VOICEFX: CLAP-BASED AUDIO QUALITY IMPROVEMENT FOR SINGING AND SPEECH

Elena Georgieva

Music and Audio Research Lab (MARL)
New York University, NY, USA
elena@nyu.edu

ABSTRACT

This project introduces an automatic method for enhancing audio quality in singing and speech. Using recordings from the LibriSpeech and Smule DAMP dataset, I applied a set of degradations and tested a set of audio effect "remedies" designed to reverse them: a high shelf filter, de-esser, noise reduction, and high-pass filter. I used the CLAP (Contrastive Language-Audio Pretraining) model to estimate recording quality and recommend remedies by comparing audio clips to descriptive text prompts in the shared embedding space. To evaluate my method, I conducted a large-scale listener study with 234 participants and 4,600 ratings. While CLAP encoded some relevant information of vocal recording quality, it often favored remedies like *noisereduce* while when listeners preferred the original clips, suggesting that suggesting that perceptual artifacts introduced by enhancement may not be captured by CLAP. My findings underscore the value of human judgment: embedding models can guide enhancement, but perceptual validation remains valuable. Audio examples are available online on this DAFx demo website.[1]

1. INTRODUCTION

Among the many sounds around us, the human voice tends to be one of the most perceptually salient [1, 2]. As recording technology has become democratized, widespread and accessible, recording quality has become inconsistent and varied [3]. Traditionally, other than re-recording, the best way to improve audio quality is to employ an audio engineer, who can apply audio effects (FX) and processing tools to correct artifacts and degradations. As recording has become more accessible, casual recordings on mobile devices—often with varied microphone quality in uncontrolled environments—have become increasingly common.

Given the volume of content, it is no longer feasible to rely on human input for quality control every time. Additionally, audio effects are often not accessible for non-experts. In this study, I propose a new method for improving vocal audio quality without listener input or expert (e.g., audio engineer) intervention. I take performances of varying recording quality from the LibriSpeech [4] and Smule DAMP datasets [5], apply audio effects as degradations to the files, and use the CLAP embedding space [6] to estimate recording quality and propose remedies. CLAP (Contrastive Language-Audio Pretraining) is a model that embeds both audio and text into a shared space, enabling comparisons between sound clips and text prompts. My remedies were audio effects designed to reverse degradations in the audio files, specifically: a high shelf filter, de-esser, noise reduction tool, and high pass filter. To evaluate the success of my methods, I conducted a listener study with 234 participants and a total of 4600 evaluations, and compared my findings to the CLAP results.

[1] https://ccrma.stanford.edu/~egeorgie/projects/voicefx.html

2. RELATED WORK

VoiceFX builds on a growing body of work in semantic audio production research, which seeks to bridge the gap between natural language descriptions of sound and audio processing parameters [7]. For example, using semantic analysis, users may be able to manipulate audio processing by describing perceptual changes to be applied (e.g., "make this sound 'warmer'"). This VoiceFX builds off a recent paper by Chu et. al that does just that [8]. Researchers used the CLAP embeddings to control audio effects, such as equalization and reverberation, using open natural language prompts. I extend that work to estimate audio quality and suggest effect remedies. While the Text2FX project used a diverse set of 30 audio stimuli, I extend my experiment to 940 audio recordings, and narrow my focus to singing and speech.

In another work (the SocialFX study), participants listened to an audio recording both before and after an effect was applied, and were asked to provide a free-response description. The study showed how such a dataset can help with communication in audio production [9]. Coming out of another crowdsourcing study, the Audealize interface allows users to modify audio by selecting descriptive terms—such as "warm" or "bright" which are then mapped to underlying parameters like EQ or reverb [10].

There are a variety of research studies on the topic of automatic recording quality enhancement for recorded singing voice and speech, specifically. The framework VoiceAssist is designed to improve the clarity of voice recordings and enhance the accuracy of voice command recognition [11]. Meanwhile, denoising and dereverberation are long-standing challenges in speech processing, with entire fields of study dedicated to addressing them due to their prevalence in real-world recordings [12, 13, 14].

In addition to research papers and textbooks, there are a selection of commercial plugins and tools for audio engineers to enhance the sound of the voice.[2] The novel contribution of VoiceFX is applying semantic audio techniques via the CLAP embedding space to enhance audio quality.

[2] https://www.izotope.com/en/products/rx.html, https://www.waves.com/plugins/clarity-vx-pro, https://podcast.adobe.com/en/enhance

3. METHODS

This study investigates whether specific audio enhancement techniques can bring degraded vocal recordings closer to the perceptual ideal of "an isolated voice, well recorded." The audio files selected were 1,000 15-second clips– 500 of audiobook narration from LibriSpeech [4] and 500 of karaoke performance from the Smule DAMP dataset [5], the same used in [15].

I degraded the audio files using a selection of audio effects in order to diversify the sound quality in the dataset and make difference between processed versions of the same recording more apparent to untrained listeners. 800 audio clips were manipulated to reflect five common audio quality issues: background noise, harshness (i.e., sibilance), boomy low end, excessive reverb, and dullness. Noise was applied using the following specifications. Background noise was introduced using SoX-generated pink noise at two amplitude ratios (0.02 and 0.1).[3] Low noise was applied to 80 audio files (40 audiobook, 40 karaoke), and high noise was applied to a different set of 80 audio files (40 audiobook, 40 karaoke). I introduced "harshness" to 160 audio files (80 audiobook, 80 karaoke) using a parametric equalizer applied via ffmpeg[4] with a +24 dB boost centered at 6.5 kHz (Q = 2). Boomy low-end was applied to another 160 audio files (80 audiobook, 80 karaoke) through a low-shelf boost at 150 Hz with +24dB gain (Q=1.4). I applied reverb using SoX with two configurations for light and heavy reverb. Finally, I implemented dullness as a high-shelf cut centered at 4 kHz, attenuated by 24 dB with Q = 0.5, applied to 160 audio files (80 audiobook, 80 karaoke). All clips were EBU normalized using ffmpeg. 200 audio files were left without any applied degradation, only normalization, categorized as "clean."

Next, the Contrastive Language-Audio Pretraining (CLAP) audio/text embedding space measured each audio file's similarity to six pre-defined text prompts, each describing a different vocal characteristic [6]. The six CLAP classes, each preceded by the prompt "this is a sound of..." were:

- An isolated voice, well-recorded
- A vocal recording that sounds dull or muffled
- A vocal recording that sounds harsh or has exaggerated s-sounds
- A vocal recording with extraneous background noise
- A vocal recording with noticeable echo or room sound
- A vocal recording with boomy low-end and mic pops

Each of the CLAP queries was created to match a specific degradation, described above: "An isolated voice, well-recorded" matches the clean condition; "A vocal recording that sounds dull or muffled" matches the "dull" low shelf filter; "A vocal recording that sounds harsh or has exaggerated s-sounds" matches the "harsh" parametric EQ boost; "A vocal recording with extraneous background noise" matches the added pink noise; "A vocal recording with noticeable echo or room sound" matches the added reverb; and "A vocal recording with boomy low-end and mic pops" matches the "boomy" low shelf degradation.

Next, I applied a set of audio remedies; again, each remedy matched a degradation type. Each remedy came in two settings called "medium" and "strong." Noise and reverb were mitigated using the same *noisereduce* Python library [16], with the proportion decrease parameters set to 0.7 and 1.0 for medium and strong, respectively. The *noisereduce* library performs spectral gating to reduce background noise; I call this "noise reduce". "Boomy low-end" was addressed with first-order high-pass filters, applied via ffmpeg, at cutoff frequencies of 80 Hz and 260 Hz, respectively. I call this "de-boom". Dullness was reversed through high-shelf boosts at 4 kHz (+12 dB and +24 dB) (dedull), and harshness was reduced using ffmpeg's de-essing tool with threshold values of 0.5 and 1.0 (de-ess). After each remedy was applied, audio files were normalized for loudness. All eight remedies were applied to all 1,000 pre-remedy audio files, resulting in a total of 8,000 remedied audio files, in addition to the original 1,000 files.

Rating	Displayed Text
+2	The audio is more like *"an isolated voice, well-recorded"* than the reference.
+1	—
0	The audio is neither more nor less like *"an isolated voice, well-recorded"* (neutral).
-1	—
-2	The audio is definitely not more like *"an isolated voice, well-recorded"* than the reference.

3.1. Listener Study Methodology

I conducted a listener study to determine whether these embedding-based results aligned with human judgments. Data were collected online through the Prolific platform.[5] Participants were recruited with eligibility restricted to ages 18–40 to optimize for hearing ability, and were required to have fluent English proficiency. Experiments were hosted on Qualtrics.[6] The Qualtrics survey had audio playback functionality, allowing participants to listen to each 15-second clip before responding. All responses were collected anonymously. The experiment was approved by the Institutional Review Board (IRB), and listeners were compensated for their time. Participants were first presented with the informed consent and then with instructions about the task to be completed. Participants were required to use headphones and underwent a headphone test to determine if they were doing so [17].

Each trial started with listening to the pre-remedy recording, whether it was clean or from one of the five degraded categories (i.e., boomy, reverb, noisy, harsh, dull). Next, participants were presented with five audio files and asked to evaluate them one at a time compared to the original. The five audio files were four remedied versions and a duplicate of the original audio (used as a catch trial). The four remedies were, as detailed above: noise reduction, de-boom, de-dull, and de-ess. For each remedy, either the medium or strong setting was randomly selected.

Next, I calculated the "overall best-performing" remedy (or no remedy), and made sure it was included as one of the five options presented in the trial. The "overall best" was defined as the file within a group that had the highest CLAP similarity score when compared to the target class. If the "overall best" audio was not already randomly selected in the remedies, nor was it the original, the "best" was swapped in for its counterpart.

Participants were instructed to evaluate how much more or less each audio file sounded like "an isolated voice, well-recorded" compared to the original. Ratings were collected on a Likert scale from –2 to +2, as displayed below:

[3] https://sourceforge.net/projects/sox/
[4] https://www.ffmpeg.org/
[5] https://www.prolific.com/
[6] https://qualtrics.com

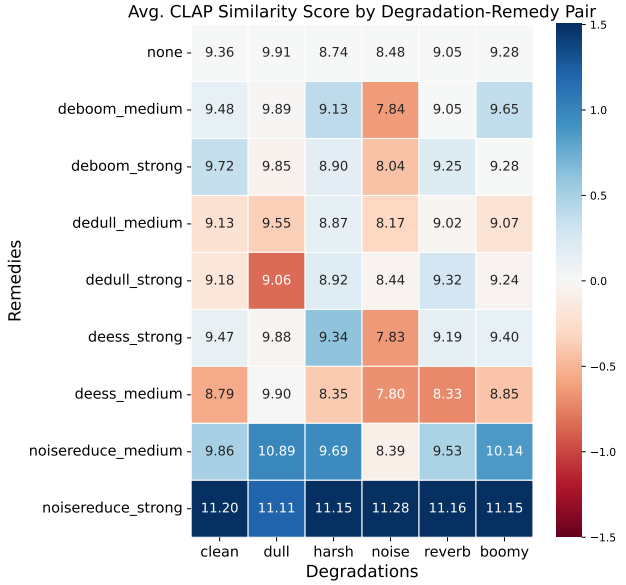

Figure 1: Heatmap of average CLAP similarity scores for all degradation–remedy pairs. Each cell shows the similarity between an audio clip with a given degradation and remedy (or none) and a text prompt. Shading is based on each cell value's change between the pre-remedy and remedied values, on average. Higher values (more blue) indicate greater CLAP similarity to 'an isolated voice, well recorded' after the remedy was applied.

4. RESULTS

A heatmap of average CLAP similarity scores to 'an isolated voice well-recorded' for all degradation–remedy pairs is illustrated in fig. 1. I analyzed whether the intended remedy—chosen to address a specific degradation type—produced the highest similarity to "an isolated voice, well-recorded" among all available options for that particular degradation. I report the average CLAP score change for each remedy within its intended degradation category, and those results are summarized in Table 1. The strongest improvement came with noisereduce_strong paired with noisy files. Similarly, noisereduce_strong yielded a substantial improvement for reverberant recordings. Moderate gains were also observed for noisereduce_medium on reverb, deess_medium on harsh recordings, and deboom_medium on boomy files. In contrast, dedull_medium and dedull_strong both reduced similarity scores on dull recordings, and deess_strong also produced a decline on harsh recordings. Deboom_strong had negligible effect, while noisereduce_medium slightly decreased scores for noisy files.

4.1. Results: Listener Study

I collected data from 369 participants on the Prolific platform. First, I removed 70 participants who failed three or more of the six headphone check questions. Next, I excluded 58 participants who missed five or more of the 20 catch trials and seven participant who self-identified as having non-typical hearing. I was left with data from 234 participants.

The average participant age was 28.5 with a standard deviation of 5.3. 134 of the participants identified as male. The majority of participants identified as White (134 participants; 57%) or Black or African American (79 participants; 34%). 19 participants (8.1%) identified as Hispanic or Latino. When asked about their mother tongue, 167 people responded with English and 102 with another language (some participants selected both English and another option). All participants identified as fluent in English. Each file was evaluated an average of 4.68 times. 63 files only evaluated once or twice were dropped from further analyses.

Figure 2 includes a heatmap of the average user ratings for all degradation-remedy category pairs. I analyzed each degraded audio file matched with its intended remedy (i.e., denoise for noisy audio files), and results are summarized in table 2. The strongest improvement came with deboom_strong for boomy files, followed closely by dedull_medium for dull recordings and dedull_strong. Noisereduce_medium applied to reverberant recordings yielded a modest positive effect, while deboom_medium for boomy files and deess_medium for harsh recordings also showed small but consistent gains. In contrast, deess_strong yielded only a slight improvement on harsh recordings. Listener ratings were negative for noisereduce_strong on noisy files, and even lower for noisere-

Table 1: Mean change in CLAP similarity score for each intended remedy, compared to the untreated degradation condition.

Degradation	Remedy Type	Av. Change	SD	n
High Noise	noisereduce_strong	+2.81	1.56	160
Heavy Reverb	noisereduce_strong	+2.11	1.89	160
Harsh	deess_med	+0.59	1.77	160
Light Reverb	noisereduce_med	+0.47	2.17	160
Boomy	deboom_med	+0.37	1.49	160
Boomy	deboom_strong	0.00	1.63	160
Low Noise	noisereduce_med	–0.08	1.85	160
Harsh	deess_strong	–0.40	1.98	160
Dull	dedull_med	–0.36	1.84	160
Dull	dedull_strong	–0.85	1.83	160

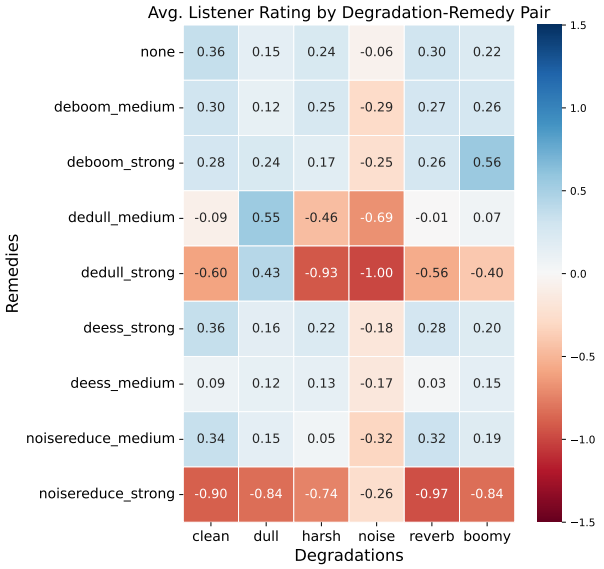

Figure 2: Heatmap of average user ratings for all degradation–remedy pairs. Each cell shows the rating of all audio clip with a given degradation and remedy (or none) combination. Higher values (more blue) indicate greater the listeners thought the audio was more like 'an isolated voice, well-recorded'.

duce_medium on noise, though notably the medium setting was only evaluated on three audio files. The lowest ratings occurred when noisereduce_strong was used to address reverberation, producing a decline in perceived quality.

Table 2: Mean listener rating for each intended remedy–degradation pair, sorted by rating.

Degradation	Remedy Type	Avg. Rating	SD	n
Boomy	deboom_strong	0.56	0.52	83
Dull	dedull_medium	0.55	0.51	74
Dull	dedull_strong	0.43	0.67	76
Reverb	noisereduce_medium	0.32	0.41	39
Boomy	deboom_medium	0.26	0.44	68
Harsh	deess_medium	0.22	0.46	68
Harsh	deess_strong	0.13	0.45	76
Noise	noisereduce_strong	-0.26	0.70	153
Noise	noisereduce_medium	-0.32	0.62	3
Reverb	noisereduce_strong	-0.97	0.56	105

5. DISCUSSION AND CONCLUSION

The thresholds for noisreduce_medium and deess_medium were likely set too low for listeners to consistently distinguish them from the references. Deboom was implemented as a first order highpass filter, which is a subtle audio effect if there is little low-frequency noise in the recording. The remedies that were certainly audible were all scored as much less similar to the clean prompt, on average; none were scored positively as much more similar.

Noisereduce_strong was the most distinguishable effect for listeners, and those audio files were rated almost a full point lower than reference recordings, on average. While remedies using the *noisereduce* Python library were less positive in the listener study, the CLAP-based evaluations came out differently. In the CLAP-based analysis, the two settings of noisereduce resulted in the most positive average change in CLAP similarity to the clean prompt after their application among all remedies. This suggests that while *noisereduce* may improve certain objective or embedding-based measures, human listeners have the opposite experience. For example, *noisereduce* may remove rustle or reverb from a recording, but listeners seem to find the gating of the sound to be distracting, and less subjectively clean. This highlights the value of perceptual validation in the evaluation of audio enhancement methods. Dedull_medium and dedull_strong, applied to dull audio, tended to result in a negative mean change for CLAP but a positive change for human listeners (see tables 1 and 2).

6. ACKNOWLEDGMENTS

Thank you to Brian McFee and Pablo Ripollés for their advice on this project, and to Iran Roman for initial discussions.

7. REFERENCES

[1] A. M. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: the relevance of vocals in the minds of listeners," in *International Society for Music Information Retrieval*, 2018.

[2] M. Bürgel, L. Picinali, and K. Siedenburg, "Listening in the mix: lead vocals robustly attract auditory attention in popular music," *Frontiers Media S.A.*, 2021.

[3] P. Harkins and N. Prior, "(dis)locating democratization: music technologies in practice," *Popular Music and Society*, vol. 45, no. 1, pp. 84–103, 2022.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, South Brisbane, Australia, 2015.

[5] Smule, Inc., "DAMP-VPB: Digital Archive of Mobile Performances," Nov. 2017.

[6] B. Elizalde, S. Deshmukh, Y. Al Ismail, and H. Wang, "CLAP: learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023.

[7] D. Moffat, B. De Man, and J. D. Reiss, "Semantic music production: a meta-study," *Journal of the Audio Engineering Society*, vol. 70, no. 7/8, pp. 548–564, 2022.

[8] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2FX: harnessing CLAP embeddings for text-guided audio effects," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2025.

[9] T. Zheng, P. Seetharaman, and B. Pardo, "SocialFX: studying a crowdsourced folksonomy of audio effects terms," in *ACM Intl. Conference on Multimedia*, Amsterdam, The Netherlands, 2016.

[10] P. Seetharaman and B. Pardo, "Audealize: crowdsourced audio production tools," *Journal of the Audio Engineering Society*, vol. 64, no. 9, pp. 683–695, 2016.

[11] P. Seetharaman, G. Mysore, B. Pardo, P. Smaragdis, C. F. Gomes, S. A. Brewster, G. Fitzpatrick, A. L. Cox, and V. Kostakos, "Voiceassist: a framework for speech enhancement and voice command recognition," in *ACM Conference on Human Factors in Computing (CHI)*, 2019.

[12] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Audio Signal Processing and Applications*, 2013.

[13] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013.

[14] P. Naylor and N. Gaubitch, *Speech Dereverberation*. Springer London, 2011, vol. 59.

[15] E. Georgieva, P. Ripollés, and B. McFee, "A listener-evaluated dataset of amateur karaoke singing and audiobook narration," in *AES AI/ML for Audio Conference*, 2025.

[16] T. Sainburg, "timsainb/noisereduce: v1.0," https://doi.org/10.5281/zenodo.3243139, 2019.

[17] K. J. P. Woods, M. H. Siegel, J. Traer, and et al., "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, pp. 2064–2072, 2017.