

Cross-Attention with Confidence Weighting for Multi-Channel Audio Alignment

Ragib Amin Nihal¹, Benjamin Yen^{1,2}, Takeshi Ashizawa¹, Kazuhiro Nakadai¹

¹Systems and Control Engineering, Institute of Science Tokyo, Japan ²RIKEN BDR, Japan

Abstract—Multi-channel audio alignment is a key requirement in bioacoustic monitoring, spatial audio systems, and acoustic localization. However, existing methods often struggle to address nonlinear clock drift and lack mechanisms for quantifying uncertainty. Traditional methods like Cross-correlation and Dynamic Time Warping assume simple drift patterns and provide no reliability measures. Meanwhile, recent deep learning models typically treat alignment as a binary classification task, overlooking inter-channel dependencies and uncertainty estimation. We introduce a method that combines cross-attention mechanisms with confidence-weighted scoring to improve multi-channel audio synchronization. We extend BEATs encoders with cross-attention layers to model temporal relationships between channels. We also develop a confidence-weighted scoring function that uses the full prediction distribution instead of binary thresholding. Our method achieved first place in the BioDCASE 2025 Task 1 challenge with 0.30 MSE average across test datasets, compared to 0.58 for the deep learning baseline. On individual datasets, we achieved 0.14 MSE on ARU data (77% reduction) and 0.45 MSE on zebra finch data (18% reduction). The framework supports probabilistic temporal alignment, moving beyond point estimates. While validated in a bioacoustic context, the approach is applicable to a broader range of multi-channel audio tasks where alignment confidence is critical. Code available on: <https://github.com/Ragib-Amin-Nihal/BEATsCA>

Index Terms—Multi-channel audio alignment, cross-attention, confidence weighting, BEATs, bioacoustic monitoring

1. INTRODUCTION

Multi-channel audio recording systems serve applications ranging from professional spatial audio production to scientific bioacoustic monitoring using automated recording units (ARUs) [1]. These systems deploy multiple synchronized devices to capture spatial information, enable source separation, and provide measurement redundancy. Maintaining precise temporal alignment between recording channels presents a significant technical challenge.

The primary obstacle is clock drift between independent recording devices. Oscillator variations due to manufacturing tolerances, temperature changes, and component aging cause temporal desynchronization that accumulates over time [2]. This drift is often nonlinear and unpredictable in field deployments with variable environmental conditions. Applications requiring sub-millisecond accuracy, like bioacoustic localization, need post-processing correction [3].

Figure 1 shows the alignment task. For temporally drifted stereo recordings, the system must predict corresponding timestamps using sparse keypoint annotations. This requires modeling nonlinear drift while providing confidence estimates.

Traditional methods for multi-channel alignment include cross-correlation techniques. Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [4] identifies time delays through correlation peaks between channels. These efficient methods assume constant time shifts but fail with nonlinear drift [5]. Dynamic Time Warping (DTW) handles nonlinear relationships but has $O(N^2)$ complexity and produces unrealistic many-to-one alignments [6]. Modern DTW variants like Memory-Restricted Multiscale DTW reduce computation but lack uncertainty quantification [7]. Speech processing tools such as forced aligners enable automatic phoneme-level segmentation [8]. However, these approaches primarily address single-modality tasks,

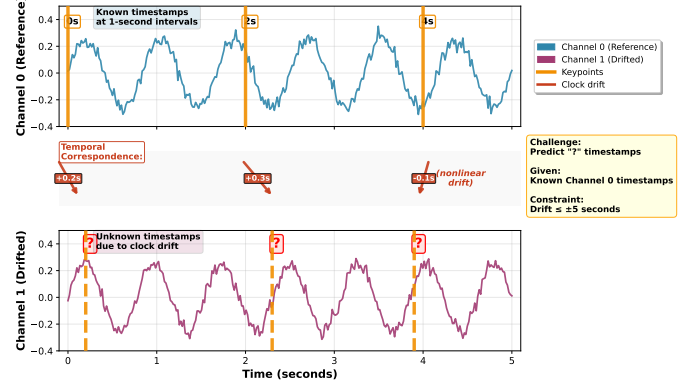


Fig. 1: Multi-Channel Audio Alignment Problem. Illustration of temporal desynchronization between two audio channels due to nonlinear clock drift. Channel 0 (top) provides reference timestamps at known 1-second intervals, while Channel 1 (bottom) contains corresponding but unknown timestamps (marked “?”). Arrows indicate temporal correspondences with drift values of +0.2s, +0.3s, and -0.1s, demonstrating the nonlinear nature of clock drift. The challenge is to predict the unknown Channel 1 timestamps given only the Channel 0 reference, with drift constrained to ± 5 seconds.

leaving multi-channel audio alignment underexplored.

Deep learning approaches have emerged as promising alternatives for temporal alignment. The Audio Spectrogram Transformer (AST) [9] uses attention for audio classification, while Conformer architectures [10] combine convolution and self-attention for temporal modeling. BEATs [11] employs iterative pre-training for diverse audio tasks. Self-supervised approaches like wav2vec 2.0 [12] advance speech representation learning. However, existing deep learning alignment systems typically formulate the problem as binary classification—predicting whether audio segments are aligned or misaligned.

This binary classification approach has two limitations that our work addresses. First, methods process channels independently, ignoring correlated clock drift patterns between synchronized devices. Second, binary classification omits reliability estimates needed for scientific applications like bioacoustic analysis [13].

Uncertainty quantification has proven valuable in related sequence alignment domains. Methods like GUIDANCE [14] provide confidence scores for alignment regions in bioinformatics applications. Similar probabilistic frameworks are needed for audio alignment, where decisions should be weighted by confidence rather than treated as binary choices.

Current alignment methods cannot model inter-channel dependencies (signal processing) or quantify uncertainty (deep learning). We hypothesize that explicitly modeling inter-channel temporal relationships through cross-attention mechanisms, combined with confidence-weighted scoring that utilizes full prediction distributions, can improve alignment accuracy while providing meaningful uncertainty estimates. Cross-attention learns relationships between temporal patterns across channels, capturing correlated clock drift. Confidence weighting uses

continuous model outputs rather than binary thresholds, enabling certainty-based decisions.

We extend the BEATs encoder [11] with cross-attention layers for inter-channel interaction before alignment prediction. Binary counting metrics are replaced by a confidence-weighted function incorporating prediction confidence, top-quartile averaging, and sigmoid-transformed scores. The approach maintains compatibility with existing candidate generation while processing decisions through learned dependencies and probabilistic confidence.

We evaluate our method on the BioDCASE 2025 Task 1 challenge, which provides stereo audio recordings with temporal keypoints from two distinct acoustic domains: field-deployed automated recording units and controlled laboratory recordings. The challenge constrains drift to ± 5 seconds and uses mean squared error for evaluation.

Our work contributes:

- 1) Cross-attention mechanisms that model inter-channel temporal dependencies for audio alignment
- 2) A confidence-weighted scoring framework using full prediction distributions to quantify alignment uncertainty.

Experimental results show improved performance on the BioDCASE 2025 benchmark.

2. PROBLEM FORMULATION

We define multi-channel audio alignment as predicting temporal correspondences with uncertainty quantification. Given a stereo audio recording $\mathcal{A} = (\mathcal{A}_0, \mathcal{A}_1) \in \mathbb{R}^{2 \times T}$ that has been affected by nonlinear clock drift. The goal is to predict corresponding timestamps between channels while providing confidence scores.

2.1. Mathematical Setup

Temporal correspondences are established through keypoints $\mathcal{K} = \{k_0, k_1, \dots, k_{N-1}\}$, where each $k_i = (k_{i,0}, k_{i,1})$ consists of timestamps from Channel 0 (reference) and Channel 1 (drifted). Channel 0 timestamps occur at fixed 1-second intervals: $k_{i,0} = i$ for $i \in \{0, 1, \dots, N-1\}$.

The clock drift function $\mathcal{D} : \mathbb{R} \rightarrow \mathbb{R}$ maps Channel 0 to Channel 1 timestamps:

$$k_{i,1} = \mathcal{D}(k_{i,0}) = \mathcal{D}(i),$$

subject to the constraint $|\mathcal{D}(t) - t| \leq 5$ seconds.

2.2. Training and Inference

Training: The system accesses complete annotations $\mathcal{K}_{train} = \{(k_{i,0}, k_{i,1})\}_{i=0}^{N-1}$ and learns alignment decisions from audio segments extracted around keypoints. For each keypoint k_i , we extract audio segments of duration $\tau = 2$ second:

$$\begin{aligned} \mathcal{S}_{i,0} &= \mathcal{A}_0[\lfloor k_{i,0} \cdot f_s \rfloor : \lfloor (k_{i,0} + \tau) \cdot f_s \rfloor], \\ \mathcal{S}_{i,1} &= \mathcal{A}_1[\lfloor k_{i,1} \cdot f_s \rfloor : \lfloor (k_{i,1} + \tau) \cdot f_s \rfloor], \end{aligned}$$

where f_s is the sampling frequency and $\lfloor \cdot \rfloor$ denotes integer indexing.

Inference: Given only Channel 0 timestamps $\mathcal{K}_{test} = \{k_{i,0}\}_{i=0}^{N-1}$, the system predicts Channel 1 timestamps $\{\hat{k}_{i,1}\}_{i=0}^{N-1}$ by evaluating candidate alignments with learned scoring.

2.3. Candidate Generation Strategy

Following the baseline approach, we model the drift as affine: $\mathcal{D}(t) \approx \alpha t + \beta$ where α represents drift rate and β represents constant offset. The candidate generation creates a discrete set $\mathcal{C} = \{(\alpha_j, \beta_j)\}_{j=1}^M$ by sampling:

$$\begin{aligned} \alpha_j &\in [1 - \frac{\delta_{max}}{T_{dur}}, 1 + \frac{\delta_{max}}{T_{dur}}] \\ \beta_j &\in [-\delta_{max}, \delta_{max}], \end{aligned}$$

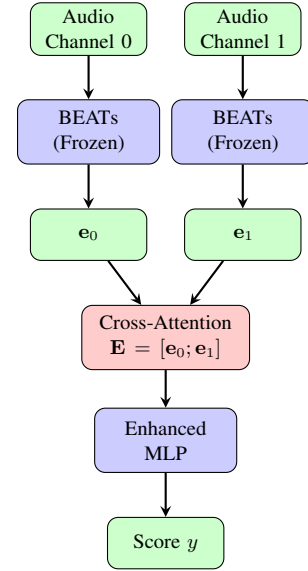


Fig. 2: System architecture with cross-attention enabling inter-channel interaction before alignment prediction.

where $\delta_{max} = 5$ seconds and T_{dur} is audio duration. For each candidate, predicted timestamps are $\hat{k}_{i,1}^{(j)} = \alpha_j \cdot i + \beta_j$, generating candidate sets $\mathcal{K}_j = \{(k_{i,0}, \hat{k}_{i,1}^{(j)})\}_{i=0}^{N-1}$.

2.4. Evaluation

Performance is measured using Mean Squared Error (MSE):

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=0}^{N-1} (k_{i,1} - \hat{k}_{i,1}^*)^2. \quad (1)$$

The benchmark combines scores from two datasets: $\mathcal{L}_{final} = \frac{1}{2}(\mathcal{L}_{MSE}(\mathcal{D}_{ARU}) + \mathcal{L}_{MSE}(\mathcal{D}_{zebra}))$.

3. METHODOLOGY

We modify BEATs with: (1) cross-attention for inter-channel dependencies, (2) conservative augmentation during training, and (3) confidence-weighted scoring. Figure 2 shows the architecture.

3.1. Cross-Attention Architecture

Using BEATs [11] with frozen encoders, we combine channel embeddings $\mathbf{e}_0, \mathbf{e}_1 \in \mathbb{R}^{768}$ as $\mathbf{E} = [\mathbf{e}_0; \mathbf{e}_1]$. Multi-head attention [15] computes, $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{E}\mathbf{W}_{Q,K,V}$ with learned projections $\mathbf{W}_{Q,K,V} \in \mathbb{R}^{768 \times 768}$.

The enhanced MLP processes attended embeddings $[\mathbf{e}'_0; \mathbf{e}'_1]$ through:

$$\begin{aligned} \mathbf{h}_1 &= \text{GELU}(\text{LayerNorm}(\mathbf{W}_1[\mathbf{e}'_0; \mathbf{e}'_1] + \mathbf{b}_1)), \\ \mathbf{h}_2 &= \text{GELU}(\text{LayerNorm}(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2)) + \mathbf{h}_1, \\ y &= \mathbf{W}_4(\text{GELU}(\text{LayerNorm}(\mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3))) + \mathbf{b}_4, \end{aligned}$$

with layer dimensions $256 \rightarrow 128 \rightarrow 64 \rightarrow 1$ and residual connections.

3.2. Training Procedures

Keypoint Sampling: To avoid temporal clustering in random sampling, we select every 20th keypoint per file. This covers 5% of keypoints uniformly while maintaining efficiency.

Conservative Data Augmentation: Applied to 30% of training samples with channel synchronization:

Table 1: BioDCASE 2025 Challenge Results (Test Set)

Method	ARU	Zebra Finch	Average
Nosync	0.850	1.840	1.350
Crosscorr	1.100	5.550	3.320
DL Baseline	0.620	0.550	0.580
Ours (BEATsCA)	0.14	0.45	0.30
Improvement vs DL	77.4%	18.2%	48.3%

- *Amplitude scaling*: Random factor $\alpha \sim \mathcal{U}(0.9, 1.1)$ applied identically to both channels
- *Gaussian noise*: $\text{SNR} \in [40, 50]$ dB
- *Consistency constraint*: Identical random seed ensures synchronized augmentation across channels

The augmentation preserves temporal alignment by applying identical transformations:

$$\begin{aligned}\tilde{S}_{i,0} &= \alpha \cdot S_{i,0} + \mathcal{N}(0, \sigma^2), \\ \tilde{S}_{i,1} &= \alpha \cdot S_{i,1} + \mathcal{N}(0, \sigma^2),\end{aligned}$$

where σ^2 depends on target SNR. This preserves temporal alignment while improving robustness.

3.3. Confidence-Weighted Scoring

Traditional methods use binary counting, simply summing the number of aligned predictions ($\sum_i \mathbf{1}[f_\theta(S_{i,0}, S_{i,1}) > 0]$) discarding valuable confidence information. A prediction of 0.01 is treated identically to 0.99, despite vastly different certainty levels. We replace binary counting with a scoring function that incorporates prediction distributions:

$$S_{conf}(\mathcal{K}_j) = 0.4\mu_{pos}r_{pos} + 0.3\mu_{top} + 0.2\sum_i \sigma(p_i^{(j)}) + 0.1\mathcal{E}_{exp} \quad (2)$$

The weights (0.4, 0.3, 0.2, 0.1) were chosen based on intuitive design principles: highest weight for positive prediction quality ($\mu_{pos} \cdot r_{pos}$) as it directly measures alignment confidence, second-highest for top quartile focus (μ_{top}) to emphasize reliable predictions, and lower weights for the supporting probabilistic and exponential components. The components:

- *Positive Confidence Weighting* ($\mu_{pos} \cdot r_{pos}$): Average confidence of positive predictions weighted by their prevalence
- *Top Quartile Focus* (μ_{top}): Mean confidence of highest-scoring 25% predictions
- *Probabilistic Coverage* ($\sum_i \sigma(p_i^{(j)})$): Sum of sigmoid-transformed probabilities
- *Exponential Amplification* (\mathcal{E}_{exp}): $e^{4(p-0.5)}$ for $p > 0.5$, else 0

Figure 3 illustrates the scoring approach with an example.

3.4. Inference Pipeline

Candidate Generation: Following the baseline, we assume affine drift $\mathcal{D}(t) \approx \alpha t + \beta$ and sample parameters: $\alpha_j \in [1 \pm 5/T_{dur}]$, $\beta_j \in [-5, 5]$, generating candidates $\hat{k}_{i,1}^{(j)} = \alpha_j \cdot i + \beta_j$.

Selection: For each candidate \mathcal{K}_j : (1) extract 2-second segments at predicted locations, (2) compute cross-attention scores, (3) calculate $S_{conf}(\mathcal{K}_j)$, (4) select maximum-scoring candidate.

4. EXPERIMENTAL SETUP

4.1. Datasets and Evaluation Protocol

We use BioDCASE 2025 Task 1 datasets [1] containing stereo audio with temporal keypoints in two domains:

- **ARU dataset:** 36 training files, 12 validation files from field recordings

Table 2: Validation Set Performance Comparison

Method	ARU	Zebra Finch	Average
Nosync	0.976	1.315	1.146
Crosscorr	6.861	10.029	8.445
DL Baseline	0.516	1.262	0.889
Ours	0.099	0.521	0.310
Improvement vs DL	80.8%	58.7%	65.1%

- **Zebra finch dataset:** 108 training files, 16 validation files from laboratory recordings

Both datasets contain keypoints at 1-second intervals with drift constrained to ± 5 seconds. The challenge evaluates using MSE (Equation 1) between predicted and ground-truth Channel 1 timestamps, with final scoring computed as the average MSE across both test datasets. All training uses only the provided data without external resources, ensuring fair comparison with baseline methods.

4.2. Implementation Details

Model Architecture: We use the pre-trained BEATs encoder (BEATs_iter3_plus_AS2M_finetuned_on_AS2M_cpt2.pt) with frozen parameters during training. The cross-attention module employs 8 attention heads with 768-dimensional embeddings. The enhanced MLP processes 1536-dimensional concatenated embeddings through layers of dimensions $256 \rightarrow 128 \rightarrow 64 \rightarrow 1$ with GELU activations, layer normalization, and residual connections.

Training Configuration: We train for up to 100 epochs with early stopping (patience 25) using AdamW optimizer with learning rate 2×10^{-4} , weight decay 0.01, and ReduceLROnPlateau scheduling (factor 0.7, patience 3). Batch size is set to 32 to prevent $O(B^2)$ memory scaling in the pairwise training loss. Conservative data augmentation is applied to 30% of samples with amplitude scaling ($\pm 10\%$) and additive noise (40-50 dB SNR).

Inference Setup: During inference, we sample 100 candidate drift parameters within constraint bounds. For each candidate, we extract 2-second audio segments, compute cross-attention predictions, and select the candidate with maximum confidence-weighted score.

Hardware and Reproducibility: Training is performed on NVIDIA H100 GPU with CUDA optimization and memory management (cache clearing every 5 batches). All experiments use fixed random seeds for reproducibility.

5. RESULTS

5.1. Competition Performance

Our method achieved 0.30 MSE on BioDCASE 2025 test data (Table 1), improving upon the deep learning baseline by 48.3%.

5.2. Validation Set Analysis

Validation results (Table 2) demonstrate consistent improvements across both datasets, with particularly strong performance on ARU data (0.099 MSE vs. 0.516 baseline).

The validation-to-test performance correlation (validation: 0.310, test: 0.30) indicates robust generalization with minimal overfitting, validating our conservative training approach.

6. ABLATION STUDIES

We conduct comprehensive ablation studies to understand component contributions. Results show that both architectural innovations and confidence-weighted scoring contribute to performance improvements.

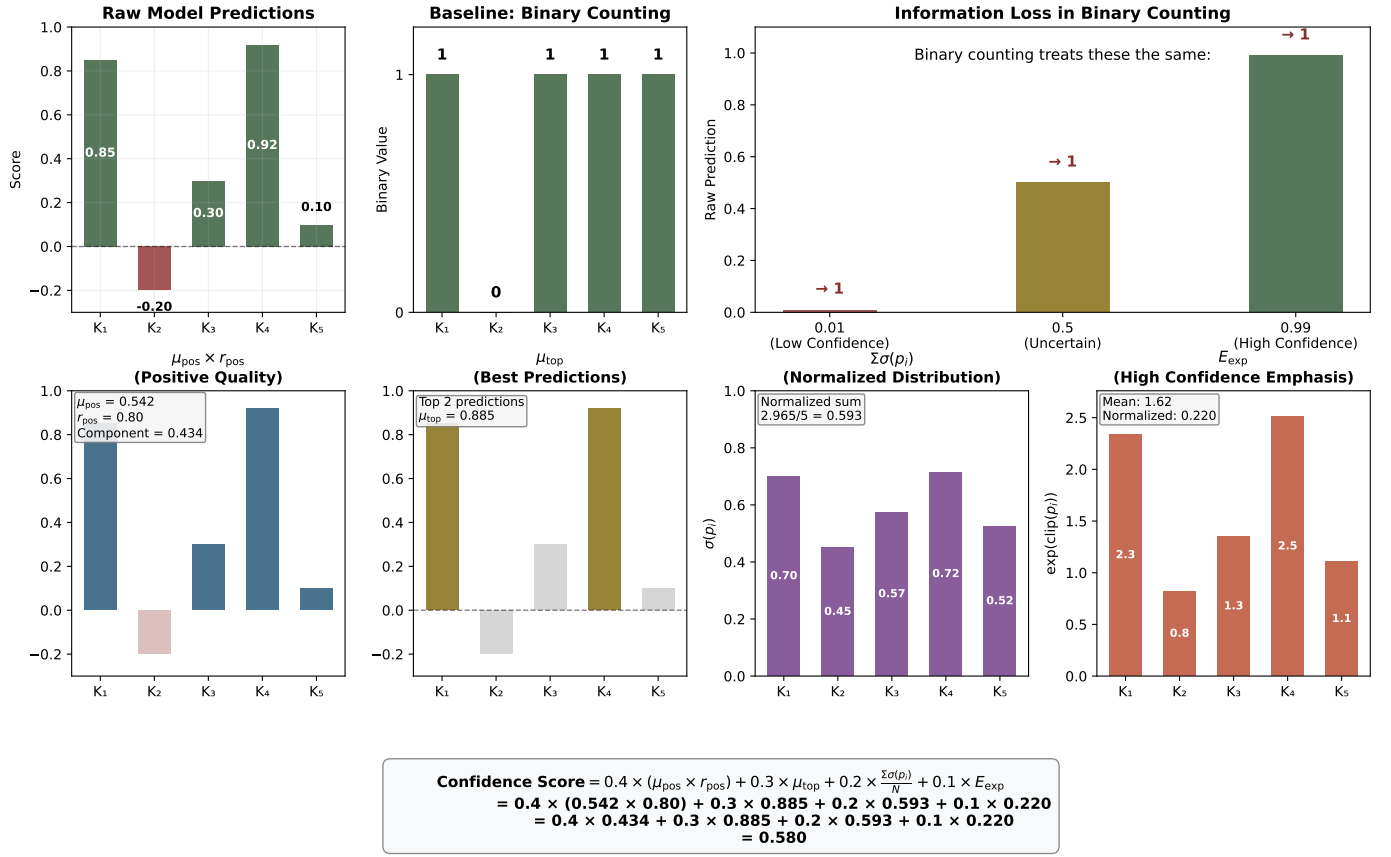


Fig. 3: Confidence-weighted scoring methodology for multi-channel audio alignment. **Top row:** Raw model predictions (left) are converted to binary values (0.80 average) in the baseline approach (center), demonstrating information loss where predictions with vastly different confidence levels are treated identically (right). **Middle row:** Our confidence-weighted approach incorporates four complementary components: positive prediction quality ($\mu_{\text{pos}} \times r_{\text{pos}} = 0.434$), top quartile averaging ($\mu_{\text{top}} = 0.885$), normalized sigmoid distribution ($\sum \sigma(p_i)/N = 0.593$), and exponential emphasis of high-confidence predictions ($E_{\text{exp}} = 0.220$). **Bottom:** The weighted combination (0.4, 0.3, 0.2, 0.1 weights respectively) produces a confidence score of 0.580, demonstrating how confidence weighting utilizes the full prediction distribution compared to binary counting for alignment candidate selection.

Table 3: Progressive Component Addition Analysis (Validation MSE)

Configuration	ARU MSE	ZF MSE	ARU Δ from Baseline	ZF Δ from Baseline
Original Baseline	0.516	1.262	—	—
+ Enhanced MLP	0.380	1.050	26%	17%
+ Cross-Attention	0.152	0.680	71%	46%
+ Confidence Scoring	0.099	0.521	81%	59%

Table 4: Confidence Scoring Weight Ablation

Weight Config	μ_{pos}	μ_{top}	$\sum \sigma(p_i)$	e^μ	ARU MSE	ZF MSE
Proposed	0.400	0.300	0.200	0.100	0.099	0.521
Mean+Top	0.500	0.500	0.000	0.000	0.161	0.673
Mean only	1.000	0.000	0.000	0.000	0.161	0.673
Equal weights	0.250	0.250	0.250	0.250	0.162	0.718
Sigmoid+Exp	0.000	0.000	0.500	0.500	0.162	0.718
Top only	0.000	1.000	0.000	0.000	0.224	0.754

6.1. Component Contribution Analysis

Table 3 shows the incremental improvement from baseline to our final method through systematic component addition. Cross-attention provides the largest single improvement, while confidence scoring adds substantial refinement.

6.2. Confidence Scoring Weight Analysis

Table 4 examines different weight configurations for the confidence scoring function (Equation 2). Our weighting (0.4, 0.3, 0.2, 0.1) performs best, with μ_{pos} and μ_{top} providing primary benefits.

7. CONCLUSION

This work demonstrates that cross-attention mechanisms and confidence-weighted scoring can significantly improve multi-channel audio alignment. By modeling inter-channel temporal dependencies and utilizing the full prediction distribution rather than binary thresholding, our approach achieved first place in the BioDCASE 2025 challenge with substantial improvements over existing methods. The key contributions include applying cross-attention to model temporal relationships between channels and developing a principled confidence estimation framework that provides uncertainty quantification for alignment decisions. The experimental validation across ARU and zebra finch datasets confirms the effectiveness of both architectural innovations. While the current approach relies on affine drift approximations for candidate generation, the demonstrated performance gains suggest that probabilistic alignment frameworks represent a promising direction for the field. Future work should investigate principled optimization of confidence scoring weights through techniques such as Bayesian optimization or gradient-based hyperparameter tuning. Additionally, the confidence scoring framework could be extended with learned weighting schemes that adapt to different acoustic domains or recording conditions. The method applies to domains needing precise synchronization, including distributed sensor networks and spatial audio systems.

REFERENCES

- [1] B. Hoffman, L. Gill, B. Heath, and G. Narula, “BioDCASE 2025 Task 1: Multi-Channel Alignment,” BioDCASE Challenge, 2025.
- [2] M. Brandstein and D. Ward, Eds., *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.
- [3] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, “An improved algorithm for blind reverberation time estimation,” in *2018 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 206–210.
- [4] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
- [5] M. Müller, *Information retrieval for music and motion*. Springer, 2007.
- [6] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [7] S. Salvador and P. Chan, “FastDTW: Toward accurate dynamic time warping in linear time and space,” *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
- [8] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Proceedings of INTERSPEECH 2017*, 2017, pp. 498–502.
- [9] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” in *Proceedings of INTERSPEECH 2021*, 2021, pp. 571–575.
- [10] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Proceedings of INTERSPEECH 2020*, 2020, pp. 5036–5040.
- [11] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, “BEATs: Audio Pre-Training with Acoustic Tokenizers,” in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193.
- [12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.
- [13] M. Hamada, “Fighting against uncertainty: an essential issue in bioinformatics,” *Briefings in Bioinformatics*, vol. 15, no. 5, pp. 748–767, 2014.
- [14] O. Penn, E. Privman, H. Ashkenazy, G. Landan, D. Graur, and T. Pupko, “GUIDANCE: a web server for assessing alignment confidence scores,” *Nucleic Acids Research*, vol. 38, no. suppl_2, pp. W23–W28, 2010.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.