# AUTOMATIC FEEDBACK ON PRONUNCIATION AND ANOPHONE: A TOOL FOR L2 CZECH ANNOTATION

RICHARD HOLAJ[1] – PETR POŘÍZKA[2]

[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[2] Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic
[1]holaj@mail.muni.cz, [2]petr.porizka@upol.cz

## ABSTRACT

This paper introduces a research project that represents an innovative approach to e-learning applications targeting automatic feedback on the pronunciation of non-native speakers based on computer speech recognition (specifically for Czech). We have collected data from 187 speakers of different pronunciation levels from 36 languages, conducted a pilot project, and developed the first version of an attributive annotation system based on tagging isolated speech sounds. We briefly mention the results of this stage (especially the success rate of the trained model), which led us to change our strategy and move to the next phase of the development of the automatic speech recognition tool. In this article, we present the current and next project phases: the Anophone annotation tool, a new annotation system based on whole-word tagging (two- to four-syllable words). The result is a measurable improvement in both the model and the success rate of speech recognition.

**Keywords**: automatic feedback on pronunciation, speech recognition, annotation, Czech, e-learning

## 1. INTRODUCTION

Today, there are some e-learning tools available for L2 competency in various languages. However, they prevailingly focus on lexical and grammatical aspects, and not much attention is paid to pronunciation. When they do address pronunciation, they only offer playback of pre-recorded vocabulary (words and phrases) or, exceptionally, provide the user with the possibility to record a short text sample and then compare this recorded sample with the original pre-recording. Thus, there is still a lack of applications providing an evaluation and automatic feedback on pronunciation to help users improve or achieve correct pronunciation in the foreign language they are learning. For Czech, we are aware of only two exceptions so far: the CzechME application [1] with several lessons focused on sound discrimination (differentiation), but the current version of this tool does not provide feedback on the user's pronunciation. The second application is Duolingo [2], where an automatic speech recognition (ASR) system transcribes the recording into text and compares with the text that was supposed to be spoken. However, ASR technology is designed to "understand;" thus, even if the pronunciation is more or less incorrect, it uses a language model [3] to estimate the intended linguistic meaning. That is a problem, because we receive feedback that our pronunciation is correct (successful in communication), even though it may be more than slightly mispronounced.

To our knowledge, probably the most technologically advanced e-learning system for L2 pronunciation is currently the English-focused ELSA Speak mobile app [4]. It includes its own proprietary solution for evaluating pronunciation and providing feedback to users in the form of the sound they were supposed to pronounce and the sound the user pronounced. However, there are still a few shortcomings. It is limited to only a segmental level of pronunciation and the feedback is restricted to the sound inventory of English, even though the speech sounds pronounced by the learners often do not correspond to any correct speech sound in the target language (i.e., English).

This situation led us to the idea that more than just the sound inventory of the target L2 (in case of ELSA, English; in our case, Czech) is needed to create a successful system for providing feedback on L2 pronunciation. Moreover, according to [5], the system "often misidentifies incorrect sounds as correct," so the problem of ASR technology remains. It is also important to mention that detection and evaluation of mispronunciations without accurate feedback is much more advanced, and there are already promising results (see [6], [7]), but none of them, to our knowledge, is able to provide accurate feedback in the form of which non-native sound was pronounced and how it differs from the speech sound that should have been pronounced.

## 2. L2 ANNOTATION SYSTEM AND NON-NATIVE SPEECH RECOGNITION OF CZECH

The aim of our research project is to create an e-learning tool for automatic pronunciation feedback for non-native speakers learning Czech. Due to the

limited scope of this paper, it is not possible to present all key aspects of our research project. Therefore, for this reason, we refer here to the previous article [8] where all key elements are described.

We have completed the first (pilot) phase of the project, which is described and summarized in [8], during which we collected recordings a total of 187 non-native speakers of 36 different languages across all learning levels (using the CEFR scale from A0 to C1) and in different age categories from 18 to 73 years old, with the largest group being speakers under the age of 40 (73%; cumulative frequency: 136 speakers).

The key idea of our approach is to include non-native sounds (phonetic features from the non-native sound inventory) into the speech recognition system inventory, so that we are not limited to the most similar speech sound from the target language and can obtain less biased (more accurate) results from the actual pronunciation. We then want to provide the learners with relevant linguistic information through a mobile app not only by outputting the pronounced versus correct sound, but also by more explicitly describing what was wrong in their pronunciation (e.g., presence of aspiration, nasalized pronunciation, etc.) and how to fix it (e.g., more open mouth, vocal cords in action, etc.).

In the first phase of our research project, we tested this approach on a manually annotated sample of just under 4,000 isolated speech sounds from 32 non-native speakers of Czech using the attributive annotation system described in [8], which we developed for this purpose and which allows us to label not only standard pronunciation but also various deviations from standard pronunciation (see 2.2). Other important phenomena from the project are presented in [8], such as the characterization of the data, the methodological framework that reflects the phonetic basis of Czech, the relevant phonetic specificities of foreign languages, and the most common pronunciation errors of non-native speakers learning Czech.

### 2.1. Test models for the speech recognition of non-native speakers of Czech

For a project based on ASR technology and data (segmented and annotated audio recordings) that contain "incorrect speech sounds" of non-native speakers, we needed an annotation system that allows us to distinguish the differences between the prompt for the desired speech sound and the user's resulting attempt at making the speech sound. Likewise, we needed a tool based on annotated data and capable of recognizing from the recordings of the speech sound and corresponding annotation.

For this purpose, we have developed a tool for speech recognition of individual speech sounds: a Python script for ASR and model evaluation based on the Persephone library [9] and, more recently, the Anophone annotation tool (see 2.3). Our data model consists of three parts in the following ratio: (a) training set – 90%, (b) validation set – 5%, and (c) test set – 5%.

Persephone contains a tool which extracts several audio features, and in both the old and new experiment we used the LMFB (Log Mel Filterbank) with delta and delta-delta features. Extracted features along with their corresponding labels were then split by library into non-intersecting train (90 %), validation (5 %) and test (5 %) sets (see above our data model). "The underlying model used is a long short-term memory (LSTM) recurrent neural network [10] in a bidirectional configuration [11]. The network is trained with the connection's temporal classification (CTC) loss function [12]." [9]

In both the old and new models, we used the same default three-layered architecture with 250 hidden nodes (cf. [8]). As we mentioned in [8], it has to be "trained with pre-processed data for at least 30 epochs. Training stops when one of those conditions is met:

(a) training LER (learning error rate) is lower than 0.1% and the validation LER is lower than 1%
(b) validation LER has not improved in the last 10 epochs
(c) after 100 epochs

In the last step, we test our trained model against the test data set."

### 2.2. Pilot project results (annotation of isolated speech sounds)

In the first phase of the research project, we developed an attribute–value annotation system that works with manually annotated isolated speech sounds and is based on systematically categorized pronunciation errors from different languages or language groups (see [8] for more detail).

This annotation system specifies two groups of attributes: (1) fixed, with a binary value of 0 or 1 for phonological features (such as quantity, voicing, etc.), and (2) variable, with the possibility to add additional values as needed (phonetic features such as palatalization, nasalization, etc.). It also includes a special tag for replacing one speech sound with another.

The annotation tag (label) is divided by a colon into two main parts: (1) the part before the colon indicates the sound to be pronounced; (2) the attributes after the colon indicate deviations in

pronunciation from the standard/correct phonetic form of the speech sound (using possible attribute values). If the pronunciation is correct, only the part before the colon is used. In case of incorrect or non-standard pronunciation, any number of attributes may follow the colon (see examples below).

**a** [standard pronunciation of vowel *a*]
**a:k1** [short vowel *a* pronounced as long]
**a:k1vN** [short vowel pronounced as long and nasalized]
**E:A** [*ɛ* pronounced like *a*]

*Explanatory notes*:
k = quantity (0 = short, 1 = long).
v = non-standard pronunciation variants (N = nasalization)

The pilot experiment [8], included 3,717 labelled sounds from 32 non-native Czech speakers with two versions of annotation AV1 and AV2 (improved by the identification of duplicate tags and some corrections of system errors in the annotation).

The results of those annotation models, which work with isolated speech sounds, are summarized in Table 1:

| error rate | training | validation | test |
|---|---|---|---|
| model AV1 | 43% | 42% | 51% |
| model AV2 | 15% | 37% | 41% |

**Table 1**: Old models – based on the attributive system and annotation of isolated speech sounds (approx. 3,700 samples) [6]

The error rate for AV1 was huge, while adjusted annotation model AV2 was more successful, but error rates (especially validation and test) were still quite big, see [8] for more detailed description.

### 2.3. Anophone annotation system and preliminary results of the new ASR model (word annotation)

The results of the pilot test led us to change our strategy and the whole annotation system to achieve more satisfactory outputs and a more reliable speech recognition model. This change meant moving from an attributive annotation system to a different approach and way of processing speech data. For the next phase of testing, we created new annotated data that annotated whole words instead of isolated speech sounds, and most importantly, we created a new annotation system, Anophone [13].

This is a very flexible tool that allows the uploading of audio segments to a database for further processing: it allows not only the annotation of data, but also the creation of new sets of tags or modifications of existing tagsets according to the user's needs, also assigning these tags to the data (audio segments) through a web interface. In addition, Anophone allows one to build a new annotation task independently of an existing one (for example, for the needs of another project or language). Thus, this tool can be used by other researchers for similar projects: to create their own annotation tasks, and upload and annotate data.

Anophone works with four datasets:
(1) Recordings: segmented sounds for annotation, which can be further filtered by four categories – language, speaker, repetition, and word.
(2) Tasks: annotation tasks, the way we annotate the data – e.g., at the level of vowels, phonetic features, etc.
(3) Labels: the tags (set of tags) that we create for a particular annotation task(s) and through which we annotate the data.
(4) Annotations: custom/individual annotation labels assigned to audio segments in the database (stored under a unique annotator ID).

Anophone selects data to annotate randomly, even repeatedly with the same sounds, to ensure multiple annotations of the same word (segment) by different annotators. Alternatively, intentional filtering of the data by a defined category/attribute (language, speaker, repetition, word)[1] can be used in combination with regular expressions. Thus, we can select only speakers of a particular language or language group, or all speakers regardless of native language, etc. In the administration menu, we can perform a final annotation of the audio segment, which is intended for export to machine learning, considering the existing annotation variants (the annotator is in the system under a certain ID, e.g., to detect systematic annotation errors, to track the correspondence between annotators, etc.).

For this phase of testing the annotation model, we selected more homogeneous data, limited to 4 groups of speakers, namely German, Russian, Ukrainian, and Vietnamese speakers. Apart from the homogeneity of the data, the reason for the choice was that these are the largest groups of non-native speakers in the Czech Republic (we did not include Slovaks because of their interlingual proximity to Czech).

We manually annotated whole words (not isolated speech sounds), specifically more than 6,500 samples (from 47 speakers) of two- to four-syllable words in which the given speech sound occurred in different positions (initial, middle, final) and in different phonemic contexts (vowels, obstruents, sonorants). Each segment was assigned metadata: speaker code, gender, age, nationality, native language, level of

Czech. The sub-segment (word) is named according to the specific meaning of the word and this label is then compared with the text output of the manual annotation, which represents the actual pronunciation of the segment, in the next phase of testing the annotation model. Thus, the word label (file name) is compared with the annotated segments (pronounced sounds, including non-standard variants or phonetic features from foreign languages). For the annotation, two sets of sound-level labels are defined in Anophone: (1) from the standard Czech inventory, (2) or containing various non-standard sounds produced by non-native speakers (in the label set color-coded into two categories: (ad 1) native vs. (ad 2) non-native sounds or features such as aspiration, etc.).

In the current phase we have created three models with different training batch size (one of the neural network parameters). We call those models B72 (batch size 72), B36 (batch size 36), and B18 (batch size 18).

The results of the current phase, including the success rate of the trained model, are summarized in Table 2 (cf. Table 1 and the success rate of the pilot model):

| error rate | training | validation | test |
|---|---|---|---|
| model B72 | 5% | 23% | 24% |
| model B36 | 1% | 23% | 23% |
| model B18 | 11% | 25% | 25% |

**Table 2**: New models – based on Anophone and whole word annotation (approx. 6,500 samples)

Despite fact that the task in the current phase is more difficult than in the initial phase (individual sounds vs. whole words), we see that the results are better, since we achieved much lower error rate in all models. From three tested models, B36 with a medium batch size is the most promising.

### *2.3.1. Detailed evaluation*

Since evaluation is phoneme based (correct vs. incorrect sound in output), to truly see the quality of the tool, we need to look closer at the incorrectly transcribed words.

Table 3 shows a few of those words. In the left part there are mistakes in validation phase, while in the right part there are mistakes in test phase. Both parts consist of expected output on the left side and real output on the right side.

| VALIDATION | | TEST | |
|---|---|---|---|
| **expected** | **output** | **expected** | **output** |
| b ou̯ r̥ k a | p ɔ ʃ k a | m a t k a | s n a p k a |
| ʃ tʲ ɛ s tʲ i | ʃ ɛ s ɟ ɪ | b ɔ : r ʃ k a | b ou̯ : r ʃ k a |
| k ɔ : l ɔ | k ɔ l ɔ | s k ou̯ ʃ k a | s p u ʃ k a |
| ʒ a k | ʐ̩ a k | m a t k a | a k a |

**Table 3**: Incorrectly transcribed words

In Table 3, both in test and validation, mistakes are often make in between two words with perceptive similarity, often interchanging similar speech sounds, differences in length, voice or manner of articulation. Missing segment errors are also common. This implies that since the evaluation does not make any difference for the type of mistake, the results are actually slightly better than we would expect from an error rate alone, since the errors encountered are usually less severe.

## 3. CONCLUSION

Our system has promising results that can lead to a tool which will be able to automatically transcribe the speech of non-native speakers of Czech and thus provide a base to create better feedback for learners of L2 Czech. Although results are getting better, the model is still not good enough to be applied in an e-learning app. To change this, we need to collect more data and optimize our annotation and model. Furthermore, we need to include L2 Czech speakers from another native languages.

We also want to experiment with tier-based annotation which would work with different phonetic tiers and annotate them separately (including tones, stress, and another suprasegmental features that can have many effects on individual segments or on the speech production of non-native speakers in general).

Following our current findings, we believe that a reliable tool for automatic L2 Czech pronunciation feedback will be possible within few years.

## 7. REFERENCES

1. EVE Technologies, s.r.o. 2021. CzechME (1.0.5) [Mobile app]. https://play.google.com/store/apps/details?id=cz.evetech.czechme.
2. Duolingo, Inc. 2021. Duolingo (5.1.5) [Mobile app]. https://play.google.com/store/apps/details?id=com.duolingo
3. Kuhn, R., De Mori, R. 1990. Cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6), 570–583.

4. ELSA Co, Ltd. 2021. ELSA Speak: Online English Learning & Practice App (6.2.1) [Mobile app]. https://play.google.com/store/apps/details?id=us.nobarriers.elsa.
5. Becker, K., Edalatishams, I. 2019. ELSA Speak – Accent Reduction [Review]. In: J. Levis, C. Nagle, and E. Todey (eds), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, Ames, IA: Iowa State University, 434–438.
6. Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., Calamaro, S., and Kostek, B. 2021. Weakly-supervised word-level pronunciation error detection in non-native English speech. *Proceedings of Interspeech 2021*, 4408–4412.
7. Yang, L., Fu, K., Zhang, J., and Shinozaki, T. 2021. Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning. *Neural Networks*, 142, 597–607.
8. Holaj, R., Pořízka, P. 2021. L2 Czech Annotation for Automatic Feedback on Pronunciation. *Jazykovedný časopis / Journal of Linguistics* 72 (2), 510–519.
9. Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., et al. 2019. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *LREC 2018 (Language Resources and Evaluation Conference)*, Miyazaki, Japan, 3356–3365.
10. Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8), 1735–1780.
11. Schuster, M., Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
12. Graves, A., Fernandez, S., Gomez, F., Schmidhuber, J. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd international conference on Machine Learning*, 369–376.
13. Anophone: an annotation tool for phonemes. [Software]. https://anophone.evetech.cz/#/intro.

[1] Readings: the texts were read twice by non-native speakers, first with and then without instructor assistance. The segments were then matched accordingly with the assigned markers v1 (repetition after instructor) vs. v2 (learners pronouncing without any support). Pronunciation quality for most recorded speakers differs noticeably between versions.