2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

Beyond Semantics:
A Perception-Based CNN Model for Quantifying Phonetic Rhythm in Speech

Mohammad Mahdi Peyravi , Alireza Talebpour , Zeynab Hajimohammadi
PhD student in Computer Engineering, Artificial Intelligence major
Professor of Computer Engineering, Artificial Intelligence, Shahid Beheshti University
mohammadmahdi.peyravi@iau.ac.ir

Abstract

Quantifying subjective, prosodic features of speech, such as rhythm, remains a significant challenge in Natural Language Processing. A primary obstacle is "semantic bias," where a listener's perception of sound is heavily influenced by the meaning of the word. This paper proposes a novel methodology to overcome this challenge and develop a robust computational model for objectively quantifying phonetic rhythm. Our method involves three key steps: 1) Operationally defining rhythm within a continuous two-dimensional space (Percussiveness/Softness and Tempo/Slowness) based on phonetics. 2) Collecting perceptual ground-truth data from native speakers via an online survey. 3) Critically, to eliminate semantic bias, we designed and utilized a corpus of phonotactically valid meaningless sentences. These sentences were generated using a heuristic function to ensure full coverage of the rhythmic spectrum. Finally, a Convolutional Neural Network (CNN) was trained on this perception data, learning to predict rhythmic scores directly from phoneme sequences. Experimental results demonstrate that the CNN architecture significantly outperforms a baseline LSTM model, confirming our hypothesis that phonetic rhythm is a local phenomenon best captured by CNNs. This work provides a foundational tool for computational stylistics and advanced Text-to-Speech (TTS) systems.

Research method

To completely isolate phonetic sound from semantic meaning, we designed and implemented a novel methodology centered around the creation of a specialized corpus consisting of phonotactically valid Persian nonsense sentences. These generated sentences strictly adhere to the phonological and grammatical structure rules of the target language but convey absolutely no semantic information or literal meaning, thereby ensuring that human evaluators and machine learning models are exposed exclusively to pure acoustic properties. To populate this corpus effectively and avoid a natural clustering around a neutral mean, a heuristic generation strategy based on sonority scores was utilized to create a bimodal U-shaped distribution, capturing extreme representations of "very soft" and "very percussive" rhythms. For the computational modeling, we proposed a deep learning architecture utilizing a 1D Convolutional Neural Network (1D-CNN). The model accepts sequences of International Phonetic Alphabet (IPA) phoneme embeddings as input. It then processes these sequences through parallel convolutional layers featuring varying kernel window sizes of 3, 4, and 5 phonemes, designed to act as local n-gram feature detectors before passing through global max-pooling and dense layers for regression.

conclusion

In conclusion, the empirical findings of this study strongly support the linguistic hypothesis that human perception of speech rhythm is primarily a local acoustic phenomenon rather than a global, long-range sequential dependency. While recurrent architectures like LSTMs attempt to capture long-term context, they inadvertently introduce a "blurring" effect that obscures the critical local rhythmic signals. Conversely, the 1D-CNN architecture successfully mimics the human auditory system by utilizing local convolutional windows to detect the precise "clumping" of plosives or the smooth flow of sonorous liquids. By filtering out semantic bias through the deployment of a phonotactically valid nonsense corpus, this research successfully demonstrates that deep learning can be trained to evaluate pure aesthetic and phonetic qualities of language. This framework establishes a powerful, objective, and robust computational foundation for fields such as computational stylistics, literary analysis, and affective Text-to-Speech (TTS) systems, enabling machines to synthesize speech with human-like rhythmic expression and emotional depth. Future work will expand this perception-based approach to cross-lingual rhythm analysis.

The purpose of the research

The primary objective of this research is to develop a robust, automated framework to quantify abstract acoustic and rhythmic features—specifically categorized into percussiveness, softness, and tempo—completely independent of semantic interference. By bridging the gap between computational linguistics and deep learning architectures, this study aims to map raw phonetic sequences directly to human auditory perception scores. Traditional models struggle with this because text-based features carry heavy semantic baggage. This research seeks to overcome these limitations by establishing a standardized, two-dimensional rhythmic space. The first dimension contrasts percussiveness and softness, which is deeply grounded in the phonological concept of sonority, where high-sonority phonemes contribute to a soft perception and low-sonority plosives generate a percussive feel. The second dimension measures tempo versus slowness, capturing the perceived pace of speech influenced by syllable count and consonant cluster complexities. Ultimately, this research aims to provide a pure, perception-based metric that can evaluate the phonetic aesthetics of speech, laying a scientific foundation for advanced computational stylistics and more natural, emotionally expressive speech synthesis technologies.

Practical or simulation results

The simulation and empirical results demonstrated that the proposed 1D-CNN architecture achieved superior performance compared to the traditional Long Short-Term Memory (LSTM) baseline model. In quantitative evaluations, the 1D-CNN model significantly minimized prediction errors, achieving a remarkably low Validation Mean Squared Error (MSE) of 0.142, whereas the LSTM model lagged behind with an MSE of 0.236. Furthermore, the proposed model successfully demonstrated a high goodness-of-fit, accounting for over 81.5% of the total variance in human rhythmic perception data with an R-squared (R^2)score of 0.815, compared to the baseline LSTM's R^2score of 0.640. To validate the internal mechanics and interpretability of the model, saliency map analysis was performed. The results confirmed that the parallel convolutional filters effectively learned to identify specific phonetic patterns; filters associated with high percussiveness became highly sensitive to local plosive clusters, while filters tracking softness learned to detect the continuous flow of sonorous liquid consonants and vowels, proving the model's alignment with human auditory perception.

References

1. Peyravi, M. M., Talebpour, A., & Hajimohammadi, Z. (2025). Beyond Semantics: A Perception-Based CNN Model for Quantifying Phonetic Rhythm in Speech. Research Institute for Interdisciplinary Quran Studies, Shahid Beheshti University.

2. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1251-1258.

3. Dellwo, V. (2006). Rhythm printing: A method for the visualization of speech rhythm. Proceedings of the 3rd International Conference on Speech Prosody, 113-116.

4. Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal.

5. Cognition, 73(3), 265-292.Grabe, E., & Low, E. L. (2002). Durational variability in speech and the characteristics of a speech rhythm hierarchy. Laboratory Phonology, 7, 515-546.

6. Clements, G. N. (1990). The role of the sonority cycle in core syllabification. Papers in Laboratory Phonology, 1, 283-333.

7. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.