# Computer-assisted pronunciation training in Icelandic (CAPTinI): developing a method for quantifying mispronunciation in L2 speech

Catlin Richter[1], Branislav Bédi[2], Ragnar Pálsson[3], and Jón Guðnason[4]

**Abstract**. We are developing a Computer-Assisted Pronunciation Training (CAPT) system for learners of Icelandic as a second and foreign language (L2). Based on pre-designed tasks in pronunciation exercises, this system will provide corrective feedback on learners' speech. One of the main features we are implementing is a new method for automatic pronunciation scoring to provide immediate feedback on learners' errors. We report promising results for a pilot study of this method in Norwegian, where the pronunciation score successfully distinguishes between native speakers and adult learners, and we discuss how this method informs our continued development of Icelandic CAPT.

**Keywords**: pronunciation training, computer-assisted pronunciation training, speech error detection, pronunciation scoring, Icelandic as a second or foreign language.

## 1. Introduction

Currently, there are no mobile applications, websites, or other online tools on the market that enable learners of L2 Icelandic to practise real speaking skills. According to a recent review of online tools for L2 Icelandic, there are about 15 tools offering pronunciation exercises or theoretical explanations about pronunciation to learners (Bédi, 2022). The available online tools that aim to help

1. Reykjavík University, Reykjavík, Iceland; caitlinr@ru.is; https://orcid.org/0000-0002-4491-0423
2. The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland; branislav.bedi@arnastofnun.is; https://orcid.org/0000-0001-7637-8737
3. Reykjavík University, Reykjavík, Iceland; ragnarp@ru.is
4. Reykjavík University, Reykjavík, Iceland; jg@ru.is; https://orcid.org/0000-0001-6560-5543

**How to cite this article**: Richter, C., Bédi, B., Pálsson, R., & Guðnason, J. (2022). Computer-assisted pronunciation training in Icelandic (CAPTinI): developing a method for quantifying mispronunciation in L2 speech. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022* (pp. 334-339). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1480

learners with practising spoken Icelandic offer only pre-recorded sounds of letters, words, phrases, or shorter texts with full sentences which learners can listen to and, if they wish, repeat aloud. As a result, no corrective feedback about learners' mispronunciation is provided.

However, current technology for delivering such feedback has unsatisfactory performance, even in the most commonly taught L2s such as English. Using state-of-the-art methods to detect L2 pronunciation errors, only six of ten identified 'errors' were actually mispronounced (i.e. 60% precision), while just 40% to 80% of incorrect pronunciations are detected (i.e. recall; Korzekwa, Lorenzo-Trueba, Drugman, & Kostek, 2022). Furthermore, these technologies depend on excellent Automatic Speech Recognition (ASR) or Text-To-Speech (TTS) support for the target language, so performance grows even worse in languages without highly developed ASR or TTS.

Therefore, high-quality automatic pronunciation feedback for Icelandic must be created. This article presents development towards a system for CAPT in Icelandic (CAPTinI) that will deliver such feedback to learners, including a method to immediately detect mispronounced parts of learners' productions, and an initial evaluation of this method in a related language.

## 2. CAPT design

The CAPTinI system is applied as part of a series of lessons which give the learners practice with selected examples of vowels, consonants, words, phrases, and sentences. As such, the system simulates a classroom setting although only virtually, and corresponds with Crabbe's (2003) learning-opportunity framework, which includes elements of comprehensible input and output, interaction exercises, feedback, and rehearsal opportunities, all of which lead to language understanding and possibly learning. The lesson content and sequencing coordinate with the course design of *Icelandic Online*, a freely available web-based course series developed by the University of Iceland and launched in 2004, which has about 80,000 active users (Arnbjörnsdóttir, Friðriksdóttir, & Bédi, 2020). In our interactive pronunciation exercises, learners at different levels may listen to correct pronunciations and repeat them, or attempt to read text aloud without errors, and in either case continue practising with feedback after each attempt until they achieve accurate pronunciation. We first used Norwegian learners' speech to validate the method for detecting pronunciation accuracy, and it will be integrated with CAPTinI later.

## 3. Scoring method

We propose the *RelativeDTW* method of scoring pronunciation accuracy, using Dynamic Time Warping (DTW) to quantify how closely learners' speech matches both native (L1) and other L2 speakers. DTW measures similarity between samples of speech by aligning and comparing corresponding elements. It accurately quantifies L2 accent strength by comparing L2 speakers with a set of L1 ('reference') speakers. DTW has previously been applied in different ways for other CAPT systems (Bartelds, Richter, Liberman, & Wieling, 2020; Yue et al., 2017) and therefore is suitable here.

To help factor out confounds, *RelativeDTW* also compares the learner's pronunciation to a set of L2 references and determines pronunciation accuracy by a difference-to-sum ratio of the two DTW scores (Figure 1). This reflects the observation that accurate pronunciations are relatively closer to L1 than L2 pronunciations, regardless of absolute DTW values, so the ratio facilitates consistent mispronunciation detection across learners. Similar two-way comparisons are often beneficial for CAPT (Fu, Chiba, Nose, & Ito, 2020; Jia et al., 2014).

Figure 1. Equation for *RelativeDTW* score: the difference-to-sum ratio of DTW values from L2 and L1 speakers' reference recordings

$$RelativeDTW = \frac{DTWL2 - DTWL1}{DTWL2 + DTWL1}$$

## 4. Experiment and results

Since DTW is comparative, evaluating a learner's utterance requires reference recordings of other speakers saying the same sounds or words. While data collection for these Icelandic recordings is still in progress, we performed a pilot study validating the method in Norwegian using the NB Tale corpus[5], which includes 240 Norwegian L1 speakers representing maximum dialect diversity and 117 L2 speakers from various backgrounds.

*RelativeDTW* is evaluated on ability to distinguish between native and non-native speakers' pronunciations, so the overlapping coefficient (OVL) of score

5. NB Tale – Speech Database for Norwegian https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-31/

distributions for L1 versus L2 test recordings expresses the method's effectiveness. To estimate practical utility, we also report recall for detecting L2 speech with the score threshold set to correctly accept at least 95% of L1 speech. As a baseline we evaluate the method of Bartelds et al. (2020), measuring DTW distance to L1 references only, using wav2vec-2.0 speech representations for all evaluations.

Results in Table 1 show that L1 and L2 speakers receive quite different pronunciation scores from our method; some individual L2 phonetic segments are indistinguishable from L1 speakers', but errors contribute to overall non-native-like accents in longer speech. In all cases, *RelativeDTW* outperforms the baseline.

Table 1. Evaluation of how effectively our *RelativeDTW* score and the Bartelds et al. (2020) baseline distinguish L1 and L2 speakers in the NB Tale corpus

| | **Overlap** | | **L2 recall at 95% L1 acceptance** | |
|---|---|---|---|---|
| | **RelativeDTW** | **Baseline** | **RelativeDTW** | **Baseline** |
| Sentences | 0.10 | 0.33 | 97% | 68% |
| Words | 0.34 | 0.62 | 63% | 30% |
| Phonemes | 0.45 | 0.76 | 46% | 17% |

## 5. Discussion and conclusions

The *RelativeDTW* difference ratio is sensitive to non-native-like pronunciations of L2 speakers, without rejecting valid L1 variation, and it performs better than the Bartelds et al. (2020) baseline. Although our task of classifying L2 but not L1 speakers as 'non-native-like' is not directly comparable to other previous work (e.g. Korzekwa et al., 2022), we can report 90% precision with 46% recall for phonemes, or 93% precision with 63% recall for words, i.e. 63% of non-native Norwegian speakers' words were identified as sounding non-native-like while under one in ten of so-called 'mispronunciations' were native speakers. These results use 25-speaker reference sets, but even reduction to three speakers identified mispronunciations in 28% of L2 words (OVL=0.54).

The NB Tale speech database provides a challenging first test for *RelativeDTW*, because these Norwegian learners are already advanced and fluent while the native speakers' dialects are highly diverse. Icelandic contains far less dialect variation, and CAPTinI is integrated with lessons for less advanced learners, so these pilot results provide assurance that *RelativeDTW* should perform well in Icelandic once we complete data collection. The requirement for reference recordings, and

consequent impossibility of scoring spontaneous speech, is a main limitation of this method. However, no other corpora in the target language are needed, so it is feasible even for languages with limited existing speech technology. Other pronunciation scoring methods are more flexible at the cost of needing hundreds of hours of training data or specialist annotation (Korzekwa et al., 2022).

Currently, we have begun implementing scoring for a selection of Icelandic pronunciation exercises, and will continue development and evaluation with this new dataset. We are collecting more Icelandic recordings through the *Samrómur* platform, the main speech data gathering platform for Icelandic language technology. Across all projects, 4,000 hours of speech from nearly 30,000 L1 and L2 Icelandic speakers have been collected so far. Our preliminary results using Norwegian samples promise a positive language-learning experience in CAPTinI for improving learners' pronunciation in L2 Icelandic.

## 6. Acknowledgements

We thank our colleagues in developing the Icelandic pronunciation training lessons: Eydís Huld Magnúsdóttir, Júlíus Reynald Björnsson, Kolbrún Friðriksdóttir, Marc Daníel Skipstað Volhardt, Róbert Kjaran, Safa Jemai, and Staffan Hedström. This project was funded by the Language Technology Programme for Icelandic 2019-2023.

## References

Arnbjörnsdóttir, B., Friðriksdóttir, K., & Bédi, B. (2020). Icelandic Online: twenty years of development, evaluation, and expansion of an LMOOC. In K.-M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds), *CALL for widening participation: short papers from EUROCALL 2020* (pp. 13-19). Research-publishing.net. https://doi.org/10.14705/rpnet.2020.48.1158

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020). A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence, 29*. https://doi.org/10.3389/frai.2020.00039

Bédi, B. (2022). Development of online tools supporting the learning of Icelandic as a foreign and second language. In *Tungumál í víðu samhengi: Afmælisrit til heiðurs Birnu Arnbjörnsdóttur,* (pp. 47-56). Reykjavík.

Crabbe, D. (2003). The quality of language learning opportunities. *TESOL Quarterly*, *37*(1), 9-34. https://doi.org/10.2307/3588464

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication 116*, 86-97. https://doi.org/10.1016/j.specom.2019.12.002

Jia, J., Leung, W.-K., Wu, Y.-H., Zhang, X.-L., Wang, H., Cai, L.-H., & Meng, H. M. (2014). Grading the severity of mispronunciations in CAPT based on statistical analysis and computational speech perception. *Journal of Computer Science and Technology, 29*, 751-761. https://doi.org/10.1007/s11390-014-1465-2

Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., & Kostek, B. (2022). Computer-assisted pronunciation training—speech synthesis is almost all you need. *Speech Communication, 142*, 22-33. https://doi.org/10.1016/j.specom.2022.06.003

Yue, J., Shiozawa, F., Toyama, S., Yamauchi, Y., Ito, K., Saito, D., Minematsu, N. (2017). Automatic scoring of shadowing speech based on DNN posteriors and their DTW. *Proc. Interspeech* 2017 (pp. 1422-1426). https://doi.org/10.21437/Interspeech.2017-728

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net



**Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022**
**Edited by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton**

**Publication date: 2022/12/12**



Typeset by Research-publishing.net
Cover photo by © 2022 Kristinn Ingvarsson (photo is taken inside Veröld – House of Vigdís)
Cover layout by © 2022 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-38372-015-7 (PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

**Legal deposit, France**: Bibliothèque Nationale de France - Dépôt légal: décembre 2022.