FSEEL: A Platform for Listening-Based Evaluation of Forensic Speech Enhancement

Vincent Aubanel, Helen Fraser

Research Hub for Language in Forensic Evidence, University of Melbourne
vincent.aubanel@unimelb.edu.au, helen.fraser@unimelb.edu.au

Abstract

This paper introduces FSEEL, a platform for evidence-based evaluation of Forensic Speech Enhancement (FSE). Current practices in handling poor-quality audio evidence in legal settings are often inadequate, leading to potential miscarriages of justice. FSEEL addresses these issues by providing an integrated interface for comparing and assessing enhanced speech samples. The platform emphasises the need for quantifiable data, reproducible methods, and explicit processing chains to bridge the gap between audio engineering expertise and legal requirements. By combining FSE with forensic transcription in a single interdisciplinary approach, FSEEL aims to improve the reliability of enhanced audio evidence in court.
Index Terms: forensic speech, speech enhancement, speech intelligibility

1. Background

Poor quality audio is frequently presented as crucial evidence in court. However, current legal practices for handling such evidence are problematic, often leading to serious injustices [1]. The legal system treats audio enhancement and transcription as separate processes, both of which are frequently mishandled [2]. This situation persists largely because legal professionals – lawyers, judges, and jury members – evaluate these elements without specialised knowledge in speech perception. A key issue is the lack of recognition of the complex interaction between contextual knowledge, audio enhancement, and transcription. The legal approach, based on precedent rather than scientific principles, fails to account for these crucial relationships. Australian linguists have been actively working to reform these laws, aiming to ensure all forensic audio evidence is provided with demonstrably reliable enhancement and transcription before trial [3]. However, implementing such reforms is challenging as current language and speech sciences cannot provide an off-the-shelf solution for forensic speech enhancement in legal contexts [4]. To address these challenges, the Research Hub for Language in Forensic Evidence at the University of Melbourne proposes to combine Forensic Speech Enhancement (FSE) and Forensic Transcription (FT) into a single interdisciplinary project. This approach uses a listening-based methodology, whereby rigorously tested listeners follow an evidence-based method to produce demonstrably reliable transcripts to accompany responsibly processed audio. This solution represents a significant shift from current practices, and requires collaboration between audio engineering experts, linguists, and legal professionals. It aims to create a scientifically grounded approach to forensic audio evidence, addressing the current disconnects between enhancement and transcription in particular. This paper focuses on the enhancement branch of this project, exploring how an integrated, evidence-based approach to FSE can contribute to more reliable outcomes in legal proceedings.

2. The problem of FSE

2.1. Enhancement in Speech Research

In speech enhancement research, continued efforts in the last decades have achieved impressive results [5, 6], with recent advances using Machine Learning techniques being able to virtually removing mild levels of noise from a speech signal [7, 8, 9, 10]. On the main, speech enhancement research has been driven by common real-life applications, such as removing audible artefacts, attenuating background noise, or separating competing sources. That is, most of the efforts have be aimed at improving speech *quality*, typically occurring with positive Signal-to-Noise Ratios (SNRs). A subset of that research has focused on improving speech *intelligibility*, tackling more adverse conditions (i.e. worse SNRs), with there too positive results, albeit using more controlled speech material [11, 12, 13].

Recent work using causal diffusion networks [14, 15] make it possible to consider new types of distortion outside the traditional additive noise and reverberation scenarios, as well as ability to operate at worse SNRs, making it suitable to tackle more forensic-like situations. While this is a promising approach, we note that suggested solutions to overcome current limitations of these approaches involve adding supplementary information originating from the target material – which is by definition not an option when handling a forensic speech recording. More generally, the question of the evaluation has become central to modern enhancing techniques [16] – and is especially relevant in the forensic context.

2.2. Forensic Speech is special

As introduced in [4], forensic speech departs in significant ways from the speech commonly considered in general speech enhancement research.

Recording conditions are generally far worse than the ones used for typical speech enhancement research programmes. This is because forensic speech is not a research object, with carefully designed degradations. The criterion that makes a recording a sample of interest is that it is deemed to contain incriminating evidence. The acoustic characteristics are therefore largely independent of any theoretical research question, and the levels of degradation can often exceed the level of recoverability.

The difference in degradation is not only quantitative, but also qualitative. Indeed, the type of degradation typically encountered in forensic settings are extremely variable, which make it difficult to propose a unified enhancing approach, but rather calls for ad-hoc solutions. In fact, this dependence on

essentially non-replicable procedures exerts pressures on FSE practice to escape scientific approach, by preventing the collection of large number of examples of similar quality, in order to develop efficient research methods.

The lack of ground truth is both a technical limitation, as the absence of a reference signal or transcript prevents the use of existing scoring methods, and also a conceptual one as it makes harder to specify the objective, and formulate successful enhancing strategies.

Another specificity of forensic speech is that it entertains a much broader context than traditionally-considered speech material. That is, intelligibility can sometimes only be established on the basis of information external to the recording itself – a fact that is usually ignored, or conveniently controlled for, in traditional speech enhancement frameworks. Linked to this point is the fact that while the target "enhanced" speech form is often assumed to be intelligible, with the challenge of enhancing being to overcome externally applied distortions, large part of forensic speech material has very low *intrinsic* intelligibility owing to talker factors, such as poor articulation, interrupted speech or foreign accent, to mention just a few examples. This is important to keep in mind when defining the objective of any enhancing method.

Finally, an essential difference between speech enhancement and FSE relates to their very purpose: while in speech enhancement aim is to model human speech perception, pushing knowledge boundaries at every iterations, the purpose of FSE is strikingly different, as it seeks to support reliable transcription of a specific potentially ambiguous speech production, to enable the court to understand what was said, in the most unambiguous way possible. A corollary of this is that while removing background noise could certainly be helpful in the general situation, and help retrieving what was said by increasing SNR, noise reduction should not constitute an objective of FSE by any measure: firstly noise reduction is commonly found to introduce artefacts [17], usually resulting in a net intelligibility *decrease*, but in some forensic scenarios, the noise itself might contain evidence or contextual information. Therefore, focussing too strongly on noise reduction might disserve the overall purpose of forensic speech enhancement. Further, it has been found that enhancement treatments that loose sight of the forensic purpose result in intelligibility *decrease*: speech perceived to have been enhanced receive increased credibility, and while fewer words are identified experimentally in this condition, it is the enhanced form that is preferred by listeners – with obvious negative consequences in real court situations [18, 1].

The purpose of FSE might therefore include relying on strong parsimony measures in terms of performing change on the speech signal, and should perhaps by guided by an overall aim of intelligibility increase, while at the same time minimising affecting the samples' identity [19].

2.3. Current practices in Forensic Speech Enhancement

In addressing the specific challenges of forensic speech defined above, practitioners have employed what could be called a "toolbox approach" to speech enhancement, that is the application of a set of mitigation strategies tailored to a set of challenges [20]. This has been served by plugins in commercially available software solutions (e.g. iZotope RX, ACON Digital Acoustica, WAVES Audio) or specific forensic systems (CEDAR Audio, Salient Sciences, Cube-Tech). Proposals to structure these strategies and make them more reproducible have been made, for example by suggesting sequential application of processing [21]. In practice however, and without the explicit formulation of objective intelligibility-increase goals informed by evidence-based research methods, the evaluation of speech enhancement is often left to rely on unspecified subjective listening by otherwise highly-talented practitioners, that is, with skills mismatched to that of intelligibility evaluation.

A number of Artificial Intelligence (AI)-inspired tools are also becoming increasingly commercially available, either integrated in the software suites mentioned above or as standalone online resources (e.g., https://hance.ai, https://www.lalal.ai/voice-cleaner/, https://dolby.io/products/enhance/). In contrast to the "toolbox approach", these solutions offer very limited set of control parameters – if any – to enhance speech samples. While the result can sometimes be spectacular, these approaches can be problematic, but from an opposite reason: the practitioner is deprived of the knowledge of the enhancement deployed and cannot justify or explain the modifications that the speech sample has undergone.

The current resulting situation is that enhanced samples which end up in court can be of surprising low quality. This also arises as a consequence oflegal procedures whereby practitioners are required to demonstrate adherence to specific submission criteria, as opposed to scientifically defined criteria for listenability or intelligibility increase [4] .

2.4. The requirements for Forensic Speech Enhancement

There is a need to establish proper evaluation methods for forensic speech enhancement output. On one hand, mature methodology exists in speech research to evaluate enhancement output. On the other hand, expert and ad-hoc approaches are by nature more difficult to capture in replicable and quantitative ways. The goal of the current work is therefore, alongside other work concerned with improving forensic transcription, to establish quantifiable evidence-based method for FSE.

3. FSEEL: Description of the platform

3.1. Definition

Forensic Speech Enhancement for Expert Listening (FSEEL) is a platform aiming at centralising speech enhancement procedures in an integrated interface, enabling streamlined and explicit comparison of processed samples, with a view to fostering a reproducible evidence-based approach. Its purposes and features are described in detail below, but before it is perhaps useful to highlight what it is *not*, to emphasise its functionalities in contrast to existing tools. First, FSEEL is not a *signal* editor, which would allow to explore interactively (i.e., zooming in, selecting portions of the signal etc.). Here, while some degree of flexibility is allowed to interact with samples, the focus is to limit distractive interaction with the signal, and instead to provide visual information, to minimise uncontrolled subjective interpretation of enhancement. Second, FSEEL is not a *transcription* platform. It should rather seen as a preparatory platform, which aim is to process a sample that will undergo a transcription in a later stage. In fact, in has been designed to integrate with Soundscribe, a dedicated platform for transcription of forensic speech (in development in our group). . Finally, FSEEL is not a *modification* platform. Here, all enhancement processes are performed offline. One reason is practical: given the variety of speech enhancement that the platform is aiming to include, it can be difficult to enable real-time access to specific resources. The main reason however is conceptual: we wanted

to keep the focus on an evidence-based workflow in contrast to promoting iterative trial-and-error modification loops.

3.2. Purposes

In its early stage, it is as much an exploratory as a conceptual tool: while it allows users to simulate and visualise the effect of enhancement in very concrete terms – as it enables to listen to the different versions at all stages of processing and read the associated analysis data – it also fosters thinking about the processing chain as a whole, and develop an understanding of the various processes and their impact at different stages of processing, specifically in terms of relative intelligibility and quality gains.

One of the main goal of the platform is to provide a unifying interface for sometimes disparate sources of sound processing. For example, a practitioner may have some preference for specific processing with third-party application, and may wish to include this processing alongside standardised formatting. The current solution aims to provide the integration of third-party enhancement as widely as possible, through the modular specification of processing. It can interact natively with most popular sound processing libraries and tools (e.g. *sox*, *ffmpeg*) and currently supports commercially available plugins in the VST3 and AU format.

Extending its organisation as a processing chain, the platform is also destined to be used "in reverse", that is, as a way to generate degraded speech in forensically meaningful ways. This is an important dimension of our project, and will build on earlier localised initiatives (e.g. [22] and [14])

Finally, we designed the interface to be as explicit as possible with regard to changes made to the speech signal. The user should have a clear idea of exactly the type of processing the sample goes through, and is aided with relevant quantified numerical information associated with each step. This is to help constraining and guiding the expectations of enhancement, and help evaluating their effect at any given step.

3.3. Features

Figure 1 shows a screenshot of the platform, illustrating current features for comparing and exploring various speech enhancements.

The platform allows to handle a set of samples, which is a typical situation, e.g., several extracts taken from different point in time of a lengthy recording. All processing steps are identical for each sample, but the value of parameters can be specified individually for each sample – e.g., it can be desirable to specify different level adjustment strategies for different portions of a recording.

As a first source of visual information, a detailed list of information is displayed relating to the sample, in its original (unprocessed) form, including technical information such as overall duration, file format, file size or sampling rate. Other metadata related to the content of the recording are displayed here as well, such as the number of talkers or the topic of conversation. If a transcription is available, a range of metrics can be computed automatically and displayed here, such as proportion of speech, or speech-to-nonspeech signal-to-noise ratio. Importantly though, this section is meant to be curated in a careful way, in order to control contextual information that is available to the user at this stage of the handling of forensic audio. The architecture is flexible and allows to implement specific context-management strategies (e.g. [23]).

Below the list of sample-specific information, an audio

Figure 1: *A screenshot of the platform. See text for description.*

player is inserted which allows to playback the sample. The list of processing steps is displayed as a table where each row is a module. This emulates the processing chain that is commonly found in Digital Audio Workstations, and indicates strict sequential processing. For each module, its name and a short description are given in the columns with those respective names, and the resulting effect can be listened to by clicking on the corresponding button in the *Result* column, which loads the corresponding sample in the audio player.

The list of modules is thought to be flexible, to allow the inclusion of processing from a wide range of sources, from custom scripts, commercial third-party VST3 plugins to offline manual editing. Additionally, any module can marked to be excluded from being displayed, while still being included in the processing chain.

Adjusting the parameters of each module and visualising their effect is an essential feature of FSEEL in supporting expert listening, and they are colour-coded, in blue and orange respectively. Only a selection of input parameters and output indicators are displayed with their units in their respective column, and their value is displayed in the *Result* column for quick visual evaluation.

Forking is one of the main feature of the platform. It allows to compare the differential effect of a module by varying a specific parameter, or a combination of parameters, and propagate the effect down the chain, therefore fostering replicable and tractable enhancement, while maintaining full flexibility. Forking is implemented visually simply by splitting cells below where the forking is specified, and each branch is visually identified in the button name by appending a letter to the label of the playback button. For example, varying three parameters for a given module will be displayed by inserting suffixes 'a', 'b' and 'c' to each branch down the module list respectively.

The implementation of the platform currently consists of a processing backend currently written in MATLAB, as many speech-related applications are available in that language, and a frontend consisting of a html webpage, for wide compatibility. Other languages can be envisaged to optimise future function-

ality (e.g., interactive processing) and interface with external speech processing libraries.

3.4. Preliminary outline of our first project

We sketch here a typical use of FSEEL for selecting samples to include in a future perception experiment, demonstrating an evidence-based approach to the evaluation of FSE. We employed the Griffith Corpus of Spoken Australian English (GCSAusE) Collection [24], an Open Access resource of conversational speech, which satisfies a number of criteria representative of forensic scenarios: multiple speakers, varying quality of recording, diverse range of background noises and uncontrolled topic of conversation. Importantly, Conversation Analysis (CA)-style transcripts are available, offering a detailed account of the content of the conversation, which we can use as a reference transcription, upon checking its accuracy.

As seen on Figure 1, a few samples from the Griffith corpus are loaded and for the selected sample, basic technical as well as contextual information are displayed, allowing to quickly acquire an understanding of the auditory sample and guide, in evidence-base ways, the expectations of enhancement processes.

This example shows six stages of processing, sequentially ordered from top to bottom. Forking is illustrated in both steps 5 and 6, and the user can intuitively navigate the resulting output at each step and each branch of this processing tree. Selected parameters (in blue) and their resulting measured effect on the sample (in orange) allow for a rapid, objective evaluation at a glance, even before listening, of the relevant parameter being manipulated at each step. We expect that this explicit and exhaustive approach to the enhancing chain will contribute to minimise subjective evaluation effects that are bound to happen when relying exclusively on the auditory modality. In particular, it will assist in focusing attention to specific and explicit signal manipulations, while allowing the user to relate their individual effect to the general objective of (potential) global net intelligibility increase. This approach could also help by reducing the need for replay which, while having been found to increase intelligibility, can also increase overconfidence [25].

Following critical listening, promising samples are selected to form a stimuli base for a listening experiment, to be deployed through Soundscribe (see above). Promising samples in the current example involve identifying regions for which a differential in intelligibility is identified, i.e., when a difference in intelligibility is detected between different steps and branches, which could be attributed to a specific enhancing process.

We think that this explicit approach to enhancing can be of interest to the various and diverse communities involved in forensic enhancing, by illustrating to each the various exponents of speech perception. To the audio engineering community it can suggest a replicable methodology and can emphasise the interaction of enhancing with transcription and context. To legal parties it may clarify that enhancing is not the result of a single on/off button, and does not always result in net intelligibility gains.

3.5. Potential later developments

In its current state, the purpose of the platform is to attract interest from various communities involved and initiate a discussion to solve issues around forensic speech enhancement in an end-to-end approach. Future developments will therefore hopefully be shaped by those interactions. We can however already forsee various improvements that can be done, some of which we detail hereafter.

There is an increase of User Generated Recordings (UGRs, [26]) in forensic scenarios which often come with a video modality. Including capacity to toggle video display could be useful in assisting critical listening, provided context is managed appropriately [23]. Efficient synchronisation of different sources in that context, or of externally enhanced versions of samples would seems a needed feature, which could be largely automatised with existing libraries ([27], [28]).

Finally, an important function of the platform will be to generate an exhaustive contextualised report, to be carried forward alongside the resulting enhanced audio sample. The specifications of the content and the format will have to be jointly established with various actors of the domain.

4. Discussion and further planned developments

In a more conceptual perspective, we plan to extend the functionality of the platform by incorporating to the *Output parameters* a selection of evaluation metrics for each stage of processing. In addition to the battery of traditional objective speech quality metrics such as PESQ, it would be interesting to integrate, for every module, an estimation of *expected* intelligibility or quality increase based on available experimental listener data.

Given the rapid pace of developments in Artificial Intelligence (AI) and its application to speech technology, it is not unreasonable to think that the general speech enhancement problem may be reformulated in the future in terms of a speech separation problem (see, e.g. [29, 15]), at least for favourable signal-to-noise ratios, and for commonly encountered disruptions. In this reformulation, target speech would be a stem in the mixture of sources to separate, and the aim of enhancement would be to segregate each source in order to identify potential incriminating information. The modules in the platform could as a consequence look quite different, i.e. Declicking and Denoising processed might be replaced by relative mixing of varying sources. The functionality of the platform will remain the same in providing evaluating capacities to enhancing processes. In fact, the purpose of a platform such as FSEEL, appears to be even more crucial as the performance of speech enhancement progresses. Indeed, the type of output made in the general framework of generative AI commonly equates or exceeds naturally produced stimuli. We can therefore surmise that there will be increased level of confidence associated with such enhanced output – including when the enhanced version is misleading.

5. Conclusion

In this paper we introduced FSEEL, a platform dedicated to address the challenges posed by Forensic Speech Enhancement. By focusing on quantifiable data and reproducible methods, it seeks to establish the basis for a listening-based evaluation of forensic speech enhancement, which we propose is a necessary step for subsequent reliable transcription. Ultimately, this will permit a more robust, transparent, and just approach to handling forensic audio evidence in the legal system.

6. Acknowledgments

We thank two anonymous reviewers for their helpful comments.

7. References

[1] Fraser, H. and Kinoshita, Y., "Injustice Arising from the Unnoticed Power of Priming: How Lawyers and Even Judges can be Misled by Unreliable Transcripts of Indistinct Forensic Audio," *Criminal Law Journal*, vol. 45, no. 3, pp. 142–152, 2021.

[2] Fraser, H., "Enhancing Forensic Audio: What Works, What Doesn't, And Why," *Journal of Law and Human Dignity*, vol. 8, no. 1, pp. 85–102, 2020.

[3] ——, "Forensic Transcription: Legal and scientific perspectives," in *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, ser. Studi AISV, Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E., and Schmid, S., Eds. IT: Officinaventuno, 2022, vol. 8, pp. 19–32.

[4] Fraser, H., Aubanel, V., Maher, R. C., Mawalim, C. O., Wang, X., Pocta, P., Keith, E., Chollet, G., and Pizzi, K., "Forensic speech enhancement: Towards reliable handling of poor-quality speech recordings used as evidence in criminal trials." *Journal Of The Audio Engineering Society*, 2024.

[5] Yu, D., Gong, Y., Picheny, M. A., Ramabhadran, B., Hakkani-Tür, D., Prasad, R., Zen, H., Skoglund, J., Černocký, J. H., Burget, L., and Mohamed, A., "Twenty-Five Years of Evolution in Speech and Language Processing," *IEEE Signal Processing Magazine*, vol. 40, no. 5, pp. 27–39, Jul. 2023.

[6] O'Shaughnessy, D., "Speech Enhancement—A Review of Modern Methods," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 110–120, Feb. 2024.

[7] Luo, Y. and Mesgarani, N., "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[8] Schröter, H., Rosenkranz, T., Escalante-B., A. N., and Maier, A., "DeepFilterNet: Perceptually Motivated Real-Time Speech Enhancement," May 2023.

[9] Strauss, M., Pia, N., Rao, N. K. S., and Edler, B., "SEFGAN: Harvesting the Power of Normalizing Flows and GANs for Efficient High-Quality Speech Enhancement," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2023.

[10] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., and Gerkmann, T., "Speech Enhancement and Dereverberation with Diffusion-based Generative Models," Jun. 2023.

[11] Cooke, M., Mayo, C., and Valentini-Botinhao, C., "Intelligibility-enhancing speech modifications: The hurricane challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[12] Aubanel, V. and Cooke, M., "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. Interspeech*, 2013, pp. 3592–3596.

[13] Cooke, M., King, S., Garnier, M., and Aubanel, V., "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[14] Serrà, J., Pascual, S., Pons, J., Araz, R. O., and Scaini, D., "Universal Speech Enhancement with Score-based Diffusion," Sep. 2022.

[15] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., Peer, T., and Gerkmann, T., "Causal Diffusion Models for Generalized Speech Enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 780–789, 2024.

[16] de Oliveira, D., Welker, S., Richter, J., and Gerkmann, T., "The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement," Jun. 2024.

[17] Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M., "Effects of noise suppression on intelligibility. II: An attempt to validate physical metrics," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 439–450, Jan. 2014.

[18] Fraser, H., "Don't believe your ears: 'enhancing' forensic audio can mislead juries in criminal trials," *The Conversation*, 2019.

[19] Aubanel, V., "A system to enhance the listening process," Master's thesis, University of Limerick, 2003.

[20] Scientific Working Group on Digital Evidence (SWGDE),, *Best Practices for the Enhancement of Digital Audio*. (v. 20-A-001-2.0), https://www.swgde.org/documents/draft-released-for-comment/, 2023.

[21] Zjalic, J., "A Proposed Framework for Forensic Audio Enhancement," Master's thesis, University of Colorado, 2017.

[22] Sharma, D., Hilkhuysen, G., Gaubitch, N. D., Brookes, M., and Naylor, P., "C-Qual—a validation of PESQ using degradations encountered in forensic and law enforcement audio," *Journal Of The Audio Engineering Society*, no. 8-1, Jun. 2010.

[23] Dror, I., Thompson, W., Meissner, C., Kornfield, I., Krane, D., Saks, M., and Risinger, M., "Context Management Toolbox: A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making," *Journal of Forensic Sciences*, vol. 60, pp. 1111–1112, Jul. 2015.

[24] Haugh, M. and Chang, W.-L. M., "Collaborative creation of spoken language corpora," in *Pragmatics and Language Learning*. T. Greer, D. Tatsuki & C. Roever (Eds.): University of Hawai'i at Mānoa: National Foreign Language Resource Center, 2013, vol. 13, pp. 133–159.

[25] Hilkhuysen, G., Lloyd, J., and Huckvale, M., "Effects of replay on the intelligibility of noisy speech," in *Proc. of 46th AES Conf.*, 2012.

[26] Miller, B. F., Robertson, F. A., and Maher, R. C., "Forensic Handling of User Generated Audio Recordings," *Journal Of The Audio Engineering Society*, no. 10515, Oct. 2021.

[27] Joren, S. and Leman, M., "PANAKO - A Scalable Acoustic Fingerprinting System Handling Time-Scale And Pitch Modification," *Proc. of 15th ISMIR*, 2014.

[28] Ellis, D. P. W., "Robust Landmark-Based Audio Fingerprinting," 2012.

[29] Pons, J., Liu, X., Pascual, S., and Serrà, J., "GASS: Generalizing Audio Source Separation with Large-scale Data," Sep. 2023.