Pronunciation Variants and ASR of Colloquial Speech: A Case Study on Czech

David Lukeš, Marie Kopřivová, Zuzana Komrsková, Petra Poukarová

Institute of the Czech National Corpus, Charles University, Faculty of Arts, Czech Republic

{david.lukes, marie.koprivova, zuzana.komrskova, petra.poukarova}@ff.cuni.cz

Abstract

A standard ASR system is built using three types of mutually related language resources: apart from speech recordings and orthographic transcripts, a pronunciation component maps tokens in the transcripts to their phonetic representations. Its implementation is either lexicon-based (whether by way of simple lookup or of a stochastic grapheme-to-phoneme converter trained on the source lexicon) or rule-based, or a hybrid thereof. Whichever approach ends up being taken (as determined primarily by the writing system of the language in question), little attention is usually paid to pronunciation variants stemming from connected speech processes, hypoarticulation, and other phenomena typical for colloquial speech, mostly because the resource is seldom directly empirically derived. This paper presents a case study on the automatic recognition of colloquial Czech, using a pronunciation dictionary extracted from the ORTOFON corpus of informal spontaneous Czech, which is manually phonetically transcribed. The performance of the dictionary is compared to a standard rule-based pronunciation component, as evaluated against a subset of the ORTOFON corpus (multiple speakers recorded on a single compact device) and the Vystadial telephone speech corpus, for which prior benchmarks are available.

Keywords: speech recognition, pronunciation variants, spontaneous speech

1. Introduction

One of the components of an automatic speech recognition (ASR) system is a pronunciation dictionary, which provides a mapping between a conventional symbolic transcript of speech, which can exhibit varying degrees of arbitrariness, and an acoustically / phonetically motivated one. On the one hand, during the training phase (phone-level alignment, creation of acoustic models, henceforth AMs), the phonetic transcription provides information about where to expect recurring patterns in the acoustic signal, which will be abstracted away and generalized in the acoustic models for the individual (tri)phones. On the other hand, during decoding, it mediates between (to borrow a neurolinguistic analogy) the top-down (predictive) processes based on the linguistic knowledge (the language models, henceforth LMs), and the bottom-up (recognitive) processes based on spotting acoustic patterns in the incoming sound data.

The complexity of the mapping between the standardized orthographic representation of a word and an encoding of its pronunciation as a string of phones depends on the characteristics of the writing system of the language in question. The possibilities range from more or less complete arbitrariness to varying degrees of systematicity in the relationship between graphemes and phonemes (see Sampson (2015) for a book-length account). The variety present in this respect in the world's languages is best attested by sketching a few illustrative points along this continuum:

- Chinese logographic writing bears little to no relation to the phonetic form of words; phonetically speaking, two Mandarin Chinese words such as *mother* and *hemp* may have the same segmental content ([ma]) and differ "only" in their suprasegmental features (1st vs. 2nd tone), yet their traditional written forms will betray none of their acoustic similarity (媽 vs. 麻);

- in general, alphabetic writing systems exhibit higher degrees of consistency in the relationship between graphemes and phonemes, because they are historically built on modeling regularities in speech sounds, but variation is still possible: for instance, contemporary English orthography, though it has evolved over the centuries, still bears the marks of earlier pronunciations and dialectal variants which have been weeded out by sound change in speech but became fossilized in writing, as well as numerous inconsistencies introduced by lexical loans borrowed from other languages (and thus other writing systems);

- conversely, the orthography of a language like Czech is fairly predictable based on pronunciation and vice versa, even increasingly so over the centuries thanks to several spelling reforms (see Kučera (1998, Fig. 5, p. 195), and the whole article for a treatment of the evolution of the efficiency and complexity of Czech spelling), although loanwords with preserved original spelling, mostly coming from English, are a recent disruptive factor.

In the case of Chinese, where the grapheme-to-phoneme correspondence is arbitrary, the only way of creating a pronunciation dictionary is compiling it manually. For English, where the correspondence is obscured by layers of historical development, it is the only *practical* solution (popular resources are the CMU Pronouncing Dictionary[1] for American English and the BEEP Dictionary[2] for British English), because creating a rule-based system would be too time-consuming, error-prone and above all exception-ridden, which means it would have to rely on extensive lists of lexical items to be treated specially anyway, so why not directly store the transcriptions for all items. A sequence-to-sequence mapping tool like *SequiturG2P* (Bisani and

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

Ney, 2008) can then optionally be used as a second step to automatically analyze recurring patterns in the dictionary items (pairs of orthographic and phonetic transcripts) and generate a stochastic grapheme-to-phoneme converter. This can be used as a less reliable[3] fallback procedure, allowing the system to handle out-of-vocabulary (OOV) items in a fairly cost-effective manner, as a "free" side-product of putting together the dictionary itself.

For Czech however, creating such a set of rules is a comparatively easy task (see (Psutka et al., 2006)), so this has become a *de facto* standard approach for the language in NLP applications in general and ASR in particular. Three potential problems with it come to mind:

- irregularities: mild (latinate words) and heavy (English loans, irregularly spelled / pronounced named entities); these may or may not be of interest to the application at hand, and if they are, exceptions for them may be hardcoded *ad hoc* into the rules;
- variants: the higher the frequency of a word, the more syllables it has "canonically" and the less formal the situation, the higher the likelihood that this word will have formally reduced pronunciation variants (see Klimešová et al. (2017, p. 153) for examples from Czech and Ernestus and Warner (2011) for a discussion of phonetic erosion in spontaneous speech in general); of course, dialectal variation can also be subsumed under this heading;
- connected speech processes: a potential problem for any system which considers transcription of lexical items atomically, without paying attention to context.

The inclination to ignore variation may be particularly strong with languages like Czech where a reasonably accurate rule-based pronunciation component can be built with comparatively little effort, but in general, whether the grapheme to phoneme component of the ASR system be lexicon- or rule-based, the latter two sources of variability tend to be downplayed. Since at the same time, they play a significant role in colloquial, spontaneous speech, it makes sense to ask whether this might be a handicap for ASR in some settings.

To explore this topic, the research presented in this paper leverages the ability of current ASR systems to handle a 1-to-many mapping in their pronunciation dictionary components. Whereas traditionally, these have often been "armchair language resources" (in analogy to the notion of "armchair linguist"), our approach was to sift through manual phonetic transcriptions of spontaneous speech and compile a lexicon of commonly attested pronunciation variants. Of course, the 1-to-many mapping addresses only some of the variability issues that ASR systems have to contend with. For a comprehensive overview, see Strik and Cucchiarini (1999); for example, differences in the temporal and spectral properties of instances of the "same" phone belong to the domain of the AM.

[3]Because it can only follow statistical trends and has no way of knowing about lexical exceptions.

2. Data and method

The procedures and overall setup are heavily based on the recipes for the Czech portion of the Vystadial telephone speech corpus (Korvas et al., 2014), more specifically their versions destined for the Kaldi ASR toolkit (Povey et al., 2011). For the most part, changes to the existing code were only cosmetic (adapting transcript and recording file preparation scripts for non-Vystadial data, fixing common typos in transcripts), except for the routines which generated the pronunciation lexicons, which were at the heart of our present undertaking (see 2.1.).

As far as data is concerned, Czech Vystadial data was also used in some of the experiments to offer a few basic comparisons, but the main focus was on recordings from the ORTOFON corpus of informal spontaneous spoken Czech (Kopřivová et al., 2014b). This is a corpus of intimate discourse recorded on a compact device (a Sony ICD-UX5xx series recorder) in natural settings (at home, at work, in cafés etc.), between groups of two or more well-acquainted people. It features two manually created transcription layers (a basic one, which is fairly close to accepted Czech orthography, and a phonetic one), as well as some additional paralinguistic annotation and sociolinguistic metadata (Kopřivová et al., 2014a). Data collection is ongoing and a sociolinguistically balanced sample of about 1M running words in length has been published in June 2017 via the online query interface at `https://korpus.cz`. It is also available for download in two different formats from the LINDAT/CLARIN repository (Kopřivová et al., 2017a; Kopřivová et al., 2017b).

The nature of the material entails a variety of challenges which can be faced with some success by human annotators, but are at present mostly insurmountable for automatic processing. These are generally related to the constraint that recordings are made on a single, preferably inconspicuously placed device, which is in turn dictated by a desire to capture natural linguistic behavior:

- finding a good compromise for placing the recording device and setting microphone sensitivity is hard, some speakers involved in the communication situation tend to be too close, others too far away;
- on a related note, depending on the setting, the speech can occasionally be drowned in noise (domestic appliances, vehicles passing by, café chatter), because microphone sensitivity needs to be set to pick up mid-range distance signals;
- unstructured interactions inevitably result in portions with overlapping speech by multiple speakers.

We keep track of the quality of the recording as an impressionistic rating made by its original transcriber. The sample of raw ORTOFON data used in the experiments reported in this article was selected from a population of utterances which contained no overlaps and were taken exclusively from recordings with the highest quality rating. To facilitate iterative development, the sample was deliberately kept relatively small (see Tab. 1) in order to speed up training and decoding. As the main goal was not absolute performance but mutual comparisons between various approaches to generating the pronunciation dictionary, we feel this is a justified choice.

Table 1: Data sets employed in the experiments and their sizes: length of audio (hours:minutes), number of recordings, number of tokens.

data set	length	# recordings	# tokens
ORTOFON			
train	3:25	3978	38,593
dev	0:28	497	5081
test	0:26	497	4869
Vystadial			
train	15:25	22,567	126,333
dev	1:23	2000	11,478
test	1:22	2000	11,204

Following the original Vystadial experiments, all sound files were converted to mono WAV PCM sampled at a 16 kHz rate and 16 bit depth. Unlike the original Vystadial data however, ORTOFON data comes with information about speaker gender and identity across recordings, so these were specified in the relevant files where Kaldi asks for this information, instead of assuming the same gender everywhere and no recurring speakers.

2.1. Generating pronunciation variants

First of all, the vanilla rule-based pronunciation algorithm provided as part of the Vystadial scripts was used as a baseline (hereafter `vanilla`). It implements the best practice rules of Czech pronunciation as traditionally employed within the NLP community (Psutka et al., 2006) and yields exactly one pronunciation per lexicon item.

Pronunciation variants were then extracted from a working sample of the ORTOFON corpus about 1M running words in size; recall that these are hand-transcribed pronunciations spotted in naturally occurring colloquial Czech. Apart from recognizably legitimate variants, this database also turned out to contain some highly idiosyncratic ones as well as orth-to-phon alignment errors, identifiable as low frequency items or even *hapax legomena*. It quickly became clear that this variability and noise had to be trimmed down in order to become manageable by Kaldi. Two types of approaches were used: an automatic frequency based thresholding heuristic (2.1.1.) and manual filtering (2.1.2.). Consider that the two most phonetically variable lexical items, *protože* (*because*) and *prostě* (*simply*, and also a lexical filler similar to *like*), had 248 and 133 different pronunciations respectively. Given the fairly limited size of the training set, Kaldi needs help in distinguishing which of these to consider as even remotely viable candidates. In general, the goal is to span the continuous space of acoustic variability in as few discrete variants as possible, so that the ASR system only has to deal with useful and meaningful complexity and uncertainty.

Variability also leads to heightened homophony which particularly affects words that are short by nature or prone to drastic formal reduction as a result of their frequency of use. An especially insidious case of homophony is empty pronunciations. Even though these are linguistically well-motivated, because highly frequent function words might be completely elided in informal speech, as the speaker can expect listeners to be able to infer them based on context and their knowledge of the language, they were systematically removed since they might result in the spontaneous addition of words to the transcript during decoding with no corresponding acoustic evidence, if not properly constrained by a sufficiently strong LM.

The rule-based transcription procedure used as fallback for items not occurring in our variants database was based on `vanilla`, with a few small emendations. For instance, assimilation of voicing was reimplemented [4], and *j*-epenthesis between a close front vowel and another vowel was added.

2.1.1. Automatic threshold

The goal of the automatic thresholding procedure was to drastically reduce the maximum number of variants allowed for an item while retaining the spread in variability, i.e. the mapping should strive to preserve a distinction between highly, mildly and marginally variable items. Several options were explored and the following heuristics were retained in the end (L is a list of items from the pronunciation database sorted in decreasing order by their number of variants, M indicates the highest attested number of variants, i.e. the number of variants of the first element of L):

1. lexicon items were split into variability groups by dividing the interval $[0; M]$ into up to N non-overlapping intervals of size at least M/N, always extending the lower boundary to the next attested number of variants if the M/N step fell in between; groups were established based on membership of the items' variant counts in these intervals; items in the 1st group were limited to at most their N most frequent variants, the 2nd group to $N-1$, etc.; these were still subject to the additional filtering heuristics defined below;

2. *hapax* variants were discarded for items with at least one variant which had been seen multiple times;

3. variants containing rare phones, i.e. phones seen less than 10 times in the lexicon generated for a given experiment,[5], were discarded;

4. variants which were short (less than 2 phones) and homophonous (shared by multiple items) were discarded.

[4] The original algorithm uses a cascade of regular expressions. Since assimilation of voicing should spread across multiple neighboring phones, it cannot be implemented in one pass of regular expression substitutions, because these need to be linearly ordered, and new assimilation-triggering contexts may emerge in the process of applying them that would necessitate restarting from the top.

[5] This might seem like a low threshold, but in practice, there was quite a considerable gap between single-figure count phones and the rest.

Two versions of this approach were tested based on the value of N specified in heuristic 1 above, `thresh9` and `thresh4` ($N = 9$ and $N = 4$, respectively).

2.1.2. Manual filtering

A human expert in the phonetics of spontaneous Czech speech went through the pronunciation variants of the 100 most common word forms in the ORTOFON and Vystadial data sets and manually removed those that were deemed too ambiguous or poorly representative. Beyond this frequency peak, transcriptions were produced by the rule-based fallback method; in other words, in this setup, multiple pronunciation variants were allowed only for the most frequent words.

As in 2.1.1., two versions of this approach were tested. `manual1` was more lenient, allowing multiple similar variants both within and across items wherever they made sense and keeping the variants "as is". `manual2` was more aggressive, taking into account which differences are perceptually and acoustically salient and weeding out variants which were judged too similar to other ones. The variants themselves were also altered on occasion, mapping less frequently occurring phones (schwa, labiodental nasal) onto more well-attested counterparts.

We are well aware that the replicability of such a manual procedure is questionable. It would perhaps be better to characterize it by its purpose, which was to act as a sieve which is both *more aggressive* and *more intelligent* than the one defined purely based on frequency rules in 2.1.1.. The human expert acts as a post-editor of the decisions made by the original transcribers of the recordings. As such, s/he should have sufficient prior acquaintance with the material and training in phonetics in order to be able to:

- remove implausible variants
- substitute rare phones with related higher frequency phones[6]
- spot variants which are acoustically similar and retain only a single representative for the entire group

The variant lexicon is the permitted to include only such manually verified items. Together with the overall purpose defined above serving as guiding principle, these are the essential parameters of the manual filtering procedure to bear in mind when replicating it.

3. Results and discussion

AM training and LM creation followed the Vystadial recipe. We therefore performed experiments with zerogram and bigram language models[7] and the following acoustic models:

[6]For purely practical reasons: when training the ASR system, it is unlikely that a meaningful generalization would be inferred from just a handful of exemplars. Cf. the focus on "useful and meaningful complexity and uncertainty" in 2.1..

[7]Inferred from test and train data, respectively; the purpose of the zerogram model is "to evaluate solely the quality of the acoustic models without being affected by a language model or presence of out-of-vocabulary (OOV) words in the test set" (Korvas et al., 2014, p. 4426).

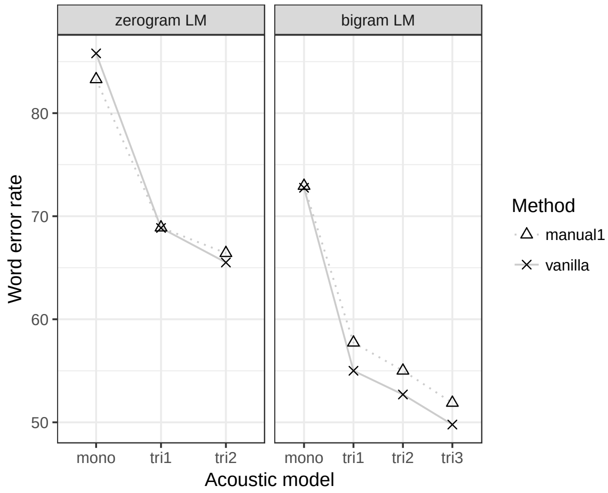

Figure 1: Word error rate results for different acoustic models, language models and pronunciation dictionary generation methods applied to the Vystadial data.

- `mono`: monophone model trained with MFCCs, Δ and $\Delta\Delta$ features;
- `tri1`: basic generative triphone model trained using Viterbi training; corresponds to `tri` $\Delta + \Delta\Delta$ in (Korvas et al., 2014);
- `tri2`: triphone model with Linear Discriminative Analysis (LDA, Haeb-Umbach and Ney (1992)) and Maximum Likelihood Linear Transformation (MLLT, Gopinath (1998)) feature transformations, trained using alignments from `tri1`; corresponds to `tri LDA+MLLT` in (Korvas et al., 2014);
- `tri3`: triphone model trained using discriminative Boosted Maximum Mutual Information (BMMI, Povey et al. (2008)) training on top of `tri2`; corresponds to `tri LDA+MLLT+BMMI` in (Korvas et al., 2014).[8]

Results are presented visually in Fig. 1 (for Vystadial data) and Fig. 2 (for ORTOFON data).

The Vystadial experiments were more computationally intensive because of the size of the data, so only `vanilla` and `manual1` methods were run. The `vanilla` results are comparable to those reported in (Korvas et al., 2014), suggesting that our basic experimental setup was sound. The word error rate (WER) for `manual1` is similar under the zerogram LM, suggesting that the acoustic models themselves are neither hampered nor (sadly) improved by having access to a wide array of pronunciation variants of highly frequent words. Under the bigram LM, except for the `mono` AMs, `manual1` is clearly outperformed by `vanilla`, suggesting that the amount of homophony

[8]Note that this model "needs a *language model* (LM) in order to compute the objective function. Here we use the [aforementioned] bigram LM" (Korvas et al., 2014, p. 4426), so the decoding performance with the zerogram LM has no clear interpretation and will be omitted from the results.

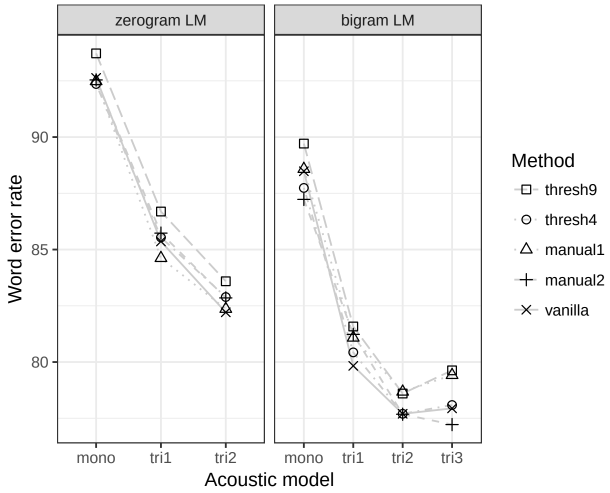

Figure 2: Word error rate results for different acoustic models, language models and pronunciation dictionary generation methods applied to the ORTOFON data.

introduced by the variants prevents the system from efficiently taking advantage of the predictive power of the LM when decoding.

With ORTOFON data, the more lenient automatic thresholding method `thresh9` evidently retains too much variability to be able to compete with the other ones even under the zerogram LM. As with Vystadial, `manual1` performs roughly on par with the others under the zerogram LM but lags behind along with `thresh9` under the bigram LM. The more stringent `thresh4` and `manual2` remain competitive with `vanilla`, with `manual2` providing the only hint that pronunciation variants might be performing useful work, as contrasted with `vanilla`, under the bigram LM and `tri3` AM: it is the only method that improves, if slightly, when BMMI is added. Of course, further work on larger data sets and more careful and systematic manual tweaks of the variants database will be needed to determine if this is indeed a viable path to follow.

In general, there seems to be a clear divide between the type of pronunciation variants human experts use to encode hints for other linguists / phoneticians to use in doing linguistic research, and the type of variation information an ASR system implementing current state-of-the-art methods can leverage. Or, to be more precise, the former needs to be thinned down quite drastically in order to become the latter, and even then, benefits are hard to glean. In light of this, it seems pointless to invest energy and resources into trying to extend the pronunciation rules for Czech to yield more than one pronunciation per item and thus account for empirically unattested but theoretically possible variants differing in voicing assimilation, vowel length or elision. However, judiciously adding attested high frequency variants should not be ruled out.

Frequency-based heuristics, whether implemented fully automatically or with manual cleanup, are fairly efficient at this sieving, but especially the manual ones, which only look at the frequency peak, also get rid of many potentially useful pronunciations of less common irregular words. Ideally, we should have additional criteria for inclusion that circumvent this drawback, e.g. based on the minimum edit distance between the hand-transcribed variant and the rule-generated variant.

Another topic for future research is that we have not yet investigated the option of providing frequency-based weights for the pronunciations, as Kaldi allows the possibility of a probabilistic pronunciation lexicon, instead of using binary thresholding mechanisms, or perhaps a combination of both approaches. Even more remotely, we may have discarded empty pronunciations as mentioned in 2.1., but they remain a valid notion from the linguistic point of view. They might just yet prove helpful in combination with stronger, more constraining LMs which do a better job at modeling language knowledge.

4. Conclusion

There is a hackneyed adage in the NLP community according to which firing the linguist / phonetician on the team invariably leads to an improvement in one's performance metric. The experiments reported in this article seem to corroborate this piece of folk wisdom, to the extent that linguists seem to be interested in details of variation which, while empirically motivated and linguistically relevant, result in confusion and an explosion of the search space when fed into an ASR system which has to bootstrap itself on limited amounts of data.

However, as we've seen with `manual2`, linguists might still have a useful contribution to bring to the table every now and then, if ever so modest, as long as they stick to another cliché saying: Less is more when it comes to pronunciation variants, at least in terms of how they fit together with the remaining components of current state-of-the-art ASR systems.

5. Acknowledgements

This research was made possible by the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Rep. within the framework of Large Research, Development and Innovation Infrastructures.

6. Bibliographical References

Ernestus, M. and Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39:253–260.

Gopinath, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. In *Proceedings of ICASSP*, pages 661–664.

Haeb-Umbach, R. and Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *IEEE International Conference onAcoustics, Speech, and Signal Processing*, volume 1, pages 13–16. IEEE.

Klimešová, P., Komrsková, Z., Kopřivová, M., and Lukeš, D. (2017). Avenues for corpus-based research on informal spoken Czech. In Piotr Pęzik et al., editors, *Language, Corpora and Cognition*, volume 51 of *Łódź Studies in Language*, pages 145–162. Peter Lang Edition.

Kopřivová, M., Goláňová, H., Klimešová, P., Komrsková, Z., and Lukeš, D. (2014a). Multi-tier transcription of informal spoken Czech: The ORTOFON corpus approach. In *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*, Olomouc Modern Language Series, Vol. 4, pages 529–544. Univerzita Palackého.

Kopřivová, M., Klimešová, P., Goláňová, H., and Lukeš, D. (2014b). Mapping diatopic and diachronic variation in spoken Czech: The ORTOFON and DIALEKT corpora. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 376–382. European Language Resources Association (ELRA).

Kučera, K. (1998). Vývoj účinnosti a složitosti českého pravopisu od konce 13. do konce 20. století [Development of orthographic efficiency and complexity in Czech texts from the end of the 13th century to the end of the 20th century]. *Slovo a slovesnost*, 59(3):178–199.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., and Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060.

Psutka, J., Müller, L., Matoušek, J., and Radová, V. (2006). *Mluvíme s počítačem česky [Speaking Czech to the Computer]*. Academia.

Sampson, G. (2015). *Writing Systems*. Equinox Publishing Limited, second edition.

Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29:225–246.

7. Language Resource References

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017a). ORTOFON v1: balanced corpus of informal spoken czech with multi-tier transcription (transcriptions). `http://hdl.handle.net/11234/1-2580`.

Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017b). ORTOFON v1: balanced corpus of informal spoken czech with multi-tier transcription (transcriptions & audio). `http://hdl.handle.net/11234/1-2579`.

Korvas, M., Plátek, O., Dušek, O., Žilka, L., and Jurčíček, F. (2014). Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4423–4428.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.