sciendo
DOI 10.2478/jazcas-2021-0047

L2 CZECH ANNOTATION FOR AUTOMATIC FEEDBACK ON PRONUNCIATION

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

HOLAJ, Richard – POŘÍZKA, Petr: L2 Czech annotation for automatic feedback on pronunciation. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 510 – 519.

Abstract: In this paper, we would like to provide a brief overview of the current state of pronunciation teaching in e-learning and demonstrate a new approach to building tools for automatic feedback concerning correct pronunciation based on the most frequent or typical errors in speech production made by non-native speakers. We will illustrate this in the process of designing annotation for a sound recognition tool to provide feedback on pronunciation. At the end of the paper, we will also present how we have tried to apply this annotation to the tool, what caveats we have found and what our plans are.

Keywords: pronunciation, L2, Czech, machine learning, neural networks, e-learning, annotation, speech recognition, automatic feedback, phonetics

1 INTRODUCTION

Over the last few decades, online language learning popularity has been growing rapidly [1]. There are dozens of e-learning applications for different languages. These include several tools focused on various languages, most notably Duolingo [2], and a large number of applications focusing on just one language or aspect of language, such as Ten Ta To [3] or CzechME [4].[1] The increasing worldwide popularity and importance of e-learning education have been accelerated even more by the current epidemiological situation caused by Covid-19 [5]. Despite this increasing importance, there is an aspect of language that does not receive as much attention in e-learning, this being pronunciation. This problem is even more critical in a less common L2 such as Czech.

2 LANGUAGE PRONUNCIATION FEEDBACK IN E-LEARNING SYSTEMS

Putting aside a few exceptions, basically the only way e-learning applications approach the teaching of pronunciation is by providing the possibility to play

[1] CzechMe was created as a result of TAČR TL01000342 – an adaptable mobile application for teaching Czech to foreigners. Both authors of the paper were participants in this project.

recordings of words and phrases. Some of those applications also provide the possibility to record users and play and compare their recording with an original record. In general, there is a lack of any feedback or lessons that would teach users how to attain correct pronunciation, or at least a certain pronunciation level. For the Czech language, we are aware of just two exceptions: CzechME and Duolingo. In the first case, there are several lessons focused on sound discrimination (differentiation) and pronunciation, however, the current version of CzechME does not provide feedback on the user's pronunciation. In the second case, an automatic speech recognition (ASR) system is used to transcribe a recording to text, which is then compared with a text that should have been pronounced.

Although some applications often use existing ASR systems (such as Google Cloud Speech or CMUSphinx) to convert speech to text in order to provide feedback to students, there is one big caveat when using ASR technology for learning pronunciation. ASR technology is designed to understand: even when the pronunciation is incorrect, it uses a language model [6] to guess what has been meant. This is a problem, since we receive the feedback that our pronunciation is correct even when it could have been more than just slightly wrong.

These types of tools are used across different L2 and it is apparently a state of the art solution for L2 pronunciation learning with one exception: a mobile application called ELSA Speak [7]. This pronunciation-only application provides an exhaustive amount of pronunciation exercises for English. It also includes a custom proprietary solution for evaluation of correct pronunciation and includes feedback to the user. The feedback is in the form of a speech sound which should be pronounced and the speech sound that the user actually pronounced. As far as we know, this is currently the technologically most advanced e-learning system for teaching L2 pronunciation, although there are still a number of issues. The system is only limited to segmental aspects, (level) of pronunciation and according to [8] the system still "often mistakenly identifies incorrect sounds as correct", thus the problem from ASR technology still partially remains. Another issue is with the feedback, which is limited to the speech sound inventory of English, despite the fact that the sounds pronounced by students often do not correspond to any sound in the target language (in this case English). The last issue leads us to the idea that we need more than a sound inventory of target L2 to create a successful system for providing feedback on the pronunciation of L2 (in our case Czech).

3 NON-NATIVE SPEECH RECOGNITION AND THE FEEDBACK APPROACH

The general idea of our approach is to include non-native sounds into an inventory of the speech recognition system, so we are not limited to the most similar speech sound from the language and thus we can obtain less distorted results of

actual pronunciation. Based on this recognition, we want to provide feedback to students that will tell what was wrong in their pronunciation and how to fix it. The feedback should not be in the form of the pronounced vs correct sound, which can be confusing for a student who usually does not know IPA. The form of feedback should be more explicit, for example, instructing students that their lips should be rounded or mouth more open, etc.

To achieve this target, we first have to collect a large amount of data and create an annotation system that will allow us to tell the differences between the speech sound that should have been pronounced and the speech sound that actually was pronounced. We will then need a tool based on annotated data and capable of recognising a speech sound and its corresponding annotation from recordings. In the first phase, we decided to test this approach on the individual speech sounds of L2 Czech.

3.1 Data: Collection and methodology

For the data collection, we had to take into account the technical aspects of the recordings, which were intentionally taken at varying levels of quality: (1) studio standard (44.1/16, wav); (2) compressed formats (mobile phone). Mobile recordings were used for the annotation and subsequent training to more closely match the quality of the recordings of the future mobile application.

187 foreigners – native speakers of 36 different languages – across all levels of language teaching (using the CEFR scale from A0 to C1) have been recorded thus far.[2] The speakers were recorded during Czech language courses for foreigners at the Summer School of Slavonic Studies at Palacký University Olomouc, as well as at the Center for Foreigners in Brno. All age categories from 18 to 73 years are represented, with the largest group being speakers under 40. The cumulative frequency is as follows: under 25 (66), under 35 (100), under 40 (136), under 50 (176), 50+ (187). In terms of gender, women predominate (114) over men (67) and over unknown (6).

The sample dataset contained isolated speech sounds, as well as two- to four-syllable words or phrases in which a given speech sound appeared in different positions (initial, middle, final) and in different phonemic contexts (vowels, obstruents, sonorants). The data was read twice by the non-native speakers – first with an instructor and then without any assistance. The students were asked to read all the Czech speech sounds in isolation at the end. Only part of the data (from 32 speakers, see below) – a set of segments with isolated speech sounds – has been used thus far to annotate the pilot testing of the recognition model.

[2] The teaching level with the dominance of the lexico-grammatical level does not have to correspond to the level of pronunciation. Representation was the following: A0–A1 (76 speakers), A0–A2 (112 speakers), others.

The design of the annotation system was based on a number of hypotheses and reflected (i) the phonetic basis of Czech, (ii) the phonetic specificities of foreign languages and the relevant phenomena and (iii) the most frequent pronunciation errors among foreigners learning Czech. These hypotheses were postulated from many years of experience with teaching foreigners by one of the authors. Deficiencies in the pronunciation of foreigners can generally be divided into a few different categories [9]:

(1) pronunciation of speech sounds that are not part of the Czech phonetic system, although the student is capable of pronouncing the Czech speech sound; these cases often stem from the written form of the language;

(2) pronouncing the speech sound is only problematic in certain positions or in close proximity to certain other speech sounds;

(3) the student is unable to distinguish two sounds – for speakers of Arabic, this can be [b] and [p], etc.;

(4) the speech sounds are not pronounced in a Czech style, such as when English-speaking students pronounce [p, t, k] with aspiration, etc.;

(5) the speech sounds cannot be pronounced by the student at all, not even approximately.

In creating the annotation system, we tried to take into account the "type and severity" of error, in the sense of: (1) slight deviations without compromising intelligibility – (2) deviations partially compromising intelligibility – (3) significant deviations compromising intelligibility (confusion of meaning, etc.). This categorisation could also be used as a way of providing the students with feedback.

3.2 Phonetic features that most often cause problems for foreigners

Certain speech sounds cause problems for foreigners regardless of their native language – they are difficult for practically everybody. At the segmental level, these are mainly the following phenomena (this is only a very brief and simplified list of the most common pronunciation issues):

- vowels – quantity: while it may be due to a lack of knowledge, certain foreigners may be applying what they are used to from their native languages; there is also the nasal production of vowels or diphthongs, "hard" pronunciation of [ɪ] or [u];

- consonants – the most difficult consonant for foreigners is the trill *ř* (whether in its voiced or unvoiced form: [r̝] [r̝̊]) and the laryngeal *h* [ɦ], which, although it exists in many languages, is not present in them in a voiced form;

- nearly all foreigners struggle with the consonants *ď* [c], *ť* [ɟ], *ň* [ɲ] (in contrast, speakers of Russian, Ukrainian or Azerbaijani frequently incorrectly soften the denti-alveolar *t*, *d*, *n*);
- pronouncing the syllable-forming consonants *l* [l̩] and *r* [r̩] – they are new to most foreigners and difficult to correctly articulate;
- issues with pronouncing Czech sibilants (*s* [s], *z* [z], *c* [t͡s], *š* [ʃ], *ž* [ʒ], *č* [t͡ʃ]) are also common;
- there are issues with distinguishing the voicedness of consonant pairs; tendencies to aspire in the pronunciation of plosives [p, t, k], articulatory issues with the lateral fricative [l].

Contextual or combinatorial phonetic phenomena are very important. The assimilation of voicedness in Czech can be, for example, a phenomenon new to many foreigners and pose issues for some; for some students, pronunciation difficulties are the result of a different articulation base or different assimilation processes (e.g., the tendency to use progressive assimilation, etc.).

4 ATTRIBUTIVE ANNOTATION SYSTEM

We created a formalised ATTRIBUTE–VALUE annotation system based on systematically categorised pronunciation errors from individual languages or language groups. The annotation label is divided by a colon into two main parts: (1) the part before the colon lists the speech sound that was supposed to be pronounced; (2) the attributes after the colon list (using the possible values of the given attribute) the deviations in pronunciation from the standard and the correct phonetic form of the speech sound. If the pronunciation is correct, only the part before the colon is used. In the case of incorrect pronunciation, any number of attributes can follow the colon (see below).

The annotation system specifies two groups of attributes: (1) *fixed*, which have a binary value of 0 or 1 for the phonological characteristics (quantity, voicedness) and (2) *variable*, with the possibility to add other values as needed (phonetic features such as palatalisation, etc.). There is a separate label for replacing one speech sound with another which has the format of X::Y where X = the desired speech sound and Y = the actually pronounced speech sound.

The *;err* tag denotes an unspecified pronunciation error. It can also be optionally supplemented with information on the acceptability of the non-normative pronunciation using the letters A or N to form *;errA* (acceptable) or *;errN* (not acceptable). The tagset labels for attribute values are unique and non-doubled, so ambiguity is not an issue.

Listed below are several examples, the format is always one speech sound per line (the meaning of the tag is explained in square brackets):

o:k1vN [short vowel *o* pronounced as long and nasalised]

a:vNvT [vowel *a* is nasalised with a hard pronunciation]

e [vowel *ɛ* is correct]

ou:vNkD_1 [both parts of the *o͡u* diphthong are nasalised, the first part is lengthened]

t':vR_t'j [the consonant *c* is pronounced in a segmented way with the inserted speech sound *j*]

Explanatory notes[3]:
attributes
k = quantity
v = non-normative pronunciation variants

*values of the **k** attribute*
K shortening
D lengthening

*values of the **v** attribute*
N nasalisation
R "segmented" pronunciation [supplemented by *aspects of value*]
T hard pronunciation

*aspects of values (can be assigned to any value of the **k** or **v** attribute)*
_1 error related to the first part of the diphthong
_2 error related to the second part of the diphthong
_xy *xy* represents the specific speech sounds in the segmented pronunciation

5 TESTING THE NON-NATIVE INDIVIDUAL SPEECH SOUND RECOGNITION

To test our annotation system, we decided to build a minimalistic tool for individual speech sound recognition. This tool was built as a Python script based on the library Persephone [10]. This library is meant as a speech recognition tool for transcription of low-resource languages and contains several parts (see Fig. 1).

[3] The explanatory notes listed below are only the ones relevant for the listed example and are not the complete set.

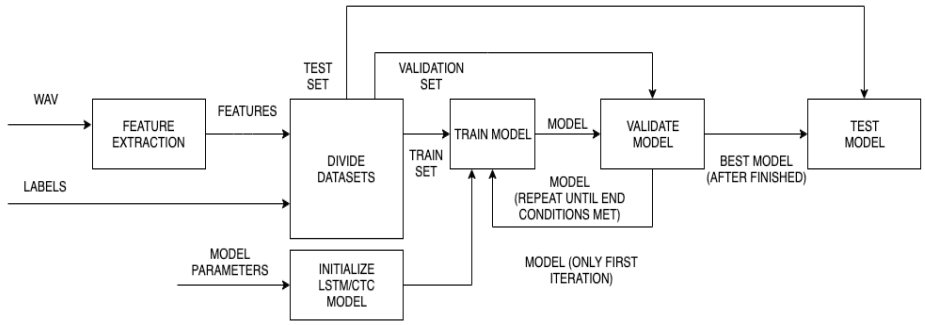

Fig. 1. Individual speech sound recognition tool architecture

The first part is audio feature extraction tools, for our experiment we have used LMFB (Log Mel Filterbank) with delta and delta-delta features. After we extracted features from *wav* recordings, we used Persephone functions to split data (features + labels) into three non-intersecting sets in the following way: 90% of data to train set, 5% to validation set and 5% to test set. We initialised our model when we had prepared our data. "The underlying model used is a long short-term memory (LSTM) recurrent neural network [11] in a bidirectional configuration [12]. The network is trained with the connection's temporal classification (CTC) loss function [13]." [10] We have used default three-layered architecture with 250 hidden nodes. This model is then trained with pre-processed data for at least 30 epochs. Training stops when one of the following 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 tested our trained model against the test data set.

This tool was initially tested on tonal languages and thus it provides the ability to label prosodic features such as tone or word stress. We decided, however, not to use those features as individual labels in the first version of our annotation system. The tool is designed to transcribe whole utterances, however, in our case our utterances are only individual sounds so the label always corresponds to a single speech sound.

For our experiment, we had 3,717 labelled sounds from 32 non-native Czech speakers. When we tried to train the tool with data labelled with the initial version of the annotation, the model stopped after 57 epochs with a huge training error rate 43.4% and a validation rate of 42.4% and an even worse error rate of 50.8% for the

test set. After checking the model results on the test set, however, we found some interesting data. Most of the incorrectly labelled data were consonants and even in those cases, the model output was often a consonant that differed from the expected consonant only in several features such as voicing, articulation position, fricative vs affricate or different variants (aspirated *t* vs "hard" *t*) as shown in Tab. 1.

Table 1 also shows one very interesting case of mislabelling that unveiled one of the issues with annotation. In the last line of Table 1, it is apparent that the expected label was *z::dz* which means that *dz* was pronounced by the speaker instead of *z* and the output label is *dz*. The problem with this is that both of those labels correspond to the same pronunciation, which is *dz*. This leads to an update in our annotation: we have changed the annotation of the incorrect sound from format X::Y (X being the expected and Y being the pronounced sound), to simply Y.

expected label	**output label**
h [ɦ]	ch [x]
s	z
m	n
ch [x]	f
c [t͡ʂ]	dz
z	dz
g	d
t:vA	t:vT
z::dz	dz

Tab. 1. Expected vs output label (consonants)

As can be seen in Table 2, the vowels were in most cases annotated correctly. There were no issues in vowel quality except for cases when diphthongs were classified as simple vowels. This happened especially for diphthongs with a shortened second component or diphthongs and lengthened vowels. There were also some issues with quantity identification. We also observed the same problem with duplicate annotation of format X::Y vs Y.

expected label	**output label**
eu:kK_2 (shortened *u*)	e [ɛ]
eu:vZ_1 (closed *e* [ɛ])	e:vZ (closed *e* [ɛ])
u:kD (lengthened)	u
au [a͡u̯]	a:kD (lengthened)
eu [ɛ͡u̯]	e [ɛ]
ou:kK_2 (shortened *u*)	ou [o͡u̯]
i:kD (lengthened)	i [ɪ]
au::a	a

Tab. 2. Expected vs output label (vowels)

Cases where consonants were annotated as vowels or vice versa were exceptionally rare and for an undiscovered reason. the most frequent error of this kind was labelling *p* as *e* [ɛ]; this could be possibly due to some mistake in the annotation.

After these findings, we decided to go through the annotation and attempt to identify duplicate labels such as the mentioned X::Y vs Y case. By unifying those labels, corresponding to the same sound, and removing a few less important features, we have dramatically reduced the label inventory size, which led to much better results.

The adjusted annotation model stopped after 70 epochs with a much lower training error rate 14.9% and validation error rate 36.8%. The test error rate also improved to 41.27%. Putting aside X::Y vs Y case, errors in the output of the new model were similar to the previous one, although they were less frequent.

There are two main consequences of those results. It is apparent that although we had quite a small data set and the model was far from an optimized one, we ended up with quite good results, although they are still not good enough to be used in a real-world application. The second one is that label inventory size has a huge impact on success rate (along with the amount of available data) and that we have to avoid different labels for the same or very similar sounds at any cost.

6 FUTURE PROSPECTS

The findings from the first testing of our approach lead us to several ideas on how to improve our system. The first plan in the future is to split the annotation into two parts. The first part would be the labels corresponding to each of the individual segments. This would be a simple identifier in the form of a number or character string. These segments would be annotated as *1* or *a1* instead of, for example, *a:kD*. The second part of the annotation will be a mapping table that will translate the identifier to its corresponding attributive annotation that will be used to obtain feedback based on speech recognition output. In this part we would consequently have the information that *1* corresponds to *a:kD*. We also want to try to split certain features such as length, nasalisation, aspiration or stress to individual segments, thus instead of *á* we would have *a:* and instead of *p:vA* (aspiration) we would have *p>* (where > means the aspiration segment). This will allow us to easily extend our annotation, shrink the size of our label dictionary, and focus on the most frequent non-native sounds in Czech.

In conclusion, the field of speech recognition in e-learning and automatic feedback on non-native speech is still in its beginnings, but our findings could become the basis for a new approach to this complex and increasingly important problem. A great deal of this research, however, still has to be done and much data has to be collected to create a system that can be used in e-learning systems. Non-native speech recognition is nevertheless a topic to be considered.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic IGA_FF_2020_021 "Czech Studies: Literary and Linguistic Overlaps and Interpretations" and MUNI/IGA/1225/2020 "L2 Pronunciation system (Linguistic Annotation System)".

References

[1] Blake, R. (2011). Current Trends in Online Language Learning. Annual Review of Applied Linguistics, 31, pages 19–35.

[2] Duolingo, Inc. (2021). Duolingo (5.1.5) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=com.duolingo.

[3] Mikušiak, L. (2015). Ten Ta To (1.7) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=com.lubosmikusiak.articuli.tentato.

[4] EVE Technologies, s.r.o. (2021). CzechME (1.0.5) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=cz.evetech.czechme.

[5] Rootstrap, Inc. (2020). Report: Online Education Industry Growth 2020. Rootstrap [Web page]. Accessible at: https://www.rootstrap.com/annual-report-online-education-statistics.

[6] Kuhn, R., and De Mori, R. (1990). Cache-based natural language model for speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(6), pages 570–583.

[7] ELSA Co, Ltd. (2021). ELSA Speak: Online English Learning & Practice App (6.2.1) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=us.nobarriers.elsa.

[8] Becker, K., and 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.

[9] Hedbávná, B., Janoušková, J., and Veroňková, J. (2009). Výslovnost češtiny u cizinců – poznámky k metodám výuky. In Sborník Asociace učitelů češtiny jako cizího jazyka 2007–2009, pages 16–23, Praha: Akropolis.

[10] 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), pages 3356–3365, Miyazaki, Japan.

[11] Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), pages 1735–1780.

[12] Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), pages 2673–2681.

[13] Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Proceedings of the 23rd international conference on Machine Learning, pages 369–376.