Vocal Unit Segmentation in White-Faced Capuchins Using Joint Multi-Species Embeddings and BIO-Style Span Detection

Anonymous ACL submission

Abstract

Much of the progress to date in understanding animal communication has relied on the ability to segment vocalizations into meaningful units, which so far has been primarily a manual process. We introduce an automatic approach for vocal-unit segmentation inspired by a human language BIO-style span-detection task, and demonstrate its effectiveness for the segmentation of white-faced capuchin Twitter calls. We compare frame classifiers, BiLSTM-CRF decoders, established baselines, and pre-trained bird-bioacoustic and human-speech embeddings, and find that structured decoding has the largest effect on unit recovery, with joint multi-species embeddings adding complementary signal when fused with spectrogram representations. The resulting system reaches a Unit F1 of 0.683 and exceeds human inter-annotator agreement on a doubly annotated subset. Our results contribute practical tools for within-call structure modeling in data-scarce wild bioacoustic settings and position NLP methods as scientific instruments for studying communication beyond human language.

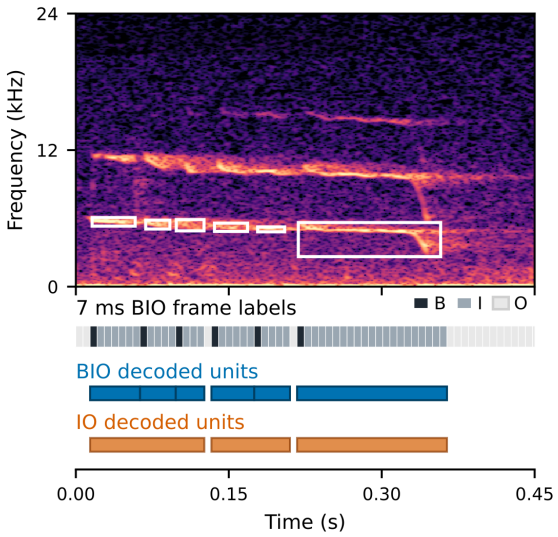

Figure 1: Manual vocal-unit annotations and BIO encoding. Raven annotations are projected onto a 7 ms frame grid: B marks unit onset, I frames inside a unit, and O background. BIO preserves adjacent units that IO would merge.

1 Introduction

The computational study of animal vocalizations has recently shifted from detecting which species are present toward understanding the information the calls carry (Sharma et al., 2024). A central question is how calls are internally structured: many consist of sequences of shorter acoustic elements whose organization may carry behavioral, individual, or communicative information (Suzuki, 2021; Odom et al., 2021; Berthet et al., 2023). Modeling within-call structure is therefore a foundational step for downstream questions in animal communication research, from testing hypotheses about linguistic-like properties of animal signals (Suzuki et al., 2017) to supporting conservation, where syllable- and song-level change can signal population decline in endangered species (Crates et al., 2021; Jepson et al., 2026; Baker and Logue, 2003).

We focus on the case of white-faced capuchin Twitter calls: like marmoset twitter calls, whose syllables are tracked by phase-locked frontal-cortex activity (Zhao and Wang, 2023), they are structured sequences rather than atomic signals. They also carry social information, with field evidence linking them to group movement coordination (Boinski and Campbell, 1995) and ontogenetic shifts in usage across age classes (Guisneuf et al., 2025), suggesting a flexible, developmentally shaped call.

Existing work on vocal-unit segmentation has relied largely on spectrogram heuristics such as blob detection, median clipping, and harmonic tracking (Fukuzawa et al., 2016). The small annotated datasets typical of wild primate recordings make supervised approaches difficult (Cauzinille

et al., 2024), with deep-learning pipelines developed only for particular species (Schröter et al., 2019; Martínez-Ortega and Sarria-Paja, 2021).

At the same time, the NLP community has developed mature methods for detecting spans in sequences. A standard formulation is BIO tagging (Ramshaw and Marcus, 1995), which labels each token as the *beginning* (B) or *inside* (I) of a span or *outside* (O) any span, so that contiguous B–I runs decode into spans; richer variants add explicit *end* (E) and *singleton* (S) tags (Ratinov and Roth, 2009). Paired with BiLSTM and Conditional Random Fields (CRF) sequence models, this approach is a strong default for span tasks such as named entity recognition (NER) (Ma and Hovy, 2016). To our knowledge, this approach has not been systematically applied to bioacoustic segmentation. Within-call units pose a representational problem that whole-call detection does not: adjacent units can abut with no intervening silence, so a model must mark boundaries, not merely distinguish vocalization from background. The simpler IO scheme is easy to learn but cannot separate touching spans, whereas boundary-aware schemes preserve that structure at the cost of rarer, harder-to-learn boundary tags. In NLP, that difficulty was resolved not by richer labels alone but by structured decoders such as CRFs that enforce valid tag transitions (Lample et al., 2016; Ma and Hovy, 2016). No existing neural segmenter for animal vocalizations applies structured decoding over boundary-aware tags: TweetyNet (Cohen et al., 2022) decodes IO-style runs of non-background frames with heuristic post-processing, and WhisperSeg (Gu et al., 2024) generates boundary tokens autoregressively.

In this paper, we use a dataset of white-faced capuchin monkey Twitter vocalizations, collected in wild field conditions, and show that these methods can be effectively used for vocal-unit segmentation for animal vocalization. The paper makes two main contributions. First, we formulate vocal unit segmentation as BIO-style acoustic span detection at 7 ms resolution, evaluating frame-level classifiers and structured BiLSTM–CRF models against established baselines and human inter-annotator agreement, scoring every system on the same recovered units. Second, we test whether pretrained representations from human speech and bird bioacoustics can contribute to the task, comparing single embeddings with joint multi-species representations fused via concatenation, summation, and a feature-selection method adapted from bioinformatics (Ding and Peng, 2005). As previous work has found found that these fusion strategies improve acoustic individual identification in capuchins Vega-Hidalgo et al. (2025), we test whether they also support within-call structure.

2 Vocalization Dataset

The Twitter recordings we use in this work are obtained from a focal-recording corpus of wild white-faced capuchins in the Taboga Reserve, Costa Rica, originally introduced by Vega-Hidalgo et al. (2025) for acoustic individual identification; we refer the reader to that work for field protocols, recording conditions, and per-individual call counts. We use a Twitter-only subset (chosen for the sequence-like structure of Twitter calls; Fig. 1) for which we produced vocal-unit annotations. An experienced bioacoustic researcher (six years of annotation) labeled 775 vocal units across 140 Twitter files in Raven Pro (K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology, 2024), averaging 5.54 units per file (SD 2.20; range 1–13). Units are short and variable, with a mean duration of $34.2\,\mathrm{ms}$ (SD $23.1\,\mathrm{ms}$; range $6.4\,\mathrm{ms}$–$364\,\mathrm{ms}$). For inter-annotator analysis and frame-size selection, a novice annotator independently re-labeled a 35-file subset (385 units) after training on held-out examples.

Frame-size selection. We sweep frame sizes from 2 to 8 ms, since larger windows become too coarse for fine-grained BIO labeling given these unit durations. On the doubly annotated subset, we compared candidate sizes by Cohen's κ (Cohen, 1960; Artstein and Poesio, 2008) and macro F1 (Fig. 2), selecting $7\,\mathrm{ms}$ as the best compromise between temporal precision and inter-annotator agreement.

Frame-label construction. We project the Raven unit annotations onto a 7 ms frame grid, labeling each unit's onset frame B, its remaining frames I, and background O. From these canonical BIO labels we derive four tagging schemes (IO, BIO, BIOE, BIOES), since added boundary tags can improve span-level reconstruction (Ratinov and Roth, 2009). IO is the exception: when two units abut with no background frame between them, it cannot mark the boundary, so projecting our BIO gold to IO merges 775 units into 711 and caps Unit F1 at 0.898. We therefore use the BIO-family

schemes for boundary-aware unit recovery, reporting TweetyNet under its own strongest scheme, IO, with this caveat.

3 Vocal-unit Segmentation

We organize our segmentation systems along two axes: the input representation and the decoder. We consider two kinds of representations. The first is a set of pretrained audio embeddings: the two encoders found in previous work to perform best for acoustic individual identification in capuchins (Vega-Hidalgo et al., 2025), which include Perch 2 (van Merriënboer et al., 2025) (animal sounds, mainly pretrained on birds) and Whisper V3 Large (Radford et al., 2022) (human speech), used here as frame-level features to test whether multi-species acoustic representations supporting whole-call identity also support within-call structure. We use each encoder by itself, and also combine them through the three joint multi-species fusions from that work: concatenation, summation, and MRMR (Ding and Peng, 2005); we refer to any of these five embedding choices as an *embedding configuration*.

The second representation is a learned spectrogram CNN front end, trained directly on the capuchin segmentation task. It converts the spectrogram into frame-aligned acoustic features for BIO decoding, following the spectrogram-based segmentation paradigm of TweetyNet (Cohen et al., 2022). In our system, however, these frame features are decoded with a BiLSTM–CRF, replacing TweetyNet's heuristic post-processing with structured decoding.

For the decoder we compare a *frame classifier*, which labels each frame independently, with a *BiLSTM–CRF*, which adds a recurrent encoder and a CRF layer that enforces valid tag transitions across the sequence. We run the frame classifier over each embedding configuration as a non-structured baseline. Under the CRF we obtain two single-representation systems: *Emb-CRF*, run over each embedding configuration, and *Spec-CRF*, run over the spectrogram front end. We then combine the two CRF branches in two ways: *score fusion* combines the two branches' CRF scores (emissions and transitions) by a convex weight α tuned on validation macro F1, and *hidden fusion* concatenates their hidden states before a shared CRF. Finally, we compare against two external baselines: *TweetyNet* (Cohen et al., 2022), a spectrogram CNN that decodes frames with heuristic post-processing rather than a CRF, isolating the effect of structured decoding from the spectrogram front end; and *WhisperSeg* (Gu et al., 2024), which generates boundary tokens autoregressively, in zero-shot and fine-tuned settings.

Experimental Setup. Models are trained by audio file under a *day-disjoint* split that holds out entire recording days, so no day appears in both training and test, preventing acoustic leakage. This experimental setup addresses any potential confounds and shortcut learning due to background noise (e.g., forest noise), as it explicitly avoids training and test data drawn from the same environment. We used 50 random train-val splits of the training days, trained with Adam and early stopping on validation macro F1; frame classifiers additionally used inverse-frequency class weighting. Unit-level scoring uses one common gold standard decoded from the manual BIO-style annotations, so every tagging scheme is evaluated against the same true units. Architecturally, frame classifiers compress each embedding configuration to 64 dimensions before the output layer, while the CRF systems use a BiLSTM encoder over the input representation followed by an emission head and CRF layer; full per-system dimensions are in Appendix A. The main text reports the primary comparison (Table 1) and paired tests against our best system (Table 4); the full grid over all tagging schemes and decoders is in Appendix Table 3.

Evaluation Metrics. We score segmentation at both frame and unit levels. Frame-level Macro F1 compares predictions and gold labels under the active scheme's label inventory. For unit-level scoring, frame predictions are decoded into spans (e.g. `B I I O B I` yields two units) and matched to gold spans decoded from the manual BIO annotations. A predicted span matches a gold span when both boundaries fall within a tolerance of t frames; we report Unit F1 averaged over $t \in \{0, 1, 2, 3\}$ frames (0–21 ms at 7 ms resolution) to avoid sensitivity to a single boundary threshold. Unit F1 is computed per file from span precision and recall, then averaged across files. Count MAE is the mean per-file absolute difference between predicted and annotated unit counts, capturing over- and under-segmentation. All values are mean $\pm$ SD over the 50 train-val splits.

System	Scheme	Emb.	Unit F1↑	Seg. err.↓	Count MAE↓
WhisperSeg (fine-tuned)	BIOE	—	0.442±0.219	0.739±0.150	2.772±1.724
Raw TweetyNet	IO	—	0.633±0.024	0.674±0.036	0.789±0.107
Spec-CRF	BIOE	—	0.665±0.020	0.608±0.032	0.737±0.106
Emb-CRF	BIOE	P	0.610±0.026	0.690±0.036	1.164±0.152
Spec+Emb (score fusion)	BIOE	Concat(W,P)	0.681±0.021	**0.583±0.028**	0.776±0.131
Spec+Emb (hidden fusion)	BIOE	Concat(W,P)	**0.683±0.024**	0.588±0.035	**0.679±0.111**

Table 1: Primary system comparison, reported as mean ± SD over 50 train-val splits on the fixed day-disjoint test set. CRF systems are reported under BIOE, their strongest scheme; baselines are reported under their strongest scheme. *Spec* denotes the spectrogram branch, *Emb* the embedding branch, *W* denotes Whisper, and *P* denotes Perch. *Score* and *hidden* fusion indicate how the spectrogram and embedding CRF branches are combined.

4 Results

Table 1 reports the main system comparison. Our structured BiLSTM–CRF systems recover vocal units more accurately than both baselines: the best system, Spec+Emb with hidden-state fusion, reaches Unit F1 0.683, ahead of the spectrogram-CNN baseline TweetyNet (0.633) and well ahead of WhisperSeg (0.442). The spectrogram-only CRF (Spec-CRF, 0.665) is already strong on its own.

The largest effect is architectural. Replacing the frame-level classifier with a CRF decoder over the same embeddings more than doubles Unit F1 (Appendix Table 3: under BIOE, Perch rises from 0.247 with the frame classifier to 0.610 with the CRF). By enforcing valid tag transitions over the sequence, structured decoding converts noisy frame predictions into well-formed spans.

The favored scheme differs by architecture (Appendix Table 3). The frame-level classifier is strongest under IO and degrades sharply under boundary-aware schemes (Perch Unit F1 0.336 at IO vs. 0.247 at BIOE), whereas the CRF exploits the explicit onset/offset structure of the boundary-aware schemes (Perch 0.527 at IO vs. 0.610 at BIOE). We therefore report our CRF systems under BIOE, their strongest scheme, and TweetyNet under IO, its strongest.

The spectrogram front end is the strongest single representation: Spec-CRF reaches Unit F1 0.665, outperforming Emb-CRF with Perch (0.610), while Whisper alone performs poorly in this setting (0.020). Fusing embeddings with Spec-CRF adds a small but reliable gain: score fusion reaches 0.681 and hidden fusion reaches 0.683, with hidden fusion significantly above Spec-CRF ($\Delta = 0.017$, $p = 4.8\times10^{-4}$; Table 4) but not significantly different from score fusion ($\Delta = 0.002$, $p = 0.65$). The pretrained embeddings are therefore weak as standalone inputs but complementary when fused with the spectrogram branch.

Human inter-annotator agreement provides a label-consistency reference, not an accuracy ceiling. On the 25-file doubly-annotated subset, human BIOE Unit F1 is 0.611 (Table 2), while our best model reaches 0.650 ± 0.032 on that same subset. The model matches or exceeds human agreement partly because it was trained against one of the two annotators (the expert) and reproduces that annotator's boundary placements at 7 ms resolution, whereas the two human annotators were not calibrated to each other at that resolution. Human agreement here reflects between-annotator variability rather than a performance ceiling.

5 Conclusion

We addressed the manual bottleneck in animal vocal-unit annotation by formulating white-faced capuchin Twitter-call segmentation as an NLP-inspired BIO-style span-detection task. This formulation lets us separate two factors in unit recovery: the decoder that turns frame predictions into spans and the acoustic representation being decoded. Three empirical insights follow. First, structured BiLSTM–CRF decoding is the dominant factor: it substantially outperforms independent frame classification, and boundary-aware schemes (BIO/BIOE/BIOES) help only when paired with the CRF. Second, the learned spectrogram front end is stronger than any single pretrained embedding for recovering fine-grained within-call units. Third, joint multi-species embeddings from bird bioacoustics and human speech add complementary signal when fused with the spectrogram branch. These results offer practical tools for within-call structure modeling in data-scarce wild bioacoustics and show how NLP methods can serve as scientific instruments for studying communication beyond human language.

6 Limitations

The main limitation is annotation scale and scope. To our knowledge, this is the first vocal-unit segmentation dataset for wild white-faced capuchins, but it remains small for modern sequence-labeling models. This scale reflects the cost of expert annotation: labeling the dataset required roughly 20 hours of annotator time, and only 25 day-disjoint files were doubly annotated. A firmer human reference would require multi-expert consensus labels, since strict span scores at 7 ms are sensitive to one- or two-frame boundary shifts. Model performance near human agreement should therefore be read as consistency with the expert protocol, not superhuman segmentation. Because we study only capuchin Twitter calls, generalization to other call types, populations, or taxa remains untested.

The models also leave open what structure is recovered. We did not exhaustively explore the architecture, loss, or fusion design space, so stronger variants may exist. The weak standalone performance of pretrained embeddings, especially Whisper, reflects this fine-grained setting rather than general unsuitability for bioacoustics. Finally, our acoustic units are modeling targets for within-call structure, not linguistic syllables; their functional significance requires further biological validation.

7 Ethical Considerations

No animals were harmed during this study. Field data collection followed applicable animal-welfare guidelines and prioritized non-invasive observation of wild capuchins. By developing methods for analyzing vocal communication from recordings, this work supports non-invasive approaches to studying species affected by habitat loss and environmental change. All data collection, annotation, and analysis were conducted by researchers or formally employed personnel under appropriate academic and labor practices.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Myron C Baker and David M Logue. 2003. Population differentiation in a complex bird sound: a comparison of three bioacoustical analysis procedures. *Ethology*, 109(3):223–242.

Mélissa Berthet, Camille Coye, Guillaume Dezecache, and Jeremy Kuhn. 2023. Animal linguistics: a primer. *Biological reviews*, 98(1):81–98.

Sue Boinski and Aimee F. Campbell. 1995. Use of Trill Vocalizations To Coordinate Troop Movement Among White-Faced Capuchins: a Second Field Test. *Behaviour*, 132(11-12):875–901.

Jules Cauzinille, Benoit Favre, Ricard Marxer, and Arnaud Rey. 2024. Applying machine learning to primate bioacoustics: Review and perspectives. *American Journal of Primatology*, 86(10):e23666.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yarden Cohen, David Aaron Nicholson, Alexa Sanchioni, Emily K Mallaber, Viktoriya Skidanova, and Timothy J Gardner. 2022. Automated annotation of birdsong with a neural network that segments spectrograms. *Elife*, 11:e63853.

Ross Crates, Naomi Langmore, Louis Ranjard, Dejan Stojanovic, Laura Rayner, Dean Ingwersen, and Robert Heinsohn. 2021. Loss of vocal culture and fitness costs in a critically endangered songbird. *Proceedings of the Royal Society B: Biological Sciences*, 288(1947):20210225.

Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.

Yukio Fukuzawa, Stephen Marsland, Matthew Pawley, and Andrew Gilman. 2016. Segmentation of harmonic syllables in noisy recordings of bird vocalisations. In *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE.

Nianlong Gu, Kanghwi Lee, Maris Basha, Sumit Kumar Ram, Guanghao You, and Richard HR Hahnloser. 2024. Positive transfer of the whisper speech transformer to human and animal voice activity detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7505–7509. IEEE.

Nicole Guisneuf, Juan Carlos Ordoñez, Marcela E Benítez, and Thore J Bergman. 2025. The ontogeny of "twitter" calls in white-faced capuchins (cebus imitator): Usage, context, and acoustic structure. *American Journal of Primatology*, 87(8):e70069.

Oliver Jepson, RT Gilman, Leah J Williams, and Rebecca N Lewis. 2026. Shifting syllable production in an ex situ population of a critically endangered songbird. *Zoo Biology*, 45(1):29–36.

K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology. 2024. Raven pro: Interactive sound analysis software (version 1.6.5). [Computer software].

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Sara Virginia Martínez-Ortega and Milton Sarria-Paja. 2021. Automatic segmentation and classification of the thamnophilidae's family of the western andean region of colombia. In *2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, pages 1–6. IEEE.

Karan J. Odom, Marcelo Araya-Salas, Janelle L. Morano, Russell A. Ligon, Gavin M. Leighton, Conor C. Taff, Anastasia H. Dalziell, Alexis C. Billings, Ryan R. Germain, Michael Pardo, Luciana Guimarães de Andrade, Daniela Hedwig, Sara C. Keen, Yu Shiu, Russell A. Charif, Michael S. Webster, and Aaron N. Rice. 2021. Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews*, 96(4):1135–1159.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third workshop on very large corpora*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Hendrik Schröter, Elmar Nöth, Andreas Maier, Rachael Cheng, Volker Barth, and Christian Bergler. 2019. Segmentation, classification, and visualization of orca calls using deep learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8231–8235. IEEE.

Pratyusha Sharma, Shane Gero, Daniela Rus, Antonio Torralba, and Jacob Andreas. 2024. Whalelm: Finding structure and information in sperm whale vocalizations and behavior with machine learning. *bioRxiv*, pages 2024–10.

Toshitaka N Suzuki. 2021. Animal linguistics: exploring referentiality and compositionality in bird calls. *Ecological Research*, 36(2):221–231.

Toshitaka N Suzuki, David Wheatcroft, and Michael Griesser. 2017. Wild birds use an ordering rule to decode novel call sequences. *Current Biology*, 27(15):2331–2336.

Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. 2025. Perch 2.0: The bittern lesson for bioacoustics. *arXiv preprint arXiv:2508.04665*.

Álvaro Vega-Hidalgo, Artem Abzaliev, Thore Bergman, and Rada Mihalcea. 2025. Acoustic individual identification of white-faced capuchin monkeys using joint multi-species embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 645–659, Vienna, Austria. Association for Computational Linguistics.

L Zhao and X Wang. 2023. Frontal cortex activity during the production of diverse social communication calls in marmoset monkeys. nat. commun. 14, 6634.

A Appendix

Per-system architecture dimensions Neural systems use a tag-emission head and, for CRF systems, a linear-chain CRF. *Frame classifier*: each single-encoder embedding is projected to 512 dimensions, then to a 64-dimensional hidden layer before the per-frame output; concatenation and summation instead project each encoder to 256 dimensions and fuse them (concat 512, sum 256) before that hidden layer. *Emb-CRF*: each single-encoder embedding is projected to 512 dimensions, encoded by a bidirectional LSTM (hidden size 128 per direction, 256-dimensional output), then passed through a 256-to-128 emission MLP before the CRF; concatenation and summation again project each encoder to 256 dimensions and fuse them (concat 512, sum 256) before the shared BiLSTM. *Spec-CRF*: a spectrogram CNN front end reduces 513 frequency bins to 8 pooled bins and produces 512-dimensional frame features, which a bidirectional LSTM (hidden size 256 per direction, 512-dimensional output) encodes before the emission layer and CRF. *Score fusion* adds no trainable parameters: it combines the trained Emb-CRF and Spec-CRF scores (emissions, CRF transitions, and start/end transition scores) by a convex weight α selected on validation macro F1. *Hidden fusion*: the embedding branch (embedding views projected to 256; BiLSTM hidden 256 per direction, output 512) and the spectrogram branch (CNN features 512; BiLSTM hidden 256 per direction, output 512) are concatenated to 1024 dimensions, reduced to 512, and decoded by a single shared CRF.

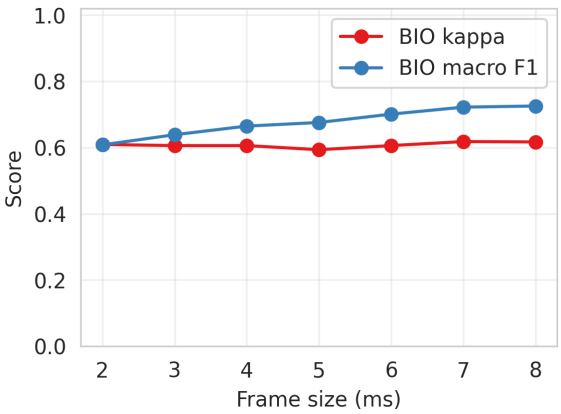

Figure 2: Inter-annotator agreement across candidate BIO frame sizes on the doubly annotated subset. Cohen's κ and BIO macro F1 both improve as frame size increases from 2 to 8 ms.

Scheme	Unit F1	Seg. err.↓	Count MAE↓	Frame F1
IO	0.379	0.931	2.600	0.819
BIO	0.611	0.725	0.680	0.737
BIOE	0.611	0.725	0.680	0.611
BIOES	0.611	0.725	0.680	0.490

Table 2: Human inter-annotator agreement by tagging scheme on the 25-file doubly-annotated subset. Unit-level metrics are scheme-invariant (recovered units do not depend on tag alphabet) except IO, which merges abutting units; frame-level agreement decreases as the label set grows.

Scheme	System	View	Unit F1 (tol0–3)	Seg. err. ↓	Count MAE ↓	Frame F1
IO	Raw TweetyNet	Spectrogram	**0.633 ±0.024**	0.674 ±0.036	**0.789 ±0.107**	0.853 ±0.010
	WhisperSeg	Zero-shot	0.020 ±0.000	1.064 ±0.000	5.324 ±0.000	0.556 ±0.000
		Fine-tuned	0.441 ±0.219	0.740 ±0.149	2.777 ±1.722	0.663 ±0.177
	Frame	Whisper	0.271 ±0.070	1.350 ±0.109	2.532 ±0.560	0.666 ±0.058
		Perch	0.336 ±0.035	1.332 ±0.053	1.969 ±0.243	0.689 ±0.012
		Concat	0.338 ±0.053	1.259 ±0.057	2.095 ±0.357	0.704 ±0.018
		Sum	0.335 ±0.050	1.243 ±0.080	2.168 ±0.353	0.704 ±0.018
		MRMR	0.375 ±0.027	1.178 ±0.044	2.024 ±0.154	0.731 ±0.005
	Emb-CRF	Whisper	0.009 ±0.045	1.146 ±0.027	4.951 ±0.370	0.612 ±0.044
		Perch	0.527 ±0.062	0.795 ±0.066	1.435 ±0.359	0.807 ±0.018
		Concat	0.529 ±0.087	0.787 ±0.076	1.454 ±0.535	0.804 ±0.026
		Sum	0.534 ±0.047	0.789 ±0.053	1.417 ±0.266	0.805 ±0.013
		MRMR	0.513 ±0.044	0.852 ±0.040	1.434 ±0.228	0.806 ±0.011
	Spec-CRF	Spectrogram	0.630 ±0.025	0.665 ±0.035	0.884 ±0.147	0.861 ±0.009
	Spec+Emb score-fusion CRF	Whisper aux.	0.617 ±0.025	0.677 ±0.032	0.962 ±0.168	0.859 ±0.008
		Perch aux.	0.618 ±0.030	0.668 ±0.036	0.982 ±0.176	0.863 ±0.009
		Concat aux.	0.623 ±0.027	**0.663 ±0.034**	0.975 ±0.164	0.864 ±0.009
		Sum aux.	0.621 ±0.027	0.666 ±0.034	0.975 ±0.167	0.863 ±0.010
		MRMR aux.	0.621 ±0.029	0.670 ±0.034	0.950 ±0.158	0.863 ±0.009
	Spec+Emb hidden-fusion CRF	Whisper aux.	0.624 ±0.031	0.674 ±0.044	0.923 ±0.170	0.862 ±0.011
		Perch aux.	0.633 ±0.026	0.665 ±0.032	0.865 ±0.143	0.861 ±0.009
		Concat aux.	0.627 ±0.031	0.667 ±0.037	0.886 ±0.199	0.861 ±0.010
		Sum aux.	0.625 ±0.026	0.673 ±0.035	0.914 ±0.155	0.861 ±0.008
		MRMR aux.	0.614 ±0.034	0.703 ±0.039	0.978 ±0.191	0.841 ±0.013
BIO	Raw TweetyNet	Spectrogram	0.560 ±0.077	1.013 ±0.141	2.622 ±0.675	0.758 ±0.052
	WhisperSeg	Zero-shot	0.020 ±0.000	1.064 ±0.000	5.324 ±0.000	0.366 ±0.000
		Fine-tuned	0.442 ±0.219	0.739 ±0.150	2.772 ±1.724	0.566 ±0.174
	Frame	Whisper	0.204 ±0.108	2.477 ±1.500	9.445 ±8.064	0.427 ±0.134
		Perch	0.251 ±0.042	2.972 ±0.941	11.083 ±5.659	0.489 ±0.037
		Concat	0.243 ±0.095	2.675 ±1.494	10.325 ±8.247	0.480 ±0.111
		Sum	0.275 ±0.072	2.519 ±1.286	8.865 ±7.518	0.510 ±0.082
		MRMR	0.327 ±0.019	2.116 ±0.191	6.658 ±1.140	0.588 ±0.013
	Emb-CRF	Whisper	0.041 ±0.117	1.127 ±0.079	4.718 ±0.774	0.436 ±0.075
		Perch	0.597 ±0.031	0.709 ±0.042	1.185 ±0.158	0.693 ±0.019
		Concat	0.593 ±0.032	0.707 ±0.040	1.243 ±0.162	0.694 ±0.019
		Sum	0.587 ±0.032	0.717 ±0.039	1.243 ±0.184	0.691 ±0.015
		MRMR	0.558 ±0.025	0.799 ±0.034	1.301 ±0.171	0.709 ±0.016
	Spec-CRF	Spectrogram	0.669 ±0.029	0.611 ±0.040	**0.695 ±0.112**	0.771 ±0.018
	Spec+Emb score-fusion CRF	Whisper aux.	0.660 ±0.025	0.621 ±0.036	0.775 ±0.137	0.770 ±0.019
		Perch aux.	**0.678 ±0.026**	**0.595 ±0.031**	0.748 ±0.138	0.773 ±0.017
		Concat aux.	0.673 ±0.029	0.603 ±0.036	0.786 ±0.150	0.773 ±0.018
		Sum aux.	0.673 ±0.026	0.604 ±0.034	0.790 ±0.139	0.770 ±0.017
		MRMR aux.	0.664 ±0.027	0.616 ±0.036	0.804 ±0.141	0.781 ±0.016
	Spec+Emb hidden-fusion CRF	Whisper aux.	0.663 ±0.027	0.620 ±0.031	0.757 ±0.132	0.773 ±0.015
		Perch aux.	0.672 ±0.035	0.608 ±0.045	0.708 ±0.157	0.774 ±0.017
		Concat aux.	0.672 ±0.028	0.607 ±0.033	0.710 ±0.141	0.777 ±0.014
		Sum aux.	0.664 ±0.031	0.614 ±0.039	0.748 ±0.154	0.777 ±0.014
		MRMR aux.	0.661 ±0.026	0.642 ±0.035	0.866 ±0.144	0.749 ±0.018

continued on next page

Table 3 continued from previous page

Scheme	System	View	Unit F1 (tol0–3)	Seg. err. ↓	Count MAE ↓	Frame F1
BIOE	Raw TweetyNet	Spectrogram	0.473 ±0.056	1.500 ±0.311	5.357 ±1.645	0.660 ±0.014
	WhisperSeg	Zero-shot	0.020 ±0.000	1.064 ±0.000	5.324 ±0.000	0.280 ±0.000
		Fine-tuned	0.442 ±0.219	0.739 ±0.150	2.772 ±1.724	0.489 ±0.157
	Frame	Whisper	0.198 ±0.077	4.014 ±2.088	17.620 ±11.972	0.350 ±0.110
		Perch	0.247 ±0.027	3.489 ±0.566	14.895 ±3.465	0.414 ±0.015
		Concat	0.213 ±0.070	3.993 ±1.720	17.811 ±9.925	0.361 ±0.089
		Sum	0.249 ±0.081	2.879 ±1.613	11.311 ±9.256	0.394 ±0.087
		MRMR	0.276 ±0.018	2.953 ±0.191	12.023 ±1.103	0.488 ±0.011
	Emb-CRF	Whisper	0.020 ±0.098	1.142 ±0.079	4.851 ±0.571	0.307 ±0.050
		Perch	0.610 ±0.026	0.690 ±0.036	1.164 ±0.152	0.612 ±0.015
		Concat	0.601 ±0.028	0.698 ±0.038	1.203 ±0.139	0.611 ±0.014
		Sum	0.602 ±0.032	0.692 ±0.046	1.232 ±0.157	0.609 ±0.016
		MRMR	0.589 ±0.026	0.742 ±0.033	1.279 ±0.138	0.619 ±0.015
	Spec-CRF	Spectrogram	0.665 ±0.020	0.608 ±0.032	0.737 ±0.106	0.660 ±0.015
	Spec+Emb score-fusion CRF	Whisper aux.	0.663 ±0.025	0.612 ±0.034	0.775 ±0.125	0.659 ±0.013
		Perch aux.	0.679 ±0.024	0.586 ±0.031	0.774 ±0.140	0.673 ±0.013
		Concat aux.	0.681 ±0.021	**0.583 ±0.028**	0.776 ±0.131	0.675 ±0.013
		Sum aux.	0.677 ±0.023	0.587 ±0.031	0.776 ±0.126	0.671 ±0.014
		MRMR aux.	0.661 ±0.022	0.609 ±0.027	0.825 ±0.130	0.675 ±0.010
	Spec+Emb hidden-fusion CRF	Whisper aux.	0.673 ±0.029	0.599 ±0.040	0.734 ±0.138	0.668 ±0.016
		Perch aux.	0.675 ±0.031	0.599 ±0.043	0.698 ±0.124	0.669 ±0.017
		Concat aux.	**0.683 ±0.024**	0.588 ±0.035	**0.679 ±0.111**	0.671 ±0.014
		Sum aux.	0.677 ±0.028	0.595 ±0.040	0.690 ±0.126	0.672 ±0.017
		MRMR aux.	0.645 ±0.024	0.654 ±0.032	0.969 ±0.150	0.644 ±0.015
BIOES	Raw TweetyNet	Spectrogram	0.475 ±0.054	1.489 ±0.301	5.285 ±1.630	0.530 ±0.009
	WhisperSeg	Zero-shot	0.020 ±0.000	1.064 ±0.000	5.324 ±0.000	0.224 ±0.000
		Fine-tuned	0.442 ±0.219	0.739 ±0.150	2.772 ±1.724	0.391 ±0.126
	Frame	Whisper	0.203 ±0.094	2.937 ±1.573	11.915 ±8.507	0.281 ±0.081
		Perch	0.249 ±0.028	3.469 ±0.585	14.839 ±3.549	0.330 ±0.014
		Concat	0.237 ±0.070	3.706 ±1.725	16.008 ±9.991	0.314 ±0.064
		Sum	0.247 ±0.064	3.252 ±1.494	13.475 ±8.708	0.327 ±0.055
		MRMR	0.279 ±0.015	2.918 ±0.192	11.806 ±1.122	0.393 ±0.009
	Emb-CRF	Whisper	0.009 ±0.061	1.146 ±0.049	4.942 ±0.358	0.238 ±0.027
		Perch	0.603 ±0.037	0.697 ±0.053	1.209 ±0.168	0.491 ±0.013
		Concat	0.606 ±0.026	0.689 ±0.037	1.183 ±0.150	0.492 ±0.013
		Sum	0.604 ±0.033	0.693 ±0.041	1.201 ±0.180	0.488 ±0.011
		MRMR	0.594 ±0.025	0.733 ±0.033	1.296 ±0.143	0.499 ±0.011
	Spec-CRF	Spectrogram	0.669 ±0.019	0.602 ±0.028	0.695 ±0.090	0.530 ±0.011
	Spec+Emb score-fusion CRF	Whisper aux.	0.658 ±0.022	0.616 ±0.033	0.776 ±0.126	0.530 ±0.014
		Perch aux.	0.678 ±0.020	0.589 ±0.030	0.752 ±0.120	0.539 ±0.013
		Concat aux.	0.679 ±0.021	**0.587 ±0.032**	0.746 ±0.104	0.540 ±0.013
		Sum aux.	0.676 ±0.023	0.591 ±0.031	0.760 ±0.122	0.538 ±0.011
		MRMR aux.	0.671 ±0.022	0.597 ±0.030	0.772 ±0.125	0.543 ±0.012
	Spec+Emb hidden-fusion CRF	Whisper aux.	0.671 ±0.027	0.606 ±0.039	0.712 ±0.115	0.534 ±0.013
		Perch aux.	**0.682 ±0.027**	0.593 ±0.037	0.695 ±0.139	0.536 ±0.012
		Concat aux.	0.676 ±0.027	0.597 ±0.041	0.712 ±0.108	0.537 ±0.015
		Sum aux.	0.677 ±0.026	0.597 ±0.036	**0.684 ±0.098**	0.535 ±0.013
		MRMR aux.	0.655 ±0.024	0.643 ±0.032	0.944 ±0.159	0.519 ±0.012

Table 3: Full results for all evaluated systems on the fixed day-disjoint test set, reported as mean ± SD over 50 train-validation splits. All systems are decoded into spans and scored against the same BIO-derived common gold; Unit F1 is averaged over boundary tolerances of 0–3 frames. *Emb-CRF* denotes an embedding-only BiLSTM–CRF, *Spec-CRF* a spectrogram-front-end BiLSTM–CRF, and *Spec+Emb* systems combine spectrogram and embedding CRF branches. In fusion rows, *aux.* identifies the embedding branch paired with the spectrogram branch.

Compared system	**Δ Unit F1 (best $-$ compared)**	**95% CI**	**Holm-corrected p**
Raw TweetyNet (IO)	0.050	[0.041, 0.058]	7.6×10^{-12}
Spec-CRF (BIOE)	0.017	[0.009, 0.026]	4.8×10^{-4}
Spec+Emb score fusion (BIOE)	0.002	[-0.006, 0.010]	0.653
Emb-CRF, Perch (BIOE)	0.073	[0.063, 0.082]	2.7×10^{-14}
WhisperSeg fine-tuned (BIOE)	0.241	[0.182, 0.303]	2.3×10^{-11}

Table 4: Paired comparisons of Unit F1 against our best system, Spec+Emb (hidden fusion), Concat(W,P), BIOE. Two-sided Wilcoxon signed-rank tests paired over the 50 matched train-val splits; Δ is best system minus compared system, with bootstrap CIs on the mean paired difference and Holm-corrected p. Unit F1 is comparable across schemes because predictions are decoded to spans and scored against a common BIO-derived gold; the TweetyNet comparison should be read with the IO caveat in Section 2.