Audio Engineering Society

Convention Express Paper 451

Presented at the AES 160th Convention
2026 May 28–30, Copenhagen, Denmark

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

The Ambisonic Denoising Paradox: U-Net Processing Degrades ASR Transcription Quality for Medical Speech

Szymon Zaporowski[1] and Bartłomiej Mróz[1]

[1]*Department of Multimedia Systems, Gdańsk University of Technology*

Correspondence should be addressed to Szymon Zaporowski (szyzapor@pg.edu.pl)

ABSTRACT

Spatial audio recording using higher-order Ambisonics offers rich directional information for medical speech capture, yet challenging hospital acoustic environments motivate preprocessing with neural denoising algorithms. This study investigates whether U-Net-based denoising of third-order ambisonic recordings improves automatic speech recognition (ASR) quality for medical applications. We developed the Medical Immersive Audio Corpus (MIAC), comprising 1,759 utterances (6.43 hours) of Polish medical speech recorded with a Zylia ZM-1 microphone in uncontrolled hospital environments, capturing 16-channel third-order Ambisonics across multiple specializations including thyroid ultrasonography, surgical procedures, and general diagnostics. We applied a U-Net architecture with dual attention mechanisms trained using the Noise2Noise paradigm to denoise the corpus, then evaluated transcription quality using ten Whisper ASR models ranging from 39 million to 1.55 billion parameters, including domain-adapted medical variants. Surprisingly, we discovered a "noise reduction paradox" where denoising degraded transcription quality for seven of ten models, with statistically significant increases in Word Error Rate (WER) and Character Error Rate (CER) for general-purpose base, small, and medium models. Only the domain-adapted `whisper-medium-68000-abbr` model showed statistically significant improvement ($p = 0.0008$), while large-scale models (large-v2, large-v3) exhibited robustness with negligible changes. Effect sizes remained small (Cohen's $d < 0.2$) across all models. These counterintuitive findings suggest modern ASR systems implicitly utilize background noise characteristics as informative features, and that preprocessing pipelines should be reconsidered for domain-specific applications. Our results provide practical guidance for medical speech processing system design.

1 Introduction

Automatic speech recognition in clinical environments presents a distinct set of acoustic challenges that differ substantially from those encountered in general-purpose telephony or broadcast scenarios. Hospital consultation rooms combine background corridor noise, equipment interference, and reverberant hard surfaces in ways that degrade speech intelligibility and confound conventional noise suppression strategies. The integration of ASR into medical documentation workflows has motivated interest in spatial microphone arrays capable of capturing directional information [1], with higher-order Ambisonics (HOA) emerging as a recording format that encodes the acoustic field across a spherical harmonic basis of configurable order. Unlike single-channel or first-order Ambisonic systems, third-order arrays provide $(N+1)^2 = 16$ channels of spatial information, enabling finer directional analysis and more targeted noise suppression.

Prior work on spatial audio corpora has produced resources of varying order and scope. The HARP dataset contains seventh-order room impulse responses generated via the image source method [2], while VibraVox offers 45.5 hours of multilingual HOA speech from 188 participants recorded under GDPR-compliant conditions [3]. The L3DAS21 corpus provides approximately 65 hours of dual first-order Ambisonic recordings designed for speech enhancement and sound event localisation and detection tasks [4], and L3DAS22 extended this to roughly 98 hours recorded in a real office environment [5]. The TAU Spatial Sound Events 2019 dataset offers four-channel first-order Ambisonic synthesised recordings across 504 source-sensor configurations [6]. In the domain of Ambisonic speech processing, transformer-based separation methods such as AmbiSep achieve a scale-invariant signal-to-distortion ratio improvement of 17.7 dB on spatial WSJ0-2mix data [7]. Neural denoising has been advanced by the U-Net architecture [8], and the Noise2Noise training paradigm [9] has enabled high-performance restoration without requiring paired clean reference recordings. For transcription, the Whisper model family [10] trained on 680,000 hours of weakly supervised audio has demonstrated strong zero-shot generalisation across languages and domains.

Despite these advances, a systematic study of whether neural denoising preprocessing benefits ASR in a HOA medical speech context has not been conducted. Existing evaluations either assess perceptual quality metrics without transcription outcomes, or evaluate ASR on conventionally recorded single-channel data. The specific interaction between Ambisonic denoising, the spectral characteristics of hospital noise, and the internal representations of large-scale ASR models trained on naturalistic data remains unexplored.

This paper makes the following contributions: (i) the development and characterisation of the Medical Immersive Audio Corpus (MIAC), a 16-channel third-order Ambisonic corpus of Polish medical speech recorded in natural hospital conditions; (ii) an attention-augmented U-Net denoising system trained under the Noise2Noise paradigm that processes all 16 Ambisonic channels simultaneously; (iii) a cross-model transcription evaluation spanning ten Whisper variants from 39 million to 1.55 billion parameters, including domain-adapted medical models; and (iv) a comprehensive statistical analysis revealing a counterintuitive degradation of transcription quality following denoising in seven of ten models tested.

2 The MIAC Corpus

The Medical Immersive Audio Corpus (MIAC) was recorded using the Zylia ZM-1 spherical microphone array, a 19-capsule device whose A-format output was converted to a 16-channel third-order B-format representation using the official Zylia Ambisonics Converter. The resulting signals follow the AmbiX convention with ACN channel ordering and SN3D normalisation. All audio files are encoded at 48 kHz with a 24-bit word length, and peak levels were normalised to -6 dBTP (true peak).

The corpus encompasses 1,759 utterances from twelve medical specialist speakers (identifiers AMV002 through AMV039), with a total duration of approximately 6.43 hours (23,142 s) distributed across 16 CSV-annotated sessions. Recordings were made in medical consultation rooms at the Medical University of Gdańsk that had received no acoustic treatment. These rooms contain hard-surface boundaries, computer equipment, and proximity to active hospital corridors, producing both stationary broadband noise and time-varying interference from adjacent clinical activity. Variable acoustic conditions between sessions reflect realistic deployment scenarios for clinical speech processing systems.

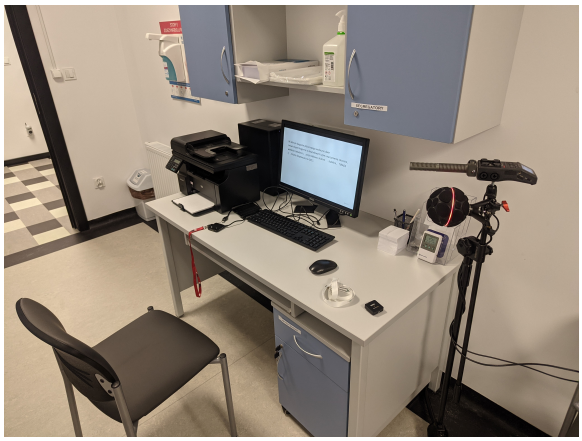


Fig. 1: Zylia ZM-1 microphone array and recording setup in a medical consultation room at the Medical University of Gdańsk.

Table 1: Distribution of utterances per speaker in the MIAC corpus.

Speaker ID	Utterances
AMV002	134
AMV003	98
AMV009	123
AMV012	251
AMV018	186
AMV019	153
AMV023	65
AMV029	183
AMV031	152
AMV032	125
AMV036	152
AMV039	137
Total	1,759

Domain coverage spans four principal areas: thyroid ultrasound diagnostics constitute the largest single domain with approximately 226 utterances, including TIRADS-classified nodule descriptions, Doppler flow narratives, and morphological measurements; surgical procedure recordings document over 50 utterances from laparotomy and cholecystectomy workflows; general medical diagnostics and pathology cover oncological and radiological reports; and breast imaging content describes BI-RADS lesion classifications. All speakers are practising clinicians, ensuring authentic prosody and domain-specific pronunciation.

Each audio utterance is paired with a CSV annotation record containing a structured filename (encoding room code, speaker identifier, microphone type, utterance identifier, and timestamp), a numeric utterance ID, a verbatim phrase transcription, optional clinical context comments, and reserved fields for mean opinion score (MOS) and perceptual quality assessment. The speaker distribution exhibits moderate imbalance, as summarised in Table 1. Speaker AMV012 contributes the largest share with 251 utterances, while AMV023 contributes the smallest with 65 utterances; stratified sampling is therefore recommended when this corpus is used for model training.

3 Denoising Architecture

Denoising was performed using an attention-augmented U-Net operating in the time-frequency domain [8]. Each of the 16 Ambisonic channels was transformed to a complex spectrogram via the Short-Time Fourier Transform (STFT) with a 1024-point FFT (yielding 513 one-sided frequency bins), a hop size of 256 samples, and a Hann analysis window. At a sampling rate of 48 kHz, this configuration yields approximately 187 frames per second. The complete input tensor thus has dimensions $16 \times 513 \times T$, where T denotes the number of time frames.

The encoder path comprises five successive DownBlocks producing feature maps of 64, 128, 256, 512, and 1024 channels respectively. Each block applies two 3×3 convolutional layers with unit padding, Group Normalisation, and a LeakyReLU activation, followed by 2×2 MaxPool2d downsampling that halves the time-frequency dimensions. Group Normalisation was preferred over batch normalisation because it computes statistics within channel groups, which aligns with the natural grouping of spherical harmonics by Ambisonic order.

The bottleneck operates at 1024 channels and incorporates a dual attention mechanism. Channel attention applies global average pooling followed by a fully connected layer and sigmoid activation to generate per-channel weighting vectors, prioritising the fundamental W, X, Y, and Z components. Spatial attention concatenates channel-wise max-pooled and average-pooled feature maps, processes them through a 7×7 convolutional layer with sigmoid activation, and produces a

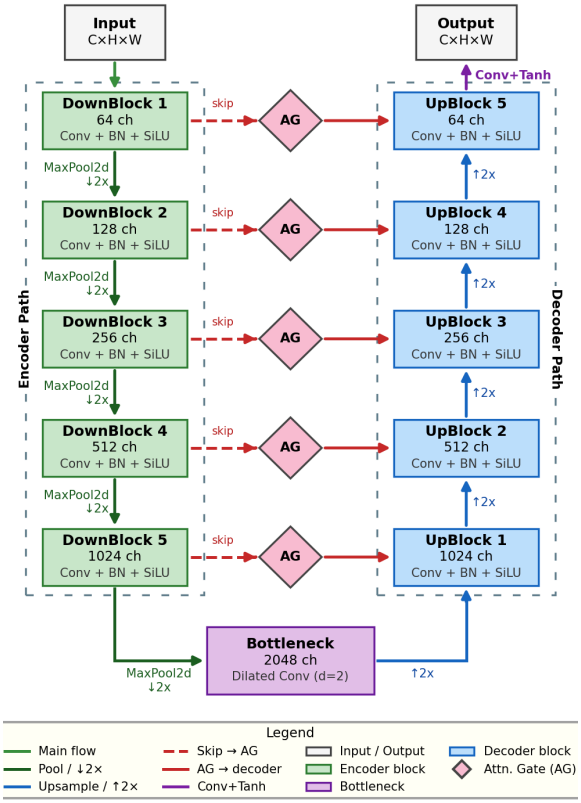


Fig. 2: Architecture of the attention-augmented U-Net used for Ambisonic denoising. The encoder path (green) progressively downsamples via MaxPool layers, while the decoder path (blue) reconstructs the signal through transposed convolutions ($\uparrow 2\times$). Red dashed lines indicate skip connections modulated by attention gates (AG). The bottleneck employs dilated convolutions with d=2 to expand the receptive field at 2048 channels.

time-frequency mask that highlights spectrogram regions containing the target speech signal.

The decoder path employs five UpBlocks. Each block applies a transposed convolutional layer with stride 2 for upsampling, concatenates the result with the corresponding encoder-level skip connection, and passes the combined tensor through two 3×3 convolutional layers with Group Normalisation and LeakyReLU. Attention gates modulate all skip connections; the gates use decoder activations from deeper layers as control signals to implement adaptive feature selection. Selective attention mechanisms are placed specifically in the last two decoder blocks corresponding to the $512 \rightarrow 256$ and $256 \rightarrow 128$ channel transitions, where discrimination between signal structures and residual noise is most critical at high time-frequency resolution. The output layer applies a 1×1 convolution reducing the channel count to two (real and imaginary mask components), followed by a sigmoid activation constraining mask values to the interval $[0, 1]$.

Training followed the Noise2Noise paradigm [9], pairing more-noisy utterances with less-noisy realisations of the same noise process, thereby enabling the network to learn the statistical structure of the interference without access to clean reference recordings. The composite loss function combined three terms: a spectral L_1 norm ensuring amplitude fidelity, spectral convergence expressed as a normalised Frobenius norm of the reconstruction error, and the scale-invariant signal-to-distortion ratio (SI-SDR) computed in the time domain after inverse STFT to enforce perceptual temporal coherence. Automatic Mixed Precision (AMP) training with gradient scaling was employed to reduce GPU memory consumption while maintaining numerical stability. During inference on full-length recordings, the network was applied with a sliding window and an overlap-add reconstruction strategy, enabling constant-memory processing of recordings of arbitrary duration.

4 ASR Evaluation

Transcription experiments were conducted on the omnidirectional W-channel extracted from the 16-channel Ambisonic signal, as this component captures the spatially integrated pressure field and corresponds most directly to the input format expected by single-channel ASR systems. All ten Whisper model variants were applied to both the original and the denoised recordings [10]. The evaluated models span the full Whisper family: tiny (39 million parameters), base, small, medium, large, large-v2, large-v3, and large-v3-turbo (approximately 1.55 billion parameters), supplemented by two medical domain-adapted variants including `whisper-medium-68000-abbr`, which was fine-tuned on Polish clinical speech data. Transcription accuracy was quantified using Word Error Rate (WER) and Character Error Rate (CER), each computed utterance-by-utterance to enable paired statistical comparisons between conditions. The per-utterance difference $\Delta WER = WER_{denoised} - WER_{original}$ served as the primary test statistic throughout the analysis.

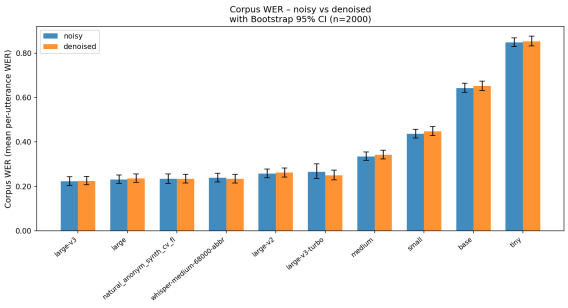


Fig. 3: Corpus-level WER comparison across Whisper model variants for original (noisy) and denoised Ambisonic recordings. Models are ordered by descending parameter count within each group; the dashed line separates general-purpose variants from domain-adapted (fine-tuned) models.

5 Statistical Analysis

Prior to applying parametric tests, the normality of the per-utterance ΔWER distributions was examined using the Shapiro-Wilk test and quantile-quantile plots, and the homoscedasticity of variances across model conditions was assessed via Levene's test. These diagnostics revealed non-normal distributions for the majority of models, motivating the adoption of non-parametric procedures as primary tests and treating parametric results as supplementary.

A sign test was applied separately to each model with the null hypothesis that the proportion of utterances showing WER improvement following denoising equals 0.5. This test is distribution-free and robust to the heavy-tailed, zero-inflated character of WER difference distributions, where a large fraction of utterances yield identical transcriptions under both conditions. The Friedman non-parametric rank-based test was applied across all model-condition combinations to assess whether systematic differences in denoising response existed across the model family. Nemenyi post-hoc pairwise comparisons with Bonferroni-type control were conducted on ranks following a significant Friedman result. Multiple comparisons at the per-model sign test level were addressed using the Benjamini-Hochberg false discovery rate procedure. Effect sizes were estimated using Cohen's d, computed via bootstrap resampling with 10,000 iterations to obtain 95 % confidence intervals that correctly account for the non-normality of the underlying distributions.

6 Results

Sign tests revealed a heterogeneous pattern of denoising effects across the model family. The domain-adapted `whisper-medium-68000-abbr` was the only model to show a statistically significant improvement in WER following denoising ($p = 0.0008$), with 597 utterances improving against 474 degrading. Seven of the ten models exhibited statistically significant degradation; this effect was most pronounced for the base, small, and medium general-purpose variants. The large-v2, large-v3, and tiny models showed no statistically significant WER change ($p > 0.05$) after Benjamini-Hochberg correction, with over 60 % of their utterances yielding zero transcription change between conditions.

Bootstrap estimation of Cohen's d confirmed that all observed effects were small in magnitude. For `whisper-medium-68000-abbr`, $d = -0.0330$ (95 % CI: -0.0834 to 0.0176), indicating a modest directional benefit. The models exhibiting significant degradation showed positive d values, all with confidence intervals bounded well below $|d| = 0.2$. The Friedman test across all model-condition combinations yielded $\chi^2 = 49.29$ ($p < 0.001$), confirming the presence of systematic differences in transcription performance across conditions. Nemenyi post-hoc tests established that statistically significant pairwise differences existed between model size groups (e.g., base versus large-v3), consistent with the large parameter count differences. Within the same model architecture, however, pairwise differences between the original and denoised conditions were not statistically significant after correction for multiple comparisons, indicating that denoising does not introduce a practically meaningful change in transcription quality relative to the variation attributable to model scale. The fine-tuned models formed a distinct performance cluster, significantly outperforming all original Whisper variants in Nemenyi pairwise comparisons.

7 Discussion

The consistent degradation of transcription quality observed for smaller general-purpose Whisper models following denoising can be interpreted through the lens of distributional mismatch. The Whisper models were trained on internet-sourced audio that encompasses a

wide variety of recording conditions, including recordings with substantial background noise and mild reverberation. Internal representations learned during this training may encode noise characteristics as informative contextual cues, analogous to the channel and environment cues that have been documented in robust speech recognition research. When the denoising U-Net removes these spectral features, the input distribution shifts away from the manifold of the training data, introducing a mismatch that manifests as increased transcription errors. This phenomenon may appropriately be termed a noise reduction paradox: the preprocessing step that would intuitively be expected to facilitate recognition actively degrades it by removing features the model has learned to exploit.

Model scale mediates the sensitivity to this distributional shift. Larger models with substantially more parameters exhibit robustness, with $> 60\,\%$ of utterances yielding identical transcriptions regardless of preprocessing. This robustness may reflect a greater diversity of learned representations that are less dependent on any single spectral characteristic. The domain-adapted `whisper-medium-68000-abbr` model presents a different profile: its fine-tuning on Polish medical speech may have replaced reliance on general noise statistics with domain-specific phonological and lexical priors, making its acoustic front end more tolerant of spectral modifications introduced by denoising. The base fine-tuned variant, lacking this adaptation, continued to exhibit degradation, underscoring that domain specialisation of the ASR model is a necessary condition for denoising to provide transcription benefit in this corpus.

The practical implications for clinical ASR system design are clear. General-purpose Whisper models, particularly those below the large scale, should not be preceded by neural denoising preprocessing when deployed on hospital Ambisonic recordings, as this degrades transcription accuracy without providing offsetting gains. Large-scale models are effectively neutral with respect to denoising and may be used with or without it based on other system constraints. For domain-adapted models, denoising benefit is conditional and requires empirical validation on the target corpus and noise profile. Denoising may remain warranted as a preprocessing step when the primary objective is perceptual quality for human listeners rather than machine transcription accuracy, a distinction that clinical workflow design must explicitly address.

8 Conclusions

This paper has introduced the Medical Immersive Audio Corpus (MIAC), a 16-channel third-order Ambisonic corpus of 1,759 Polish medical speech utterances recorded in natural hospital conditions, and has reported a systematic evaluation of the effect of attention-augmented U-Net denoising on ASR transcription quality across ten Whisper model variants. The central finding is that denoising degrades WER for seven of ten models tested, with the dominant source of transcription performance variation being model architecture and the presence of domain-specific fine-tuning rather than preprocessing. Effect sizes are uniformly small, confirming that the observed degradation, while statistically reliable, does not disqualify neural denoising from clinical pipelines in which perceptual quality is the primary objective. Recommended directions for future work include: development of multichannel ASR architectures that natively process all 16 Ambisonic channels to exploit directional information rather than discarding it in channel extraction; joint end-to-end optimisation of denoising and ASR objectives to circumvent the distributional mismatch identified here; perceptual evaluation of denoised audio quality using MOS assessment; and extension of the MIAC corpus with additional speakers and medical specialisations to improve stratification and domain breadth.

Acknowledgements

This research was supported by the Polish National Centre for Research and Development (NCBR) within the project: “ADMEDVOICE - Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process”. No. INFOSTRATEG4/0003/2022

References

[1] Pulkki, V., “Spatial Sound Reproduction with Directional Audio Coding,” *Journal of Audio Engineering Society*, 55(6), pp. 503–516, 2007.

[2] Diaz-Guerra, D., Miguel, A., and Beltran, J. R., “HARP: Higher-Order Ambisonic Room Impulse Response Dataset Generated via the Image Source Method,” arXiv:2311.00507, 2023.

[3] Hauret, J., Fontaine, M., Joubaud, T., Langrenne, C., Poirée, S., Zimpfer, V., and Bavu, É., "VibraVox: A GDPR-Compliant Multilingual Higher-Order Ambisonic Speech Dataset," arXiv:2405.14803, 2024.

[4] Guizzo, E., Marinoni, C., Pennese, M., Ren, X., Zheng, X., Zhang, C., Masiero, B., Uncini, A., and Comminiello, D., "L3DAS21 Challenge: Machine Learning for 3D Audio Signal Processing," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021, doi:10.1109/MLSP52302.2021.9596248.

[5] Guizzo, E., Gramaccioni, R. F., Jamili, S., Marinoni, C., Olivieri, E., Recchiuto, C., Uncini, A., and Comminiello, D., "L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9186–9190, 2022, doi:10.1109/ICASSP43922.2022.9746872.

[6] Adavanne, S., Politis, A., and Virtanen, T., "A Multi-Room Reverberant Dataset for Sound Event Localization and Detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 10–14, 2019.

[7] Levkovitch, A., Rafaely, R., and Schechner, Y., "AmbiSep: Ambisonic-to-Ambisonic Reverberant Speech Separation Using Transformer Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9187–9191, 2022.

[8] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Springer, 2015, doi:10.1007/978-3-319-24574-4_28.

[9] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T., "Noise2Noise: Learning Image Restoration without Clean Data," in *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974, 2018.

[10] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023.