Fine-Grained Prosody Transfer for Expressive AI-Generated Audiobooks and Digital Narration

Mariola Trost
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
trost1992@oregonstate.edu

Tripti Pandit
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
tripti452@buffalo.edu

Nela Dudko
Department of Computer Science, University of North Texas, Denton, TX, USA.
nela1994@unt.edu

Abstract

AI-generated audiobooks and digital narration require more than intelligible speech. They require local prosodic decisions that convey emphasis, suspense, pacing, character voice, and discourse structure across long passages. Current neural text-to-speech systems can control global style, speaker identity, and approximate speaking rate, but they often blur the distinction between a reference narrator's expressive timing and the target text's lexical content. This manuscript develops a methodological framework for fine-grained prosody transfer in long-form narration. The framework represents prosody as a structured sequence of log F0, F0 change, phone duration, energy, and boundary cues; aligns reference prosody to target text through hierarchical cross-attention; and suppresses speaker-identity leakage with adversarial invariance while preserving linguistically licensed prominence and phrasing. An uncertainty-weighted objective prevents ambiguous or weakly aligned prosodic targets from dominating training, and constrained interpolation provides interpretable control over transfer strength. Analytical assessment addresses intelligibility protection, identity–prosody entanglement, long-range prosodic drift, and evaluation validity. The central argument is that fine-grained prosody transfer should be treated as constrained structured prediction rather than global style embedding. This framing clarifies model design, reduces predictable failure modes, and supports production workflows in which narration quality depends on local expressive decisions rather than acoustic fidelity alone.

Keywords

prosody transfer, expressive speech synthesis, audiobook narration, style disentanglement, uncertainty modeling, neural text-to-speech

1. Introduction

Neural text-to-speech has moved from concatenative and parametric synthesis toward sequence-to-sequence acoustic modeling with neural vocoders, substantially improving naturalness and speaker similarity [1]. For audiobook and digital-narration applications, however, naturalness is only one requirement. A narrator continuously resolves local interpretive problems: which word in a phrase receives pitch accent, how long a comma pause

should be, when a sentence should accelerate toward a revelation, and where a character's voice should differ from the narrative frame. These decisions are prosodic, local, and context-dependent. They are also constrained by lexical identity, syntax, punctuation, discourse structure, and speaker identity. A system that reproduces a reference narrator's global timbre or average pitch range but fails to place prominence on the correct syllable does not achieve expressive narration.

Research on prosody transfer has established that reference-conditioned models can imitate aspects of expressive timing and intonation [2]. Style-token and global-style approaches provide useful controls over broad speaking modes, but they compress local variation into a small set of latent factors [3]. Controllable feed-forward architectures improve duration and pitch prediction while reducing alignment errors [4], and end-to-end variational models strengthen acoustic modeling and sampling stability [5]. These advances do not solve the central long-form problem: a reference performance contains speaker identity, recording conditions, lexical content, and expressive intention entangled in the same acoustic signal. Transferring the expressive component without importing the reference speaker's identity or misaligning prominence to the target text remains difficult.

The contribution developed here is a design framework for fine-grained prosody transfer in audiobook narration. It combines a structured prosodic target representation, hierarchical text-to-prosody alignment, adversarial speaker-invariance, uncertainty-aware supervision, and interpretable transfer-strength control. The framework is not presented as an experimentally validated system; rather, it is a technical synthesis of established neural TTS mechanisms and prosodic theory organized around the specific constraints of long-form narration. The analysis explains why fine-grained transfer should be formulated as constrained structured prediction, identifies likely failure modes, and proposes evaluation logic that separates intelligibility, expressive fidelity, identity leakage, and long-range consistency. The practical significance lies in shifting audiobook synthesis away from global style imitation toward controllable local interpretation that producers, narrators, and listeners can assess.

2. Related Work and Design Requirements

Out-of-domain style transfer in singing voice synthesis is instructive because it separates style from content under limited parallel data and must preserve linguistic or lyrical content while importing expressive delivery [6]. The analogy to narration is partial: singing imposes melodic constraints that audiobook narration does not, while narration imposes stronger demands on sentence-level syntax, discourse emphasis, and speaker consistency over hours of speech. The useful lesson is architectural: reference-conditioned style encoding works best when style and content are explicitly disentangled and when the transfer target is represented at multiple granularities rather than as a single vector.

Sequence-to-sequence acoustic models with attention provide the basic mechanism for mapping text to acoustic frames [7]. Neural vocoders then reconstruct waveforms with high fidelity [8], but vocoder quality cannot compensate for misplaced prominence or unnatural pause placement. Prosodic descriptors used in voice research, including F0 statistics, energy, spectral balance, timing, and jitter-related measures, offer interpretable targets and diagnostic variables [9]. Linguistic work shows that prosody guides syntactic parsing, information structure, and referential interpretation [10], and that prosodic functions include marking focus, uncertainty, affect, and discourse boundaries [11]. Experimental and theoretical reviews further emphasize that prosody is not a single acoustic dimension but a structured system relating prominence, phrasing, and tune [12]. These findings imply that fine-grained

transfer requires explicit representation of local prosodic events, not only global style embeddings.

Voice conversion research adds a second relevant mechanism: adversarial training can encourage latent representations to discard speaker identity while retaining content or style information [13]. For narration, the design objective is not to erase all speaker information, because the target narrator must remain recognizable, but to prevent reference-speaker identity from leaking through transferred prosody. The framework therefore needs a controllable invariance mechanism rather than complete invariance. It also needs long-form consistency because sentence-level models can produce discontinuities at chunk boundaries. These requirements motivate a hierarchical, uncertainty-aware architecture with explicit constraints on timing, prominence, and speaker identity.

3. Problem Formulation and Methodological Framework

3.1 Fine-grained prosodic target representation

The target of transfer is a prosodic sequence aligned to the target text, not a global style vector. For each frame or phone-aligned unit t, the local prosodic target is represented as a vector of interpretable acoustic and temporal descriptors. This representation preserves F0 level, F0 movement, duration, and energy while keeping boundary information separate because phrase boundaries have different linguistic status from within-word timing.

$$p_t = (\log f0_t, \Delta \log f0_t, d_t, e_t)$$

Here, f0_t is the fundamental frequency at unit t, Δ log f0_t is the first difference of log F0 and captures local contour direction, d_t is the duration of the associated phone or syllable nucleus, and e_t is frame energy normalized within the utterance. The vector p_t deliberately excludes speaker embedding dimensions and recording-level loudness, because those variables should be controlled by the target narrator profile rather than imported from the reference. Boundary cues are represented separately as a categorical variable b_t indicating whether t lies at a phrase-internal, phrase-final, sentence-final, or paragraph-level break. This separation matters because a rising contour near a comma and a rising contour near a question mark can be acoustically similar but pragmatically different.

The representation is phrase-relative. For each sentence or prosodic phrase, F0 values are normalized by the target speaker's local median and range, and durations are normalized by the target speaker's phone-duration statistics. Without such normalization, a reference narrator with a low voice or slow tempo would impose absolute acoustic values that conflict with the target speaker's identity. The resulting target is a sequence of relative prosodic shapes: local rises, falls, lengthenings, reductions, and boundary movements that can be transferred while the target speaker's timbre remains intact.

3.2 Hierarchical encoding and cross-attention

The text side is encoded at multiple levels: phone or character, syllable, word, phrase, and sentence. The reference side is encoded as a style memory containing local prosodic units. A hierarchical cross-attention layer then retrieves, for each target text position, a reference prosodic context that is relevant to the target's syntactic and information structure. The attention weights are not fixed by lexical identity; they are conditioned on text context and on the requested transfer strength.

$$h_t \; = \; \sum_\{i = 1\}\^\{N\} \alpha_\{t, i\} \, v_i, \alpha_\{t, i\} \; = \; softmax_i(e_\{t, i\})$$

In this expression, h_t is the retrieved prosodic context at target position t, v_i is the i-th reference style-memory vector, e_{t,i} is the attention score between target position t and reference unit i, and α_{t,i} is the normalized attention weight over N reference units. The attention follows transformer-style content-based matching [16], but it is constrained by monotonic or phrase-level alignment priors so that target prominence is not assigned to an unrelated reference position. Such constraints are necessary because audiobook narration contains repeated words, character names, and formulaic phrases that can produce spurious attention matches.

The retrieved context h_t conditions a prosody predictor that outputs $\hat{p}_t$, the predicted local prosodic target. The predicted prosody then conditions an acoustic model and waveform generator; neural vocoders are capable of high-fidelity waveform reconstruction [15], but the vocoder is not asked to invent missing prosodic structure. The predictor is trained with multi-level supervision so that phone-level duration, syllable-level prominence, phrase-level contour, and sentence-boundary movement are optimized together rather than treated as independent post-processing steps.

3.3 Loss design and speaker-invariance

The prosodic loss combines frame-level contour accuracy, duration accuracy, energy accuracy, and boundary classification. This combination prevents a model from achieving low average pitch error while producing flat contours or incorrect pauses.

$$L_p = \lambda_f\, L_f0 + \lambda_d\, L_dur + \lambda_e\, L_energy + \lambda_b\, L_boundary$$

Here, L_f0 measures squared error or correlation loss on normalized log F0 contours, L_dur measures phone-duration discrepancy, L_energy measures normalized energy error, and L_boundary is a cross-entropy loss for phrase and sentence boundaries. The coefficients λ_f, λ_d, λ_e, and λ_b trade off these objectives. Duration and boundary terms should receive substantial weight in narration because pause placement and segment lengthening carry interpretive information that F0 alone cannot represent.

Speaker leakage is addressed with an adversarial classifier attached to the prosodic context representation. A gradient-reversal layer encourages the representation to predict prosodic targets well while making speaker identity difficult to classify. The classifier is trained with a standard cross-entropy objective.

$$L_grl = -\mu \sum_{k=1}^{K} y_k \log q_k(h)$$

Here, K is the number of target-speaker classes, y_k is the one-hot speaker label, q_k(h) is the classifier's predicted probability for speaker k given hidden representation h, and μ controls the strength of gradient reversal. The purpose is to prevent reference-speaker identity from being encoded in h while preserving prosodic variation that is linguistically or expressively licensed. Because prosody and identity are partially correlated, μ cannot be set arbitrarily high; excessive invariance can remove voice-quality cues that make a narrator recognizable. The design therefore treats speaker invariance as a tunable constraint, not an absolute objective.

3.4 Uncertainty-aware prosody supervision

Prosodic targets are often ambiguous. A comma may permit a short pause or no pause, a focused word may be realized with a pitch accent or with lengthening, and a reference

performance may contain microphone or breathing artifacts that are not part of expressive intent. A model trained with equal weighting on all targets will overfit to noisy or weakly aligned events. The framework therefore predicts a per-unit uncertainty and uses it to weight the prosodic loss.

$$L_u \;=\; \sum_\{t=1\}\^\{T\}\,(1/(2\sigma_t\^2))\;||p_t \;-\; \hat{p}_t||\^2 \;+\; log\;\sigma_t$$

In this expression, p_t is the target prosodic vector, p̂_t is the predicted vector, σ_t is the predicted standard deviation for unit t, and T is the number of aligned units. The first term down-weights units with high predicted uncertainty, while the log σ_t term prevents the model from assigning infinite uncertainty to every unit. This objective is a diagonal Gaussian negative log-likelihood and is appropriate when uncertainty is local and approximately independent after alignment. It is not a substitute for alignment quality control; rather, it makes the learning objective aware that some prosodic targets are intrinsically underdetermined.

3.5 Interpretable transfer control

A production system should allow a director or narrator to request subtle or strong expressive transfer without retraining the model. Linear interpolation between a neutral target-speaker prosody and a transferred reference prosody provides a simple control, provided the interpolation occurs in the normalized prosodic space rather than in raw acoustic space.

$$p_t(\beta) \;=\; p_t\^neutral \;+\; \beta(p_t\^style \;-\; p_t\^neutral),\; 0 \;\leq\; \beta \;\leq\; 1$$

Here, p_t^neutral is the target speaker's neutral prosody predicted from text, p_t^style is the prosody obtained from reference-conditioned transfer, and β is the transfer strength. At β = 0 the system produces neutral narration; at β = 1 it applies the full transferred contour. Values between these extremes allow the director to preserve intelligibility when a reference performance is highly stylized. The interpolation is meaningful only if both endpoints share the same alignment and normalization; otherwise the midpoint can produce unnatural contour mixtures rather than a controlled reduction in expressive intensity.

4. Analytical Assessment and Technical Discussion

4.1 Intelligibility protection under local transfer

Fine-grained prosody transfer can improve expressiveness while increasing intelligibility risk. Lexical stress, vowel reduction, and phrase boundaries are not optional ornaments; they support word recognition and syntactic parsing. A transfer model that moves prominence from the target word to a neighboring function word may sound expressive yet make the sentence harder to process. The proposed representation reduces this risk by separating boundary cues and by requiring attention priors that respect target syntax. Duration and energy losses also protect intelligibility because they constrain segment lengthening and prominence strength. Claims of human-level synthesis quality [17] should therefore be interpreted carefully: acoustic naturalness and semantic appropriateness are distinct endpoints, and audiobook narration stresses the latter.

The main remaining risk is alignment drift across long sentences. If the attention mechanism assigns a reference contour to the wrong target syllable, the resulting prosody can be locally natural but lexically misleading. Phrase-level normalization and boundary supervision mitigate this failure, but they do not eliminate it. Production systems should

therefore include alignment diagnostics that flag high-attention entropy, large duration residuals, or contour discontinuities at chunk boundaries.

4.2 Expressive fidelity versus speaker identity

The adversarial invariance term addresses reference-speaker leakage, but it introduces a trade-off. Prosodic style is partly speaker-specific: an intimate narration style may be realized through low energy, narrow pitch range, and slow tempo, all of which correlate with speaker identity and recording conditions. Aggressive invariance can flatten these cues, producing a neutral target voice that fails to convey the reference's expressive intention. Conversely, weak invariance can import the reference narrator's vocal persona, which is unacceptable when the target speaker is a distinct character or brand voice. The appropriate operating point depends on the application: character dialogue may permit stronger prosodic transfer than first-person narration, while instructional narration may prioritize consistency and intelligibility over expressive intensity.

Zero-shot speaker adaptation methods illustrate the broader difficulty of separating what a voice is from how it is used [18]. A target speaker embedding can preserve timbre while prosody is supplied by a separate predictor, but the predictor must not treat speaker-correlated prosodic habits as universal expressive features. The framework's phrase-relative normalization and speaker-invariance classifier provide a principled way to manage this boundary, but they do not remove the need for human review in high-stakes narration.

4.3 Long-form consistency and prosodic drift

Audiobooks expose errors that short-utterance synthesis hides. Sentence-level models can reset pitch range, pause length, or speaking rate at every chunk, producing a narration that sounds locally acceptable but globally unstable. The hierarchical encoder addresses this by conditioning on sentence and paragraph context and by carrying a slowly varying prosodic state across chunks. That state should represent narrative mode, character perspective, and pacing, not specific F0 values, because absolute F0 must remain tied to the target speaker. Boundary classification is especially important at chapter and section transitions, where a pause may be syntactic, rhetorical, or structural.

Prosodic drift can also arise from uncertainty weighting. If the model becomes confident on frequent neutral patterns and uncertain on rare expressive events, the uncertainty term may suppress the very events that make narration expressive. The objective should therefore be paired with event-level sampling or reweighting during training so that questions, exclamations, character turns, and discourse markers are not treated as noise. This is a design constraint derived from the structure of audiobook text, not an empirical result.

4.4 Evaluation logic and error modes

Evaluation should not collapse into a single mean opinion score. At minimum, it must separate intelligibility, prosodic appropriateness, expressive similarity to the reference, target-speaker similarity, and long-range consistency. A normalized prosodic distance can summarize contour, duration, energy, and boundary discrepancies, but it should be reported alongside perceptual judgments because a small average distance can hide a single catastrophic misaccent.

$$D_p = w_f\, D_f0 + w_d\, D_dur + w_e\, D_energy + w_b\, D_boundary$$

Here, D_f0 is a normalized contour distance, D_dur is a duration distance, D_energy is an energy distance, and D_boundary is a boundary mismatch penalty; w_f, w_d, w_e, and w_b

are nonnegative weights chosen for the narration task. The metric is useful for diagnostic comparison, not as a substitute for listening tests. A system may achieve low D_p while placing emphasis on the wrong word if the weighting does not reflect information structure. Evaluation should therefore include targeted probe sentences with contrastive focus, parentheticals, lists, quotations, and character dialogue.

5. Implications for Audiobook Production

For production use, fine-grained prosody transfer changes the unit of editorial control. Instead of selecting a global narrator style and applying it uniformly, a director can specify local expressive intentions: emphasize a reveal, slow a descriptive passage, sharpen a character's retort, or reduce transfer strength in dense factual sections. The model's interpretable parameters, especially transfer strength and boundary behavior, make such direction possible without retraining. Quality assurance can focus on alignment failures, identity leakage, and long-range drift rather than only on waveform artifacts.

The framework also clarifies where human narration remains valuable. Professional narrators do not merely apply a style; they resolve ambiguous prosody using narrative knowledge that may not be present in punctuation or syntax. A transfer system can imitate local contours, but it cannot justify why a particular word should be emphasized unless the text representation encodes discourse structure. The practical implication is not replacement but augmentation: synthesis can handle repetitive or provisional narration, while human review and direction concentrate on interpretive decisions.

6. Conclusion

Fine-grained prosody transfer for AI-generated audiobooks is a structured prediction problem under linguistic, speaker-identity, and long-form consistency constraints. The framework developed here represents prosody as local F0, duration, energy, and boundary targets; aligns reference prosody to target text through hierarchical cross-attention; suppresses speaker leakage with tunable adversarial invariance; weights ambiguous targets through uncertainty-aware supervision; and exposes transfer strength as an interpretable control. The analysis shows that global style embeddings are insufficient for narration because they cannot reliably place local prominence or preserve discourse boundaries. It also shows that fine-grained transfer without explicit constraints can improve expressiveness while damaging intelligibility or importing the reference narrator's identity. Treating prosody transfer as constrained structured prediction provides a coherent basis for model design, failure analysis, and evaluation in audiobook and digital-narration systems.

References

1. Taylor, P. (2009). Text-to-speech synthesis. Cambridge University Press.

2. Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., & Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In Proceedings of the 35th International Conference on Machine Learning (pp. 4693–4702). PMLR.

3. Wang, Y., Stanton, D., Zhang, Y., Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., & Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Proceedings of the 35th International Conference on Machine Learning (pp. 5180–5189). PMLR.

4. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019). FastSpeech: Fast, robust and controllable text to speech. In Advances in Neural Information Processing Systems (Vol. 32, pp. 3171–3180).

5. Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (pp. 5530–5540). PMLR.

6. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

7. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779–4783). IEEE.

8. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17022–17033).

9. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190–202.

10. Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. Language and Speech, 40(2), 141–201.

11. Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. Speech Communication, 36(1–2), 31–43.

12. Wagner, P., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. Language and Cognitive Processes, 25(7–9), 905–945.

13. Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 266–273). IEEE.

14. Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ES-1. Speech Communication, 137, 1–18.

15. Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

17. Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F., Qin, T., Zhao, S., & Liu, T. Y. (2022). NaturalSpeech: End-to-end text to speech synthesis with human-level quality. arXiv preprint arXiv:2205.04421.

18. Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Proceedings of the 39th International Conference on Machine Learning (pp. 2709–2720). PMLR.