pm$ 5.48	16.74 $\pm$ 5.27
Sniffing	16.91 $\pm$ 5.21	17.08 $\pm$ 5.15
Tapping	13.27 $\pm$ 3.83	14.09 $\pm$ 3.75
Throat Clearing	14.81 $\pm$ 4.71	15.09 $\pm$ 4.63

The multi-hot query $\mathbf{q} \in \{0, 1\}^C$ is injected by feature-wise linear modulation (FiLM) [30] at the input of every TF-GridNet block: two linear layers map $\mathbf{q}$ to a per-channel scale and shift applied to the latent channels.

We train two classes of models: (i) a *single-trigger multi-class* model over all C trigger classes with $|\mathcal{Q}| = 1$ at both train and test; and (ii) the *multi-trigger multi-hot queried* model trained with $|\mathcal{Q}| \geq 1$ as described above. We also train and compare on the same datasets with Waveformer [8], which is a time-domain based target sound extraction model.

4. EXPERIMENTS AND RESULTS

Hyperparameters. We use the OrangePi configuration in [17] with six GridNet layers and a 32-dimensional latent space, giving an algorithmic latency of 10 ms (6 ms chunk plus 4 ms lookahead). Models are trained in the same chunked mode used at inference so that train and test state handling match. We used AdamW with an initial learning rate of 1×10^{-4} , no weight decay, a batch size of 2 per rank, and gradient clipping with the norm set to 1. We halved the learning rate if validation SNRi did not improve over five epochs, and selected the best-performing checkpoint on the validation set. The multi-class single-trigger models were trained for 100 epochs with 30,000 samples per epoch (3,000 samples per trigger class), so that every class is seen within each epoch. The multi-trigger model was not trained from scratch: we initialized it from the best single-trigger checkpoint and fine-tuned for 50 further epochs under the same settings. The number of trigger sounds per mixture depends on the model type: single-trigger multi-class models see exactly one trigger per mixture, while the multi-trigger model sees 1, 2, or 3 triggers, drawn uniformly at random with equal probability. For a given mixture, trigger classes are drawn uniformly at random from the 10 classes without repetition, so all class combinations are equally likely and no class is over- or under-represented by its clip count; for each selected class, a clip is drawn uniformly at random from that class’s

Table 3: Multi-trigger suppression model with 0.496M parameters, compared to 2.020M for Waveformer baseline.

Model	Metric	# selected trigger classes		
		1	2	3
Ours	SI-SNRi	19.86 $\pm$ 5.82	18.17 $\pm$ 5.23	16.72 $\pm$ 5.10
	SNRi	19.99 $\pm$ 5.47	18.69 $\pm$ 4.82	17.78 $\pm$ 4.43
Waveformer	SI-SNRi	12.92 $\pm$ 5.06	12.30 $\pm$ 4.17	11.90 $\pm$ 3.88
	SNRi	13.67 $\pm$ 4.87	13.76 $\pm$ 4.07	14.07 $\pm$ 3.93

Table 4: Ablation study over STFT parameters for the multi-trigger model, evaluated with a single trigger queried.

$(L_C/L_B/L_F, \text{ms})$	Latency	SNRi (dB)	SI-SNRi (dB)
6/6/4	10 ms	19.99 $\pm$ 5.47	19.86 $\pm$ 5.82
8/4/4	12 ms	17.08 $\pm$ 5.02	16.76 $\pm$ 5.25
4/12/0	4 ms	13.38 $\pm$ 4.84	12.55 $\pm$ 5.08

training clips, and a random 5-second segment is extracted, resampled if it is judged near-silent.

Loss function. All models are trained with a composite loss between the ground truth residue s and its prediction $\hat{s}$:

$$\mathcal{L} = \mathcal{L}_{\text{MRSTFT}}(\hat{s}, s) + \lambda_1 \|\hat{s} - s\|_1 - \lambda_{\text{snr}} \text{SNR}(\hat{s}, s) \quad (2)$$

where $\mathcal{L}_{\text{MRSTFT}}$ is a multi-resolution STFT loss [17] and $\|\cdot\|_1$ is the L_1 loss. We set $\lambda_1 = 10$ and $\lambda_{\text{snr}} = 0.1$.

Results. Table 2 reports the results for our single-trigger multi-class model. The table shows both the per-class residual signal SI-SNRi and SNRi on the held-out test split over 3,000 test mixtures per trigger class. Impulsive, spectrally distinct triggers such as Sniffing (16.91 dB) and Sneezing (16.47 dB) are separated most reliably, while Drinking (10.73 dB) and Chewing (11.88 dB) are the hardest, as their broadband, low-energy transients overlap more heavily with background speech and ambient noise. The gap between residual signal SNRi and SI-SNRi is under 1.30 dB for every class, indicating the residual signals are close to correctly scaled rather than relying on a global gain correction.

Table 3 shows results for the multi-trigger suppression model on test mixtures with multiple triggers. Our model outperforms Waveformer [8] across all three trigger-count conditions. Its SI-SNR and SNR improvements decrease with more concurrent trigger suppression, since the model has the same parameter count as the single-trigger model but must suppress arbitrary combinations of 1–3 triggers.

The ablation study in Table 4 examines how the STFT window budget affects performance for the multi-trigger suppression model. We retrain the model by changing the current chunk size L_C , lookback L_B and lookahead L_F , while keeping their sum constant at 16 ms so the STFT length remains fixed. Removing lookahead entirely (4/12/0) yields a fully causal model at 4 ms latency, but costs 3.70 dB SNRi (4.21 dB SI-SNRi) relative to 8/4/4 window.

The table reports the algorithmic latency. The specific processing time however depends on the target hardware running the model. Our network processes each 6 ms chunk in

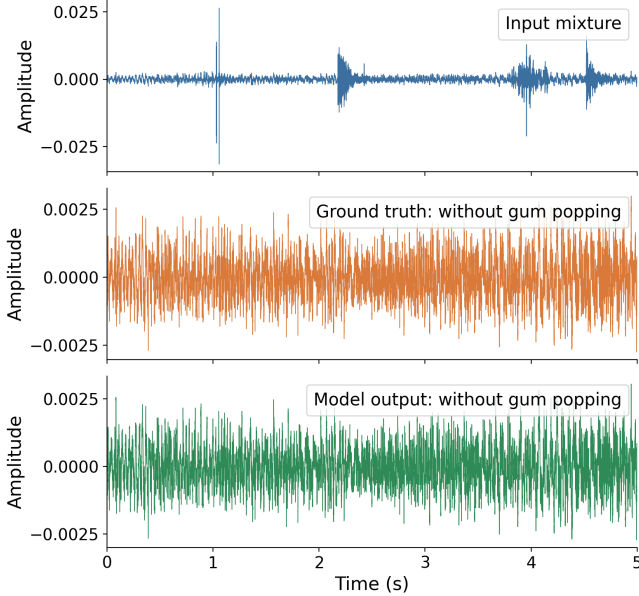


Fig. 1: Time-domain waveforms for gum popping suppression: input mixture, ground truth, and model output.

5.22 ms on average and 5.41 ms at the 95th percentile on an Orange Pi 5B (Arm Cortex-A76 at 2.4 GHz, ONNX).

In Fig. 1, we qualitatively show an average example of single-trigger suppression. The output closely tracks the temporal envelope and fine structure of the ground truth, indicating that the model recovers the full residual signal. The extracted signals are also correctly scaled, so the model preserves the amplitude of the non-trigger sounds in the mixture.

4.1. Listening Study with Misophonic Individuals

We ran an online IRB-approved listening study with $N = 30$ adults who reported chewing, a commonly reported primary trigger, as a bothersome trigger. Participants were recruited via Duke Misophonia mailing lists and social media. Participants had no bilateral hearing loss or profound vision loss. A Duke Misophonia Questionnaire (DMQ) Impairment score of 14 or higher was required as our inclusive criteria, so all participants met criteria for moderate or higher impairment [4]. Participants scored a mean DMQ Impairment of 31.3 (SD 6.0, median 32) out of a possible 48, indicating a sample skewed toward more significant symptoms.

Participants were presented with a series of ten random audio samples, each 5 seconds in length. These corresponded to different source clips from our test set, each presented in both an original version and the model output in individually randomized order. Each audio sample contained a mixture of everyday background sounds together with a chewing trigger sound. Both conditions were presented in mono, generated by the multi-trigger model, and participants were not informed which clips had been through the model. Participants rated their subjective distress (1-10 scale) immediately after

Table 5: Affective response to raw audio vs. model output (30 participants). Lower is better for distress and arousal; higher better for valence. All paired t -tests $p < .001$.

Measure	Original	Output	Δ [95% CI]
Distress (1–10)	7.68 ± 1.60	5.31 ± 2.48	-2.37 [1.45, 3.28]
Arousal (1–9)	7.02 ± 1.42	4.89 ± 2.19	-2.13 [1.36, 2.91]
Valence (1–9)	4.42 ± 2.75	6.01 ± 1.65	$+1.59$ [1.05, 2.13]

each clip, along with emotional arousal and valence (Self-Assessment Manikin scales). A relaxation/reset step was offered between clips to reduce carryover distress.

Table 5 shows that the model output reduced distress and arousal and improved valence relative to the original clips, with large effect sizes on all three measures. Distress fell by 2.37 points ($t(29) = 5.30$, $p < .001$, $d = 0.97$) and arousal by 2.13 points ($t(29) = 5.62$, $p < .001$, $d = 1.03$), while valence rose by 1.59 points ($t(29) = -6.03$, $p < .001$, $d = -1.10$). The three measures move consistently and in the expected directions, indicating that suppression shifts the broader affective response. Post-suppression distress remained above floor, likely because, triggers are idiosyncratic and extend beyond our ten classes. As noted by multiple participants, background scenes (e.g., restaurants) contain individually triggering sounds such as cutlery or typing. Since these sounds were present in both conditions, the observed reduction reflects the effect of removing chewing alone. This is also confirmed by two non-participating adults who did not report hearing chewing in the model outputs.

Interest in our technology was high (8.73/10), as was the perceived impact of losing access (8.63/10). Expected benefit was consistently near 8/10 across academic/work (8.10), home (7.93), social (8.57), and general (8.03) domains.

5. CONCLUSION

We address a practical problem for people with misophonia: suppressing the trigger sounds that distress them.

Discussion. We note that although active noise cancellation (ANC) itself requires sub-millisecond processing, recent work [14, 31] shows that deep-learning-based hearables can operate within a 10 ms algorithmic latency by running ANC to suppress the full acoustic scene and then reinjecting the processed sounds of interest into the ear canal. Our latency therefore falls within the range this architecture allows.

Limitations and future work. Our dataset was curated using online audio sources, many of which are recorded close to the microphone. Augmenting it with in-situ recordings in reverberant rooms, captured on hearable microphones, would reduce the gap between our simulated mixtures and real use. Expanding coverage to more triggers is also needed. Deploying on the AI accelerators found in hearable form factors will require further optimization along the lines of [16]. Future work is also required for training models to output binaural signals and perform end-to-end in-person evaluation.

Acknowledgements

This work was supported by a grant from the Misophonia Foundation. The authors were also supported by a Moore Foundation fellowship.

6. REFERENCES

- [1] E. Jakubovski, A. Müller, H. Kley, M. de Zwaan, and K. Müller-Vahl, "Prevalence and clinical correlates of misophonia symptoms in the general population of Germany," *Frontiers in Psychiatry*, vol. 13, p. 1012424, 2022.
- [2] M. S. Wu, A. B. Lewin, T. K. Murphy, and E. A. Storch, "Misophonia: Incidence, phenomenology, and clinical correlates in an undergraduate student sample," *Journal of Clinical Psychology*, 2014.
- [3] S. Vitoratou, C. Hayes, N. Uglik-Marucha, O. Pearson, T. Graham, and J. Gregory, "Misophonia in the UK: Prevalence and norms from the S-Five in a UK representative sample," *PLoS ONE*, 2023.
- [4] M. Z. Rosenthal, D. Anand, C. Cassiello-Robbins, Z. J. Williams, R. E. Guetta, J. Trumbull, and L. Kelley, "Development and initial validation of the Duke Misophonia Questionnaire," *Frontiers in Psychology*, vol. 12, p. 709928, 2021.
- [5] L. J. Dixon, M. J. Schadeegg, H. L. Clark, C. J. Sevier, and S. M. Witcraft, "Prevalence, phenomenology, and impact of misophonia in a nationally representative sample of U.S. adults," *Journal of Psychopathology and Clinical Science*, 2024.
- [6] P. J. Jastreboff and M. M. Jastreboff, "Decreased sound tolerance: Hyperacusis, misophonia, diplacusis, and polyacusis," in *Handbook of Clinical Neurology*. Elsevier, 2015.
- [7] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 121–136, 2023.
- [8] B. Veluri, J. Chan, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Real-time target sound extraction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [9] S. E. Swedo, D. M. Baguley, D. Denys, L. J. Dixon, M. Erfanian, A. Fioretti, P. J. Jastreboff, S. Kumar, M. Z. Rosenthal, R. Rouw, D. Schiller, J. Simner, E. A. Storch, S. Taylor, K. R. V. Werff, C. M. Altimus, and S. M. Raver, "Consensus definition of misophonia: A Delphi study," *Frontiers in Neuroscience*, vol. 16, p. 841816, 2022.
- [10] I. J. Jager, N. C. Vulink, I. O. Bergfeld, A. J. van Loon, and D. A. Denys, "Cognitive behavioral therapy for misophonia: A randomized clinical trial," *Depression and Anxiety*, 2020.
- [11] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to What You Want: Neural Network-Based Universal Sound Selector," in *Interspeech 2020*, 2020, pp. 1441–1445.
- [12] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 8, p. 1256–1266, Aug. 2019. [Online]. Available: <https://doi.org/10.1109/TASLP.2019.2915167>
- [13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.
- [14] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic hearing: Programming acoustic scenes with binaural hearables," in *Proc. ACM Symposium on User Interface Software and Technology (UIST)*, 2023, pp. 1–15.
- [15] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 458–471, 2025.
- [16] M. Itani, T. Chen, and S. Gollakota, "TF-MLPNet: Tiny real-time neural speech separation," in *Proc. Clarity Challenge, Interspeech*, 2025.
- [17] S. Oh, M. Itani, A. Gauri, and S. Gollakota, "Fine-grained soundscape control for augmented hearing," *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services*, p. 371–391, 2026.
- [18] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "Stft-domain neural speech enhancement with very low algorithmic latency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2023.
- [19] M. Itani, T. Chen, A. Raghavan, G. Kohlberg, and S. Gollakota, "Wireless hearables with programmable speech AI accelerators," in *Proc. ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2025.
- [20] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.
- [21] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia (MM '15)*, 2015, pp. 1015–1018.
- [22] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. 21st ACM Int. Conf. Multimedia (MM '13)*, 2013, pp. 411–412.
- [23] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowd-sourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, p. 156, 2021.
- [24] S. V. Shinkareva, "The misophonia audiovisual trigger archive (MATA)," *bioRxiv*, 2025, preprint. Dataset: <https://osf.io/e7qbm/>, DOI 10.17605/OSF.IO/E7QBM.
- [25] Y. Gong, J. Yu, and J. Glass, "VocalSound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022 - IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2022, pp. 151–155.
- [26] D. Inc., "Deeply Nonverbal Vocalization dataset," 2021. [Online]. Available: <https://github.com/deeplyinc/Nonverbal-Vocalization-Dataset>
- [27] T. Heittola, A. Mesaros, and T. Virtanen, "TAU Urban Acoustic Scenes 2024 Mobile, Evaluation dataset," 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.11366913>
- [28] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2001, pp. 99–102.
- [29] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-gridnet: Making time-frequency domain models great again for monaural speaker separation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [30] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, vol. 32, no. 1, 2018.
- [31] A. T. Sabin, D. McElhone, D. Gauger, and B. Rabinowitz, "Modeling the intelligibility benefit of active noise cancelation in hearing devices that improve signal-to-noise ratio," *Trends in Hearing*, vol. 28, p. 23312165241260029, 2024, pMID: 38831646. [Online]. Available: <https://doi.org/10.1177/23312165241260029>