Exploring the Utility of Automatically Generated References for Assessing L2 Prosody
Mariana Julião1,2, Helena Moniz1,3 & Alberto Abad1,2
1INESC-ID, 2IST, 3CLUL - Lisbon
mariana.juliao@inesc-id.pt

Prosody assessment is a challenging yet essential aspect of effective communication, particularly in the context of spoken language. Prominence, denoting the emphasis placed on specific words within a phrase, significantly contributes to speech intelligibility. While modern e-learning platforms have recently incorporated prosody assessment, the existing approaches remain simplistic, relying mainly on imitation tasks. In this study, we explore the possibility of a novel system designed to evaluate the quality of spoken sentences in terms of prominence. Our system, Goodness of Prosody (GoP) [Fig1] comprises two main branches: an acoustic-based branch for prosody classification, which identifies prominent words, and a text-based branch for prosody prediction, which determines the expected prominent words.

The current version of GoP utilizes forced alignment, leveraging readily available text data. However, future iterations can integrate an Automatic Speech Recognition (ASR) module to extend its usability in scenarios where text availability is limited. For prosody classification, we trained a prosodic event detector using English radio news speech (BURNC [1]) to detect pitch accents. When selecting an acoustic classifier, our primary consideration was not only its overall performance but also its independence from the speaker's proficiency levels. We compared classifiers based on wavelets, a CNN-only model, and a CNN+LSTM model. Our results demonstrated that the CNN+LSTM model outperformed the others and exhibited minimal variation across different proficiency levels.

To predict prosody from text, we followed the example of [4], who released a full annotation of LibriTTS on prominence and boundaries after classifying the corresponding speech. We annotated two large corpora (LibriTTS [5] and VCTK [6]) by classifying them with our best acoustic model, previously described. With this, we have generated labels for the text. Also based on the work of [4], we trained a network to label text for prominence. This consisted of a fully connected layer which got BERT [7] embeddings as input, and which learns the prominence labels previously assigned by the acoustic classifier. The full process is described in [Fig2].

One of the primary challenges in prosody assessment is the wide range of possible variations in speech production for the same text-intention pairs. This variability also applies to prominence, as it is often left to the speaker's sensitivity to decide which words to emphasize. To address this issue, we developed different text-based classifiers to predict the accentedness of each word in an utterance, providing reference standards for the assessment of spoken productions. We trained four different text-classifiers, each trained on a distinct corpus or corpus partition. By generating multiple references for the same utterance and selecting the best one, we aimed to account for the diverse possibilities.

Finally, we examined the agreement between the L2-corpus labels and the generated text references. In our evaluation framework, LeaP [8], the speaker proficiency levels were not absolute; rather, they indicated whether the speaker had undergone prosody training or language immersion. Consequently, we limited our comparisons to "before" and "after" categories, excluding comparisons with other speakers. We observed no correlation between speaker proficiency and reference matching [Tab2], regardless of utterance length, except for native speakers who tended to exhibit more mismatches. In particular, we notice that after going abroad the accuracy improved approximately 2%, but when comparing before and after a prosody training course, the accuracy worsened 4%. Further investigation of utterance and reference mismatches revealed that native speakers often over or under-emphasized words in unexpected ways, highlighting the subjective nature of prominence. Additionally, we identified some shortcomings in the acoustic classifier, indicating areas for improvement.

This study contributes to the advancement of prosody assessment for e-learning platforms by proposing the GoP system, which combines acoustic-based classification and text-based prediction to evaluate the prominence of spoken sentences. The findings shed light on the challenges associated with prosody assessment, particularly in L2 speech, and provide insights into the potential enhancements needed for future iterations of the GoP system.
Fig1: Goodness of Prosody model.

Tab2: Accuracy between the closest generated reference and the manual labeling of prominence. Average per utterance, speaker and level.

<table>
<thead>
<tr>
<th>Level</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>going abroad</td>
<td>0.889</td>
</tr>
<tr>
<td>after (e2)</td>
<td>0.905</td>
</tr>
<tr>
<td>prosody training course</td>
<td>0.861</td>
</tr>
<tr>
<td>after (c2+c3)</td>
<td>0.819</td>
</tr>
<tr>
<td>superlearner</td>
<td>0.861</td>
</tr>
<tr>
<td>native</td>
<td>0.797</td>
</tr>
</tbody>
</table>

References