

Exploring the potential of dynamic time warping algorithms for automated intonation evaluation in EFL learners

Yongkook Won
Hanyang University
linguistry@hanyang.ac.kr

Accurate assessment of intonation remains a challenge in second language (L2) pronunciation research and pedagogy, particularly in English as a Foreign Language (EFL) contexts. Intonation, defined as systematic variation in fundamental frequency (F0) across an utterance, plays a crucial role in conveying discourse meaning and speaker intent [1, 2]. Traditional approaches to intonation assessment rely primarily on trained human raters, whose judgments are considered the gold standard in pronunciation research [3, 4]. Despite their validity, human ratings are time-consuming, resource-intensive, and difficult to scale for large-scale assessments [5]. As a result, there has been increasing interest in automated suprasegmental assessment within computer-assisted pronunciation training (CAPT) systems [6, 7]. However, many existing automated measures depend on pointwise pitch comparisons or global summary statistics, which often fail to capture the temporal structure of intonation contours.

Dynamic Time Warping (DTW) is a time-series alignment algorithm originally developed for speech recognition to compare sequences that differ in length or speaking rate [8, 9]. By allowing nonlinear alignment between two sequences, DTW provides a principled way to compare entire F0 contours while accounting for temporal variation. Owing to these properties, DTW has been increasingly applied to speech-related tasks, including pronunciation and intonation evaluation [6, 10]. However, systematic comparisons of multiple DTW variants against human judgments of intonation, particularly across different proficiency levels, remain limited.

The present study explores the potential of DTW-based measures as automated indicators of intonation quality in EFL speech. A total of 183 read-aloud utterances produced by Korean elementary school EFL learners were analyzed. Learner productions were compared to native English reference recordings using normalized F0 contours. Seven DTW variants were implemented, including Standard DTW, Derivative DTW (DDTW), and Locally Weighted DTW (LWDTW), each of which emphasizes different aspects of contour alignment such as slope, local weighting, or shape similarity [11, 12]. To focus on intonational shape rather than absolute pitch height, DTW distance values were standardized prior to analysis.

Human ratings of intonation were obtained from trained raters using a nine-point scale, and comprehensibility ratings were included as an index of learner ability or proficiency. Correlation analyses revealed that most DTW distance measures were negatively associated with human-rated intonation scores, indicating that smaller DTW distances corresponded to more native-like intonation patterns. Among the algorithms tested, DDTW and LWDTW demonstrated the strongest associations with perceptual judgments, suggesting that approaches sensitive to contour slope and local alignment are particularly effective for modeling intonational variation in L2 speech.

Further analyses using nonparametric kernel regression showed that DDTW and LWDTW accounted for approximately 29% and 32% of the variance in human-rated intonation scores, respectively. Importantly, the strength of the relationship between DTW distance and intonation ratings varied across levels of comprehensibility, indicating that learner proficiency mediates the effectiveness of automated intonation measures. This finding highlights the need to consider proficiency-related factors when designing and interpreting automated pronunciation assessment systems.

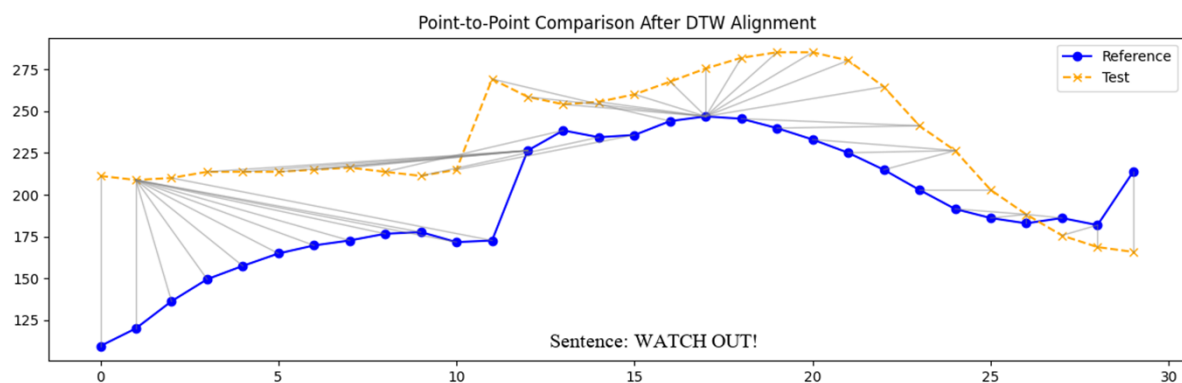


Fig. 1. DTW analysis of intonation contours.

Funding

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5B5A16052864).

References

- [1] Wells, J. C. (2006). *English intonation: An introduction*. Cambridge University Press.
- [2] Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.
- [3] Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369-377. <https://doi.org/10.2307/3588485>
- [4] Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73-97. <https://doi.org/10.1111/j.1467-1770.1995.tb00963.x>
- [5] Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51(10), 832-844. <https://doi.org/10.1016/j.specom.2009.04.005>
- [6] Arias, J. P., Yoma, N. B., & Vivanco, H. (2010). Automatic intonation assessment for computer aided language learning. *Speech Communication*, 52(3), 254-267. <https://doi.org/10.1016/j.specom.2009.11.001>
- [7] Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10), 883-895. <https://doi.org/10.1016/j.specom.2009.04.009>
- [8] Rabiner, L., Rosenberg, A., & Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(6), 575-582. <https://doi.org/10.1109/TASSP.1978.1163164>
- [9] Müller, M. (2007). *Information retrieval for music and motion*. Springer.
- [10] Sheoran, K., Bajgoti, A., Gupta, R., Jatana, N., Dhand, G., Gupta, C., Dadheech, P., Yahya, U., & Aneja, N. (2023). Pronunciation scoring with goodness of pronunciation and dynamic time warping. *IEEE Access*, 11, 15485-15495. <https://doi.org/10.1109/ACCESS.2023.3244393>
- [11] Jeong, Y.-S., Jeong, M. K., & Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition*, 44(9), 2231-2240. <https://doi.org/10.1016/j.patcog.2010.09.022>
- [12] Yuan, J., Douzal-Chouakria, A., Varasteh Yazdi, S., & Wang, Z. (2019). A large margin time series nearest neighbour classification under locally weighted time warps. *Knowledge and Information Systems*, 59(1), 117-135. <https://doi.org/10.1007/s10115-018-1184-z>