Prosody for lexical stress in Icelandic CAPT

Annotated Bibliography

Caitlin Richter, Reykjavik University, for CAPTinI

26 September 2022

References

[1] Juan Pablo Arias, Nestor Becerra Yoma, and Hiram Vivanco. Automatic intonation assessment for computer aided language learning. *Speech communication*, 52(3):254–267, 2010.

- Scores general prosody quality, and derives a score of stress
- Uses DTW to align L2 with L1 reference recording, then computes correlation of pitch across aligned frames
- Proposes equation to detect stress errors based on comparing aligned L2 and L1 pitch and intensity
- Features: pitch, intensity
- Evaluation: Word error rate 21.5% for stress detection. Very comprehensive evaluations.

[2] Hossein Bozorgian and Esmat Shamsi. Computer-assisted pronunciation training on Iranian EFL learners' use of suprasegmental features: A case study. *Computer-Assisted Language Learning Electronic Journal*, 21(1):93–113, 2020.

- Uses My English Tutor (MyET) commercial software for five Iranian learners of English, B1 level
- MyET: Learner listens to example and repeats it. Feedback is 4 scores on Pronunciation, Pitch, Timing, Emphasis, and viewing waveforms of both the example and learner's production.
- Emphasises stress as a learning goal, but not clear how or if this was implemented in CAPT
- No evaluation of how accurate MyET feedback is for this learner population

[3] Nancy F Chen and Haizhou Li. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–7. IEEE, 2016.

- Review paper on prosody error detection (not phonetic errors) in CAPT: lexical stress and lexical tone
- Classifiers are the most common method for lexical stress error detection, with duration and pitch as key features
- Conclusion: more work is needed, no clear practical recommendation or best practise

[4] Jian Cheng. Automatic assessment of prosody in high-stakes English tests. In *Twelfth annual conference of the international speech communication association*, 2011.

- Scores general prosody/fluency, to correlate with human ratings of L2 English prosody quality
- Uses k-means clustering to build a set of canonical F0 contours and energy contours for each word
- F0 sampling method removes all duration information; explicit duration models per phoneme were added back

- Features: F0, energy, duration, phone identity
- Evaluation: Correlation 0.80 between aggregated score and human rating

[5] Dorothy M Chun and Yan Jiang. Using Technology to Explore L2 Pronunciation. *Second Language Pronunciation: Bridging the Gap Between Research and Teaching*, page 129, 2022.

- Recent non-technical review for language teachers, discussing speech processing technologies for CAPT
- Considers prosody and suprasegmental features important in CAPT, but almost entirely focuses on intonation - not stress
- Finds large gaps between available speech processing technology, L2 speech research, and education in practise
- Studies that evaluate commercial CAPT software like MyET are often unclear about what the software does and what the evaluation measures
- There is also not much research on what learners do with various types of feedback such as pitch contours

[6] Nicole Dehé. The timing of nuclear and prenuclear Icelandic pitch accents. In *Speech Prosody*, 2010.

- Investigates timing and shape of pitch contours for phrase accents on different Icelandic syllable types, in a few specific types of sentences
- Collected dataset: 12 speakers x 3 repetitions each of 12 sentences, and 12 other speakers x 3 repetitions each of the same 12 sentences in different contexts which change the emphasis
- Uses ToBI for analysis; the goal of the research is theoretical/descriptive phonology
- The author has several other papers with similar focus, i.e. theoretical linguistics research with hand-annotated prosodic analysis of a small single-purpose Icelandic dataset

[7] Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, 69:31–45, 2015.

- Detects each syllable in L2-English words as primary, secondary, or unstressed, based on GMMs of features for each stress type
- Trained on 75k words L1 English, tested on L1 and L2 English (all children)
- Features: duration, pitch, energy, spectral tilt, MFCCs - extracted for each syllable nucleus, normalised per word
- Evaluation: At best 11% syllable error rate for L1 English children, 20% error rate for L1 Japanese children speaking English. When constrained to max 5% false corrections, misses about half of actual errors.
- No advantage to phone-specific models, contra previous papers - probably because of general higher speech error rate in this data

[8] Kazunori Imoto, Yasushi Tsubota, Antoine Raux, Tatsuya Kawahara, and Masatake Dantsuji. Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system. In *Seventh International Conference on Spoken Language Processing*, 2002.

- Classifies syllables in English sentences as primary stress (one per phrase), other lexical stress, or unstressed
- Detects syllable errors e.g. vowel insertion using ASR with an error vocabulary designed for Japanese learners of English
- Trained with part of TIMIT, 31 L1 English speakers

- Features: Log F0, log power, MFCCs with delta-deltas
- Evaluation: 95% accurate syllable classification for held-out L1-English speakers, 84% accuracy for L1-Japanese

[9] Heini Kallio, Antti Suni, and Juraj Šimko. Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of english with different language backgrounds. *Language and Speech*, 65(3):571–597, 2022.

- Evaluates how well measures of fluency and stress placement predict human ratings of L2-English fluency for L1 Czech, Slovak, Polish, and Hungarian speakers
- Features for stress (syllable prominence): representations related to F0, energy, and duration, extracted with continuous wavelet transform (CWT), not fully automated and requires human supervision/adjustment for each sample
- Features for fluency relate to speed (rate), breakdown (pause), and repair (false start, correction)
- Evaluation: These stress features did not capture much information related to a speaker's fluency, for any L1

[10] Okim Kang, David O Johnson, and Alyssa Kermad. *Second Language Prosody and Computer Modeling*. Routledge, 2022.

- Regardless of the general title, book is concerned with the authors' own work, with English as the L2
- Almost all experiments use only L1 English speech, minimal evaluation of methods for L2 speakers
- Most research focuses on idealised description of speech and errors, not practical automatic detection of them
- Methods do not reflect current speech processing technology, e.g. formants, segmental accuracy scoring unrelated to most methods used in CAPT or developed in computational research, ToBI labelling algorithm ignores nearly all work in this area from the past 10 years
- Tables 5.6 and 5.7 list many possible features that could be used for prosody or fluency scoring
- Figure 3.1 is a display of possible correct vs. incorrect stress visual feedback
- Evaluation: Table 5.2, for prominent/stressed syllable detection, 86% accuracy for L1-English speakers

[11] Tsuneo Kato, Quy-Thao Truong, Kohei Kitamura, and Seiichi Yamamoto. Referential vowel duration ratio as a feature for automatic assessment of L2 word prosody. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6595–6599. IEEE, 2019.

- Suggests additional prosodic feature for stress or fluency, representing whether learners have similar relative durations for adjacent syllables as native speakers
- Other/baseline features: Pitch contour, intensity contour (comparison e.g. DTW)
- Method depends on parallel L1 recordings of words to be scored for L2 speakers
- No direct evaluation on stress-detection task

[12] Daniel Korzekwa, Roberto Barra-Chicote, Szymon Zaporowski, Grzegorz Beringer, Jaime Lorenzo-Trueba, Alicja Serafinowicz, Jasha Droppo, Thomas Drugman, and Bozena Kostek. Detection of lexical stress errors in non-native (l2) english with data augmentation and attention. *arXiv preprint arXiv:2012.14788*, 2020.

- Detects lexical stress errors in L2 English
- Uses TTS to generate training data corpus of stress errors, and uses attention mechanism instead of pre-defining regions of interest from forced alignment
- Features: Pitch, intensity, and phoneme duration from forced alignment
- Training data: 12000 stress-annotated words, 8000 L1 assumed to be correct and 4000 manually annotated L2 or TTS possibly containing errors
- Also depends on pre-existing TTS for the target language, and forced alignment used acoustic features trained on LibriSpeech (960 hours) - unknown how robust this method is to less accurate alignments
- Evaluation: 95% precision and 49% recall for detecting syllable stress errors in L2 English of Slavic and Baltic speakers

[13] Björn Kristinsson. *Towards speech synthesis for Icelandic*. 2004.

- Section 2.5 discusses Icelandic prosody, including ToBI annotation, from the perspective of speech synthesis technology
- Defines initial algorithm to predict pitch contour of syllables in a sentence based on lexical and sentence stress - for synthesis only, not expected to match human production
- Duration also expected to be related to stress but not yet confirmed by research

[14] Konstantinos Kyriakopoulos. *Deep Learning for Automatic Assessment and Feedback of Spoken English*. PhD thesis, University of Cambridge, 2022.

- Section 2.3.4 reviews approaches to syllable stress detection for CAPT, both word and sentence stress.
- Aside from this background the work does not involve stress.
- Features for stress detection include vowel duration, energy, pitch, and equations to normalise these.
- Sections 2.3.3 and 2.3.5 summarise available metrics related to fluency, e.g. words per second, phones per second, disfluencies and silence duration. Some of this is very English-specific.
- Section 2.3.5 reviews use of pitch (F0) to evaluate L2 proficiency, compare with native speakers' pitch contours, or extract pitch events to compare with canonical prosody.
- Most pitch methods rely critically on ToBI annotation. This is very time consuming and I do not think there is a published ToBI standard for Icelandic anyway.

[15] Kun Li, Shaoguang Mao, Xu Li, Zhiyong Wu, and Helen Meng. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. *Speech Communication*, 96:28–36, 2018.

- Classifies English syllables as primary, secondary, or no stress in words with 3 or more syllables, using mixed Gaussian-Bernoulli Restricted Boltzmann Machine
- Trained and tested on L2-English corpus from L1 Mandarin & Cantonese speakers, annotated for syllable stress
- Features: nucleus duration, maximum loudness, min/max pitch, per syllable
- Evaluation: 88% syllable classification accuracy, or 67% total word accuracy, for L2-English words with at least 3 syllables
- Also trained pitch accent classifier with 90% accuracy

[16] AS Liberman. Stress in Icelandic. *Phonetica*, 31(3-4):125–143, 1975.

- Focus on theoretical linguistics, phonological analysis of Icelandic - probably no practical consequences

- Claims that stress does not exist in Icelandic, as the linguistic rules of stress are redundant with rules of syllable quantity (vowel/consonant length)

[17] Yoo Rhee Oh, Kiyoung Park, Hyung-Bae Jeon, and Jeon Gue Park. Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition. *Etri Journal*, 42(5):761–772, 2020.

- Scores fluency for L2 Korean and correlates this score with 5 human ratings per sentence: general proficiency, phonetic accuracy, phonologic accuracy, fluency, pitch/accent.
- Evaluation: For L2-Korean of speakers from 5 other Asian countries, correlation around 0.75 between the automatic score and most human rating types (range from 0.54 for pitch, to 0.86 for fluency rating)
- Trained native and non-native BLSTM-based acoustic models from log mel filterbanks, concatenated for context window
- Other features extracted from forced alignments, e.g. counts and durations of pauses, connected speech, syllables; ratios among these - tables 1-3 in paper

[18] Jörgen Pind. Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception & Psychophysics*, 57(3):291–304, 1995.

- Collected dataset: isolated words *gala, galla, kala, Kalla*, spoken by four native Icelandic speakers, at 5 different speaking rates, 5 repetitions of each word, i.e. 400 tokens.
- Acoustic measurements annotated by hand, using spectrograms and waveforms
- See Figure 2: a linear classifier of vowels in initial syllables as long or short, based on the ratio of this vowel's duration and the duration of the consonant after it.

[19] Jörgen Pind. Speech segment durations and quantity in Icelandic. *The Journal of the Acoustical Society of America*, 106(2):1045–1053, 1999.

- Many works cited & discussed in the introduction
- See Figure 1: a linear classifier of vowels (average) as long or short, based on the ratio of the vowel's duration and the duration of the consonant after it.
- Citations in Figure 1 include different word types, contexts, dialects, etc. and the linear classifier appears highly reliable
- See Figure 3, 4: classifier is not perfectly clean for extended connected speech, though still clear overall - unlikely to be a problem for isolated short sentences. did you train a neural network to learn a linear classifier ;———;

[20] Evgeny Pyshkin, John Blake, Anton Lamtev, Iurii Lezhenin, Artyom Zhuikov, and Natalia Bogach. Prosody training mobile application: Early design assessment and lessons learned. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2, pages 735–740. IEEE, 2019.

- Creates and evaluates mobile CAPT application to provide visual feedback on English intonation
- Displays user's pitch contour alongside model contour, and calculates a score for user representing how good their pronunciation was
- Screenshots of app in the paper, also http://studyintonation.org/
- Focus on developing algorithm for robust and user-comprehensible pitch tracking
- Qualitative pilot evaluation using native speakers only
- Evaluation: Accurate pitch tracking even of native speakers remains very challenging, but generally native speakers can use this visual feedback to match their intonation contour closer to a model over time

- Scoring method does not handle timing differences well, not useful compared to DTW which most other papers use to measure difference between pitch tracks

[21] Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. ASR Inspired Syllable Stress Detection for Pronunciation Evaluation Without Using a Supervised Classifier and Syllable Level Features. In *INTERSPEECH*, pages 924–928, 2019.

- Stress detection via forced alignment: phone set includes stress indicator on every vowel and pronunciation dictionary includes all relevant variants of the words
- Trains acoustic models from L1 speech only; no special annotation is needed except for testing/evaluation
- Tested with ISLE corpus: German and Italian speakers of L2 English
- 85-87% accuracy at L2 stress detection when trained on 960 hours L1 speech (Librispeech); 72-79% accuracy with only 30 hours training if we want to try stress detection, but expect minimum 20% errors and might be even worse for Icelandic than English since L1 stress is more regular

[22] Yong Ruan, Xiangdong Wang, Hong Liu, Zhigang Ou, Yun Gao, Jianfeng Cheng, and Yueliang Qian. An end-to-end approach for lexical stress detection based on transformer. *arXiv preprint arXiv:1911.04862*, 2019.

- English lexical stress detection using seq2seq model of audio features + phoneme sequence to stress-marked-phoneme sequence
- Trained on Librispeech English subsets 100-960 hours, tested on TIMIT L1 English and L2-ARCTIC L2 English
- Evaluation: L1 syllable classification error rate 11% with 100 hours L1 training data, 6% with 960 hours. L2 error rate 14 to 9%.
- Claims to be less influenced by inaccurate phoneme or syllable boundaries from forced alignment (but probably still needs decent word boundaries)

[23] Sandra Schwab and Jean-Philippe Goldman. MIAPARLE: Online training for discrimination and production of stress contrasts. In *Proc. 9th Int. Conf. Speech Prosody*, pages 572–576, 2018.

- Web application trains 6 French speakers to perceive and produce Spanish stress contrasts
- Uses single isolated words
- Feedback: Chart representing all syllables in the word, with the most prominent one marked
- Prominence score is based on F0 and syllable duration
- Prominence scoring method was developed for L1 French and requires significant annotation for training data
- No evaluation of how accurate the prominence detection was in this application (i.e. speaking Spanish not French)

[24] Dávid Sztahó, Gábor Kiss, László Czap, and Klára Vicsi. A computer-assisted prosody pronunciation teaching system. In *WOCCI*, pages 45–49, 2014.

- Application claims to teach intonation, stress, and rhythm
- There is no measurement, scoring, feedback, etc. for anything defined as 'stress'; only for intonation and rhythm
- Intonation visual feedback shows user's pitch contour alongside example pitch contour
- Mainly qualitative evaluation for L1 Hungarian hard of hearing children

- Describes algorithm to score user's F0 compared to reference recording - not evaluated vs. DTW for intonation score, even though they also do DTW for alignment to display visual feedback

[25] Anjana Sofia Vakil. A CAPT tool for training and research on lexical stress errors in German. In *SLaTE*, page 185, 2015.

- 2015-german-supervised-70%.pdf p. 185 Automatic classification of lexical stress errors for German CAPT
- Web app for French learners of German
- https://github.com/vakila/de-stress
- Learner reads a sentence with one highlighted word, whose lexical stress is scored
- Feedback 'options including visual feedback via abstract graphical visualizations and/or text stylization, auditory feedback via prosodic modification of the learner's utterance, verbal error/success messages, and graphical skill bars corresponding to each of the prosodic parameters analyzed.'
- Not immediately clear how much of that was implemented

[26] Anjana Sofia Vakil and Jürgen Trouvain. Automatic classification of lexical stress errors for German CAPT. In *SLaTE*, pages 47–52, 2015.

- For French speakers of L2 German, classify 2-syllable German words as correct (initial) or incorrect stress, using CART classifiers (WEKA toolkit)
- Manually annotated L2 error corpus for training: 668 bisyllabic word tokens
- Used forced alignment to segment words, syllables, phones
- Features for stress detection: Syllable and nucleus (e.g. vowel) duration, F0 min/max/range, intensity mean/max, word identity, speaker characteristics (level, age, gender)
- Duration and F0 alone are sufficient
- Classification accuracy 70%, F-measure 0.75 for 'correct' stress label

[27] Seung Hee Yang and Minhwa Chung. Self-imitating feedback generation using GAN for computer-assisted pronunciation training. *arXiv preprint arXiv:1904.09407*, 2019.

- Generates self-imitating Korean CAPT feedback, and has native speakers evaluate the generated speech on 5 criteria
- Goal is for learners to listen to resynthesis of their own voice with correct prosody and phonetic accuracy - previous self-imitating feedback could not fix both aspects
- Training data: 30000 L1 Korean recordings and 65000 parallel L2 Korean recordings from various L1 backgrounds
- Evaluation: Sound quality is somewhat low but the voice matching is reasonable, and correction of both prosody and segments is good

[28] Mahmood Yenkimaleki and Vincent J van Heuven. The relative contribution of computer assisted prosody training vs. instructor based prosody teaching in developing speaking skills by interpreter trainees: An experimental study. *Speech Communication*, 107:48–57, 2019.

- Questions pedagogical value of some common prosody feedback styles in commercial CAPT software
- Tests Accent Master software vs. conventional classroom instruction for teaching advanced English prosody to L1-Farsi speakers

- Accent Master shows learner's pitch track alongside reference pitch; also has modules focused on English consonant clusters, vowel contrasts, and lexical stress
- Measures learning outcomes according to teachers' subjective ratings for comprehensibility, accentedness, word stress and sentence stress
- Evaluation: CAPT was more effective overall, but only due to comprehensibility and accentedness improvment - CAPT was not better for learning word or sentence stress

[29] Junhong Zhao, Hua Yuan, Jia Liu, and Shanhong Xia. Automatic lexical stress detection using acoustic features for computer assisted language learning. *Proc. APSIPA ASC*, pages 247–251, 2011.

- 88.6% accuracy classifying English vowels as stressed or unstressed
- Uses features: duration, loudness, pitch, mid-frequency energy, pitch slope type (rise/fall/other), relative duration/loudness/pitch vs. other vowels in the same word, + constraint of one stress per word
- SVM classifiers trained separately for each vowel
- MIR-SD (Multimedia Information Retrieval lab, Stress Detection) dataset: 22 Taiwanese speakers with intermediate L2 English. 3000 words training data, 668 test