ColourSenseXR: Helping the Hearing-Impaired Perceive Sound Through Colour

*Submitted to IEEE ICCE 2026

1st Yawen Xiao
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
grace.xiao@mail.utoronto.ca

2nd Steve Mann
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
mann@eecg.toronto.edu

3rd Alexander Vicol
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
alexander.vicol@mail.utoronto.ca

***Abstract*—We present a real-time extended reality (XR) system, called ColourSenseXR, designed to enhance sensory perception through the integration of extended intelligence and extended reality. Like the Quest 2, we aim to create consumer electronics that connect individuals with each other and the environment. Our goal is to aid individuals with hearing impairments notice and intepret sounds in daily life. ColourSenseXR creates a colour overlay from audio, allowing users to perceive sound visually, while in passthrough mode on virtual reality (VR) headsets. The system implements and compares two audio-to-visual translation strategies. The first strategy uses rule-based mapping, and converts varying pitches directly into colour with varying hues and brightness levels; this provides interpretable, frequency-driven feedback. The second strategy uses an AI-driven emotion recognition model trained through convolutional neural network (CNN) to estimate emotion from environmental and musical sounds, which are then visualized as dynamic colour gradients. A user study conducted on the Meta Quest 2 evaluates both methods in terms of intuitiveness, emotional expressiveness, and environmental awareness. These results suggest that note-to-colour is easiest for following pitch, while emotion-to-colour is preferred for context and mood communication.**

***Index Terms*—XR, assistive technology, sound visualization, hearing impairment, sensory augmentation, emotion recognition**

I. Introduction

Hearing impairment reduces the ability for affected individuals to perceive their environment, often resulting in missed auditory cues that convey spatial, emotional, and contextual information. Traditional hearing aids amplify sound but cannot capture its affective meaning or situational context. Consequently, users may remain unaware of environmental changes or emotional tones present in conversations, music, or ambient surroundings.

ColourSenseXR follows this idea. We render sound as adaptive colour within the user's passthrough view with the aim to restore awareness without adding significant cognitive load. Research in multisensory cognition shows that visual information often dominates auditory input when both are present [1], [2]. This suggests that augmenting auditory information with visual representation can enhance environmental awareness, particularly for individuals with limited hearing ability.

An important motivation for this project is to safely include the use of AI in affecting experience extending our ability to experience in an artistic and non-biased way; it's crucial for XR technologies, which are swiftly adopting AI as inherent in their infrastructure, to remain as tools to better experience that which already exists rather than adding disharmonious stimuli. The concept of "Mersivity" of expanding technologies to connect people with each other and with nature through technology, and ColourSenseXR is an elegant example of such technology. We're using AI as a tool for accurate reality interpretation rather than adding an additional fictitious layer of virtuality.

With advances in extended reality (XR) and real-time scene understanding, it is possible to blend visual augmentation with human sensory experience. Such integration aligns with the emerging paradigm of *Extended Intelligence (EI)*, which combines human cognition with AI-driven sensory interpretation [3].

This is why a real-time XR system that enables users to perceive sound through adaptive colour overlays in passthrough vision, such as ColourSenseXR, is helpful to the extended intelligence technology community. The system aims to restore lost auditory awareness by visually encoding sound characteristics within the user's field of view. Two complementary translation strategies are implemented:
(1) a deterministic note–colour mapping that represents pitch and frequency as hue and brightness respectively, and
(2) an AI-based emotion–colour mapping that predicts valence–arousal states from environmental audio using a deep learning model trained on the DEAM dataset [4].

By allowing users to *see sound*, ColourSenseXR provides an alternative sensory pathway for understanding the environment. Beyond accessibility, this approach allows for music to be perceived visually; similar to how the world's first spatial computer, the Sequential Wave Imprinting Machine (S.W.I.M), allowed for individuals to view sound waves, ColourSenseXR

aims to expand upon that idea.

II. Related Work

Decades of sonification and visualization research have shown the ability for pitch, loudness, timbre, and rhythm to be mapped to visuals such as hue, brightness, and motion. Below, we group prior research done by the scientific community into three relevant strands, and we position *ColourSenseXR* relative to them.

A. Sound-to-visual mapping and sonification

Converting audio into visual form (and vice versa) has a long history in sonification and visualization research. Music visualizers for the Deaf and Hard-of-Hearing (DHH) show that thoughtful visual channels can increase access and enjoyment for users: combined audiovisual displays can improve perception and task performance when designed carefully [5]. Open-source projects such as AudioLux use LED-based visualizers to allow (DHH) audiences to experience music through colour and light in live performances [6]. In immersive environments, systems such as SoundVizVR demonstrate how spatial sound indicators and visual overlays can make virtual scenes more accessible to users with hearing loss by showing real-time cues for interactive and environmental sounds [7]. These projects collectively motivate the use of colour and spatial overlays as effective channels for conveying auditory information visually in real time.

B. Affective audio analysis and emotion-to-colour mapping

A second strand of research focuses on automatic recognition of affect from audio and its mapping to visual representations. Public benchmarks such as the DEAM dataset support emotion recognition from audio on valence and arousal. [4]. Deep learning models such as CNNs and transformers have achieved high accuracy in predicting emotional patterns from mel-spectrograms and other features [8]. This shows that deep learning models, and other predictive mathematical objects, are capable of being integrated with the emotional-experiential space. Several repositories, such as Vesper, provide pretrained speech or emotion models that support emotion-aware visualization in interactive systems, further supporting this claim [9]. Mapping valence and arousal to colour spaces by commonly using blue–red for valence and dark–bright for arousal is what *ColourSenseXR* intends to do.

C. XR, assistive visualization, and real-time inference

Recent advances in extended reality (XR) platforms with passthrough cameras and on-device computation have enabled low-latency, context-aware visual augmentations for accessibility. Early work on wearable computing and reality mediation by Mann [10] established the foundation for using XR as an interface for sensory enhancement. Contemporary research continues this vision, demonstrating how AI-derived information can be overlaid onto the physical world to improve human perception and awareness. Examples include industrial and research prototypes that visualize speech or sound on AR glasses for DHH users. Practical XR implementations commonly integrate neural inference through frameworks such as Unity Barracuda [11], allowing ONNX-exported models to run efficiently on standalone headsets like the Meta Quest series [12]. Studies on XR accessibility emphasize the importance of low latency, legibility, and minimal distraction when converting audio into visual feedback [13]. These works collectively establish the technological basis that *ColourSenseXR* extends by introducing a comparative evaluation of deterministic and AI-driven audio-to-colour mappings within an assistive XR environment.

III. System Design

This section presents the overall architecture of *ColourSenseXR* and then describes in detail the two sound-to-colour translation strategies implemented: (1) note-to-colour mapping, and (2) emotion-to-colour mapping.

A. Strategy 1: Note-to-Colour Mapping

In this deterministic mapping strategy, musical notes A, B, C, D, E, F, G, ... are mapped in the standard RETMA / EIA colour code used for musical notes [10]:

- A $\rightarrow$ 1 = Brown
- B $\rightarrow$ 2 = Red
- C $\rightarrow$ 3 = Orange
- D $\rightarrow$ 4 = Yellow
- E $\rightarrow$ 5 = Green
- F $\rightarrow$ 6 = Blue
- G $\rightarrow$ 7 = Violet

This chosen colour ordering is outlined which is grounded in musical-colour mapping research [14]. Note that Newton originally proposed seven colours for a seven-note octave to mirror musical structure, and more recent studies in colour–sound correspondence confirm that mapping pitch or tone to colour hue is plausible [15]. We extract the dominant note class from the input and then apply a fixed hue and brightness based on its octave. As a result, users can quickly associate a musical note or pitch with a consistent colour, aiding recognition of tonal events in the environment.

A = 1 = Brown | B = 2 = Red | C = 3 = Orange | D = 4 = Yellow | E = 5 = Green | F = 6 = Blue | G = 7 = Violet

Fig. 1. Color-note mapping from A (Brown) to G (Violet).

B. Strategy 2: Emotion-to-Colour Mapping (Valence–Arousal)

In this AI-driven mapping strategy, environmental audio is analyzed by a CNN trained on the DEAM dataset to predict two dimensions: valence and arousal (ranges from 1 to 9), where valence represents the pleasantness and arousal represents energy. We then map these dimensions into colour following two rules:

- Gradient: valences from low to high correspond to colours in order blue, cyan, lime, yellow, orange, orange-red, red.

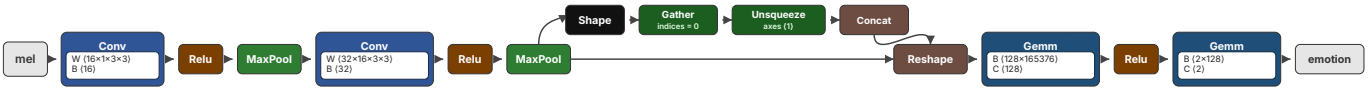

Fig. 2. CNN model structure and training workflow.

- Brightness: low arousal corresponds to dark, high arousal corresponds to bright.

From the 2-D valence–arousal space we interpret the four quadrants as:

- High valence + high arousal → upbeat, excited happiness (e.g., "I feel great and full of energy")
- High valence + low arousal → pleasant calm (e.g., "I feel peaceful and content")
- Low valence + high arousal → negative but intense (e.g., "I feel upset or angry and very worked up")
- Low valence + low arousal → negative and subdued (e.g., "I feel sad or depressed and low-energy")

Studies confirm that hue, saturation, and lightness correlate with emotional valence and arousal ratings of colour stimuli [16], and the circumplex valence–arousal model is widely used in affective computing research [1]. The system converts the predicted (valence, arousal) pair into a colour overlay in real-time, adjusting both hue and brightness smoothly as sound context changes. This enables the user to visually sense not only the presence of sound, but its emotional context.

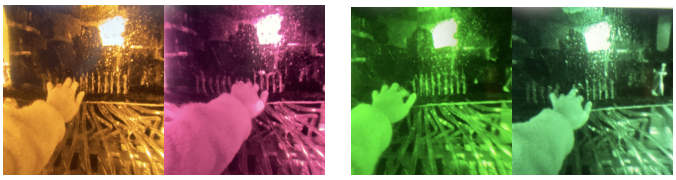

Fig. 3. Colour change (left) and brightness change (right) on the hydraulophone.

IV. Model and Implementation

ColourSenseXR's AI-driven strategy relies on a lightweight CNN, which predicts valence and arousal from musical audio clips. This section describes the model architecture, dataset processing, model training, and Unity-based real-time inference pipeline. The overall architecture is illustrated in Fig. 2.

A. Training Model

The CNN model follows a compact convolutional neural network design implemented in PyTorch and exported to the ONNX format for later deployment in Unity. The model was trained on a desktop PC (CPU: Intel Core i7-13700KF, GPU: NVIDIA GeForce RTX 3060 Ti) using the *Database for Emotional Analysis of Music* (DEAM) dataset [4], which contains 2000 songs annotated with continuous emotional dimensions of valence and arousal in the range $[1, 9]$. Each song was processed in full length (up to 30 seconds per track) to generate its corresponding emotional representation.

During preprocessing, audio was resampled to 22.05 kHz and converted into 64-bin Mel-spectrograms using the `librosa` library, with a hop length of 512 samples and a 30-second analysis window. The resulting Mel-spectrogram tensors had the shape $[1, 64, 1292]$, corresponding to one channel, 64 Mel frequency bins, and 1292 temporal frames per input.

The model consists of three convolutional blocks, each comprising a 2D convolution layer (kernel size 3×3), batch normalization, ReLU activation, and 2×2 max pooling. The flattened feature maps are passed through a fully connected layer (128 hidden units) and a linear output layer producing two continuous outputs $(\hat{v}, \hat{a})$ representing predicted valence and arousal values.

The model was trained using the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 32, and mean squared error (MSE) as the objective function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left[(\hat{v}_i - v_i)^2 + (\hat{a}_i - a_i)^2 \right].$$

Training was conducted for 10 epochs until convergence of the validation loss. After training, the resulting model (`emotion_model.onnx`) was exported to the ONNX format for compatibility with Unity's Barracuda inference engine, enabling low-latency emotion prediction on the Meta Quest 2 platform.

B. Valence–Arousal Colour Mapping in Unity

To translate the Barracuda model's output into real-time colour overlays, we implemented a calibrated mapping function that converts valence–arousal predictions into hue and brightness values. Although the model was trained using Mel-spectrograms generated by `librosa` in Python, Unity's internal audio pipeline differs in its handling of frequency scaling, frame size, and spectral normalization. Consequently, the raw valence and arousal predictions produced by Barracuda vary slightly from those computed offline during training.

To compensate for this discrepancy, we recorded continuous Unity inference outputs into a log file while playing multiple songs of varying emotional tone. The predicted (v, a) values were collected every five seconds and their lower and upper bounds were estimated. These bounds were then used to fit normalization functions that map the predictions into the $[0, 1]$ range accepted by the gradient-based colour renderer.

The final mapping equations for Unity are as follows:

$$\text{valence} = \frac{1}{1 + e^{-2.5\,(\text{valence} - 1.8)}},$$

$$\text{arousal} = \log_{10}\left(1 + 9.0 \times \frac{\text{arousal} - 0.8928}{1.2576}\right).$$

These transformations compress the raw output distributions while preserving perceptual contrast, ensuring that the full dynamic range of colour gradients is utilized.

Normalized valence values control the hue gradient (blue → red), while normalized arousal values modulate brightness (dark → bright). The mapping runs at frame rate on the Meta Quest 2, maintaining real-time performance (under 20 ms per frame). This mapping process ensures that the colours produced in Unity match the model's intended emotional output, despite small differences between `librosa` and Barracuda feature processing.

Fig. 4. Colour gradient representation of valence, transitioning from low (blue) to high (red).

V. Evaluation

A. Experimental Setup

A user study was conducted to evaluate the perceptual effectiveness of the two audio-to-visual translation strategies implemented in *ColourSenseXR*. 10 participants were recruited to complete through different strategies (1) *note-to-colour mapping*, (2) *emotion-to-colour mapping*, and (3) a baseline condition involving direct auditory perception.

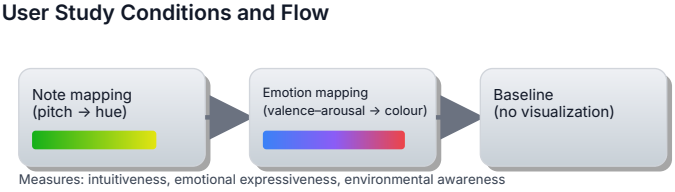

Fig. 5. Flow of User Study

Each participant wore noise-cancelling earphones to simulate a hearing-impaired condition and experienced 4 audio clips in different genres. For both XR visualization strategies, participants observed the colour-encoded music through the passthrough display and completed a short survey after each trial. The same set of four clips was then replayed in the baseline condition, where participants removed the earphones and listened directly to the music. The order of music was shuffled across all trials to prevent bias, while the direct-hearing condition was always presented last to establish a reference baseline. As a result, each participant would answer the same questions 12 times (4 audio clips × 3 strategies), and a total of 120 rows of the data collection.

Each survey collected the following measures for each trial:

- Rated valence (1–9)
- Rated arousal (1–9)
- Best word to describe (Joyful, Calm, Tense, Sad, Excited, Fearful, Neutral)
- Confidence level (1–5)
- A short note describing which visual cues influenced the judgment

B. Results

The results were analyzed by comparing the participants' reported valence and arousal ratings under the two XR strategies against their ratings during direct hearing. Accuracy scores were computed based on how closely the perceived effective responses under visual-only conditions aligned with the baseline emotional perception.

Table I summarizes the averaged quantitative results across all participants and four music conditions. In the baseline (*audio*) condition, participants directly listened to each song and expressed how they felt, serving as a reference for emotional perception. The two XR strategies: *visual_note* and *visual_emotion* were compared against this baseline using emotion match rate, valence mean absolute error (MAE), arousal MAE, and average confidence score.

TABLE I
Performance Comparison Across Strategies

	Visual–Emotion	**Visual–Note**	**Audio (Baseline)**
Number of Trial	40	40	40
Emotion Accuracy	70% (28/40)	32.5% (13/40)	100%
Valence MAE	2.10	2.90	0.00
Arousal MAE	2.78	4.08	0.00
Avg. Confidence	4.2	2.7	4.8

From the notes, participants often mentioned cues such as "fast flashing," "brightness," or "colour intensity" in the *visual_note* strategy, and "warm vs. cool colours" or "colour smoothness" in the *visual_emotion* strategy. These observations provide qualitative insight into how participants interpreted emotional cues under visual-only conditions.

VI. Discussion

The quantitative results demonstrate that the *visual_emotion* strategy was highly effective in conveying emotional information from audio. With an accuracy of 70%, participants were able to identify emotions that closely matched their direct hearing experience, indicating that the emotion-based mapping successfully translated affective cues into colour representations. The valence and arousal MAE were around 2 and 3, showing that participants' emotional ratings typically deviated by no more than ± 3 from the baseline.

Although the *visual_note* strategy achieved a lower emotion match rate of 32.5%, this does not imply it was ineffective. Many emotional categories, such as "tense" and "fearful," or "calm" and "sad," share overlapping affective characteristics, making them harder to distinguish even in direct auditory perception. When focusing on the valence MAE, the *visual_note* approach performed reasonably well, indicating that participants could still perceive the general pleasantness of the music. However, its higher arousal MAE suggests that rapid and

frequent colour changes often led participants to overestimate the music's energy, misinterpreting visual intensity as auditory excitement.

In terms of confidence, participants felt more certain when using the *visual_emotion* strategy (average 4.2) compared to *visual_note* (average 2.7), reflecting its greater intuitiveness. The note-based mapping appears to require a higher level of musical expertise and sensitivity to pitch variations, which may not be accessible to general users.

Qualitative feedback further supports these findings. In the *visual_note* condition, participants often relied on cues that were closely associated with arousal but often exaggerated the perceived excitement due to the colour change for different note. This leads to an overestimation for low-energy music. In the *visual_emotion* condition, participants could easily describe and categorize the colour as cold and warm. This aligns with well-established psychological and perceptual associations between colour and emotion. Warm colours such as red, orange, and yellow tend to evoke feelings of warmth and happiness, whereas cool colours such as blue and cyan are often perceived as calm, relaxed, or sad [1]. These intuitive associations allow users to interpret emotional states more naturally without needing to consciously decode the visual representation.

A. Task Alignment

The users' goal in our study was not just the transcription of pitch, but also real-time conversion of audio signals into visual information in the form of brightness, hue, and colour; this allows for a *feel* of the audio that expands into the visual domain. Mapping audio to a valence arousal space and then to colour demonstrates an isomorphic transformation. The pitch→colour→affect experiential pipeline aligns with the extended intelligence scope of broadening music-emotion recognition. Comparing with our questionnaire's emotional accuracy data, we are able to determine the degree to which ColourSenseXR is effective in capturing emotional responses to auditory stimuli. More specifically, participants' qualitative reports ("warm vs. cool colours," "smoothness/brightness") with emotional accuracy of 70% suggest stable cross-modal correspondences between colour and emotion. The use of a VR headset also likely decreased latency that otherwise may have impacted the emotional experience evaluated by the questionnaire.

B. Why Note-to-colour Underperformed

Pitch on its own was not a sufficient variable to have an effect, but rather, rhythm, timbre, and harmony together drive perceived emotion. Also, our deterministic mapping potentially invited arousal overstimulation due to the fast flickers and high luminance occuring at a more excited state. Lastly, more complex music may create colour mixtures that are hard to parse per note, whereas the emotion overlay compresses such mixture into a single field.

VII. Future Work

Future improvements to *ColourSenseXR* will focus on both hardware optimization and model enhancement. First, the system's auditory capture pipeline can be refined by upgrading the microphone and amplifier hardware to improve sensitivity and signal clarity. A more stable input will ensure higher-quality mel-spectrograms [17] and more accurate emotion inference, especially in real-world environments with ambient noise.

Second, instead of training on full-length songs, future models will use short, dynamic audio clips to predict time-varying valence and arousal. This approach would enable finer-grained emotional tracking and smoother visual transitions, allowing users to perceive evolving emotions more accurately in real time.

Third, adopting more advanced neural architectures. For instance, Vesper has a multimodal emotion recognition networks that could further enhance prediction accuracy and generalization [9]. Finally, the framework can be extended beyond auditory-to-visual translation to other sensory substitution or augmentation tasks, such as mapping tactile or environmental data to visual cues, broadening the potential applications of XR-based sensory intelligence.

VIII. Conclusion

ColourSenseXR demonstrates how "auditory emotion" and pitch can be visually represented in XR through colour. The system provides a promising direction for assistive technologies supporting the hearing-impaired through sensory augmentation. The ability to experience one of the senses (in this case, sound) through another one of the senses (in this case, sight) is a special way of fully immersing oneself into the environment; given the social nature of this endeavour with ColourSenseXR, this enables individuals to connect with each other and with nature through technology (Mersivity).

Acknowledgment

The authors thank the Wearable Computing Lab (wearcam.org) and MannLab (mannlab.com) for conceptual inspiration and the open-source DEAM dataset contributors.

References

[1] C. Spence, "Crossmodal correspondences: A tutorial review," *Attention, Perception, & Psychophysics*, vol. 73, pp. 971–995, 2011. [Online]. Available: https://doi.org/10.3758/s13414-010-0073-7

[2] B. E. Stein, *The New Handbook of Multisensory Processing*. Cambridge, MA: MIT Press, 2012. [Online]. Available: https://mitpress.mit.edu/9780262017121/the-new-handbook-of-multisensory-processing/

[3] Q. Xu, Y. Zhang, and B. Chen, "Extended intelligence: A new paradigm of human–ai collaboration in xr environments," *IEEE Access*, vol. 10, pp. 11 845–11 858, 2022. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3146702

[4] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Deam: Mediaeval database for emotional analysis of music," in *Proc. Int. Soc. for Music Information Retrieval (ISMIR)*, 2017. [Online]. Available: https://cvml.unige.ch/databases/DEAM/

[5] L. Caiola, P. R. Cook, and M. Wright, "A state-of-the-art report on the integration of sonification and visualization," *arXiv preprint arXiv:2402.16558*, 2024. [Online]. Available: https://arxiv.org/abs/2402.16558

[6] CymaSpace, "Audiolux: Visual music systems for the deaf and hard-of-hearing," 2014–2024. [Online]. Available: https://www.cymaspace.org/audiolux/

[7] Z. Li, Y. Chen, and R. Li, "Soundvizvr: Sound indicators for accessible sounds in virtual reality," 2022. [Online]. Available: https://3dvar.com/Li2022SoundVizVR.pdf

[8] M. Zhang, J. Chen, and Y.-H. Yang, "A survey of music emotion recognition: Approaches, databases, and challenges," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2035–2052, 2022. [Online]. Available: https://doi.org/10.1109/TAFFC.2020.3017724

[9] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," 2023. [Online]. Available: https://arxiv.org/abs/2307.10757

[10] S. Mann, "Wearable computing and reality mediation," 1999. [Online]. Available: http://wearcam.org/xr.htm

[11] NVIDIA Developer Blog, "Visualizing spoken language and sounds on ar glasses," 2023. [Online]. Available: https://developer.nvidia.com/blog/speech-ai-spotlight-visualizing-spoken-language-and-sounds-on-ar-glasses/

[12] Unity Technologies, "Barracuda neural network inference library," 2024. [Online]. Available: https://docs.unity3d.com/Packages/com.unity.barracuda

[13] Tetralogical, "Xr accessibility for people with hearing disabilities," 2024. [Online]. Available: https://tetralogical.com/blog/2024/10/01/xr-accessibility-for-people-with-hearing-disabilities/

[14] I. Newton, *Opticks: Or, A Treatise of the Reflections, Refractions, Inflections, and Colors of Light*. London: Royal Society, 1704. [Online]. Available: https://archive.org/details/opticksortreatis00newt

[15] S. Ward and D. Simner, "Exploring the relationship between sound–color synaesthesia and visual imagery," *Cortex*, vol. 113, pp. 104–118, 2019. [Online]. Available: https://doi.org/10.1016/j.cortex.2018.11.015

[16] A. J. Valdez and A. Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology: General*, vol. 123, no. 4, pp. 394–409, 1994. [Online]. Available: https://doi.org/10.1037/0096-3445.123.4.394

[17] W. Lambamo, R. Srinivasagan, and W. Jifara, "Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition," *Applied Sciences*, vol. 13, no. 1, p. 569, 2023. [Online]. Available: https://doi.org/10.3390/app13010569