Computer-Aided Quality Assurance of an Icelandic Pronunciation Dictionary

Martin Jansche
Google Inc.
New York, NY 10011, USA

Abstract

We propose a model-driven method for ensuring the quality of pronunciation dictionaries. The key ingredient is computing an alignment between letter strings and phoneme strings, a standard technique in pronunciation modeling. The novel aspect of our method is the use of informative, parametric alignment models which are refined iteratively as they are tested against the data. We discuss the use of alignment failures as a signal for detecting and correcting problematic dictionary entries. We illustrate this method using an existing pronunciation dictionary for Icelandic. Our method is completely general and has been applied in the construction of pronunciation dictionaries for commercially deployed speech recognition systems in several languages.

Keywords: Pronunciation dictionary, letter-phoneme alignment, Icelandic

1. Overview

Pronunciation dictionaries are crucial language resources for building automatic speech recognition (ASR) and synthesis systems. A pronunciation dictionary defines how the familiar orthographic form of a word relates to its pronunciations. The use of this is obvious in speech synthesis: in order to generate synthetic speech for a given input sentence, a synthesis system needs to be able to map words to sound sequences. The same problem applies inversely in ASR: a speech recognizer produces ordinary text as output, deriving it from a speech signal via intermediate phonetic representations. Here too the system needs to know about the relationship between pronunciation and orthographic spelling of words.

In the simplest case a pronunciation dictionary is a collection of orthographic words paired with their phonemic transcriptions. The construction of pronunciation dictionary typically involves human experts transcribing orthographic word forms into some form of phonemic notation, such as IPA, X-SAMPA, ArpaBet, etc. Here our focus is on methods for assuring the quality of pronunciation dictionaries. We are specifically concerned with the problem of checking if a given spelling/pronunciation pair is plausible. Such quality checks can be applied online to the work of human transcribers, providing immediate feedback about transcription quality. Or they can be applied offline to verify and improve the quality of existing dictionaries. Here we describe our method as it applies to an existing Icelandic pronunciation dictionary.

Our method is based on computing a constrained alignment between the orthographic and phonemic form of a dictionary entry. Such an alignment establishes a mapping between letters and phonemes subject to constraints about which letters and phonemes can correspond to each other. By developing a tight set of constraints, one can flag erroneous entries in which spelling and phonemic transcription do not correspond to each other in various ways. For example, the letter “u” can be pronounced in various ways in Icelandic: e.g. the suffix “-unum” is pronounced as /ɔnʏm/, with distinct phonemes corresponding to each occurrence of the letter “u”. However, the letter “u” can never be pronounced as the phoneme /t/ or similar consonants. We record facts like this in the form of automatically checkable constraints, which allows us to flag entries that violate these constraints.

2. Language and language resources

Icelandic poses a particular challenge for generating pronunciation resources, since the language is morphologically rich, with long compound words and inflected forms. Moreover the Icelandic writing system is fairly deep, in the sense that the pronunciation of a word form may not correspond straightforwardly to its spelling (unlike e.g. in Spanish), due to cluster simplification, assimilation, and other phonological phenomena. For example, “gagnvirkt” (‘interactive’) can be transcribed phonemically in IPA as /kakvɪr̥t/, reflecting the simplification of the orthographic “gnv” cluster into phonemic /kv/ and of the orthographic “rkt” cluster into phonemic /r̥t/.

We started with a version of the Pronunciation Dictionary for Icelandic developed at the University of Iceland for the Icelandic speech recognition project Hjal (Rögnvaldsson, 2004). This dictionary consists of approximately 60,000 word forms of modern Icelandic, including inflected native words forms (e.g. “utanríkisviðskiptaráðherrann”), important acronyms (e.g. “ADSL”), as well as proper names using both native (e.g. “Atlantshafsbandalagið”) and foreign/borrowed (e.g. “Netscape”) forms. The dictionary provides phonemic transcriptions in IPA and, equivalently, in a customization of SAMPA (Rögnvaldsson, 2003). We use SAMPA transcriptions in our work and the following illustrations. The phonemic transcriptions use 60 distinct phonemes plus one boundary marker. The spellings consist of the 32 letters of the Icelandic alphabet, plus a few additional letters used in foreign words (“c”, “w”, “z”, and “ø”), as well as a stop/period used in abbreviations and a dash/hyphen used in compound words.

3. Method

Consider the spelling and pronunciation of “Eyjafjallajökull”. The following table shows one plausible way to align its letters and phonemes:

```
ey  j a f j a ll  a j ö  k u ll
ei: j a f j a d+l a j 9: g Y d+l0
```

Each column is a pairing of zero or more letters (top row) and zero or more phonemes (bottom row). For example, the first two letters, "ey", correspond to the single phoneme /ei:/ and the last two letters "ll" correspond to the phoneme sequence consisting of /d/ followed by /l0/.

Alignments like this can be computed straightforwardly by dynamic programming, essentially along the lines of string edit distance and sequence alignment problems in text and speech processing and computational biology. Stochastic alignment computation requires an initial probability matrix as a side input, which specifies the probability of the atomic letter/phoneme pairings that will be considered as the basis for alignments. Such probabilities can be re-estimated via a straightforward instantiation of the EM algorithm (Ristad and Yianilos, 1998). However, this leaves open the issue of how the atomic pairings and their probabilities (i.e. the structure and parameters of the alignment model) should initially be specified.

One common solution is to enumerate all possible pairings of letter and phoneme strings up to a given length and to initialize the matrix with random positive weights and normalize appropriately. This results in a very large matrix that will become increasingly sparse with re-estimation. Crucially, this way of specifying the cost matrix is entirely non-parametric, in the sense that it will fit any data points provided, including erroneous entries in the pronunciation dictionary. If sparsity of the matrix were enforced by regularization, only non-productive entries would be set to zero, but errors in the data would persist.

Our approach starts from the opposite direction. We specify the alignment model by enumerating plausible letter/phoneme pairings by inspection of successively increasing portions of the data. Concretely, we start with manually specified pairings for all 32 letters of the Icelandic alphabet (e.g. the letter "f" is often pronounced as /f/, sometimes as /v/, etc.). We then automatically align a portion of the dictionary and keep track of complete alignment failures resulting from zero-probability events due to potential letter/phoneme pairings that are not yet part of the alignment model. An increasingly richer alignment model is built up in an iterative fashion until it fits the entire portion of the data that was set aside for model building.

Failure to align can be a valuable signal for finding problems with lexical entries. Due to the heterogeneous nature of a typical pronunciation dictionary, the interpretation of alignment failures requires careful inspection of model and data. The status of alignment failure vs. success in different situation is summarized in Table 1. Keeping in mind that we view transcription errors as the outcome which we want to detect and alignment failure as a positive signal we can distinguish the following scenarios:

1. True positives occur when a given ordinary words and its phonemic transcription do not correspond to each other and the model correctly fails to produce an alignment. This type of failure reveals actionable problems in the pronunciation dictionary.

2. True negatives occur when a given ordinary word and its phonemic transcription correspond meaningfully to each other and alignment under a given model succeeds. In quality assurance work, most dictionary entries do not contain mistakes and this scenario is very frequent.

3. Type I errors, or false positives (FP), occur when a given spelling of an ordinary word and its phonemic transcription do in fact correspond meaningfully to each other, but the model fails to produce an alignment. These types of failures tend to occur in the early stages of work, when the alignment model does not yet know about certain letter/phoneme pairings and is thus not rich enough to produce an alignment. This is a kind of deficiency in model structure or parameters that can be addressed easily, by adding the missing letter/phoneme pairing to the model specification. An important aspect of the modeling work is to repeatedly test the model against data and enrich it until until it matures to the point where false positives have been eliminated or reduced as much as possible.

4. Type II errors, or false negatives (FN), occur when a given spelling of a word and its phonemic transcription do not correspond to each other, but the model still produces an alignment. These types of error typically indicate that the alignment model is too rich and overly permissive. These errors go undetected in our setup, since only failure to align is used as a signal. Type II errors therefore reduce the usefulness of this signal, i.e. they lower the recall of problematic alignments. By constructing alignment models in a bottom-up fashion, we typically limit the amount by which we overshoot and make the model too rich. Many additional signals for model complexity are available that can help reveal latent type II errors, including: the number of alignment pairs (model parameters), how often a given alignment pair occurs in alignments, inspection of the occurrences of infrequent alignment pairs, inspecting words with many alignments and removing redundant alignment pairs, etc.

5. What we informally call a type III error in Table 1 is in fact a breakdown of basic assumptions about the generative process that the alignment model is part of: the model and the data are mismatched in a fundamental way that cannot and should not be remedied by enriching the model. This type of alignment failure occurs when a given spelling and phonemic transcription correspond to each other in a way that the model was not designed to capture. For example, "ADSL" is pronounced letter-by-letter as /a # d j E: # E s # E d l0/, but the alignment model treats it like an ordinary word and cannot explain the presence of the /E/ and /E:/ phonemes, among other things. Foreign words like "Netscape" have a transcription that a model for native Icelandic words cannot predict. In this case the correspondence between the letter "a" and the phoneme /ei/ is foreign to the Icelandic writing system. We deal with these cases by treating them

	ordinary word		other word	
	transcription error	correct transcription	transcription error	correct transcription
alignment failure	true positive	type I error (FP) model too small	lucky positive	type III error model mismatch
alignment success	type II error (FN) model too big	true negative	type II error (FN) model too big	lucky negative

Table 1: Alignment failure as signal for detecting transcription errors

as exceptions: for purposes of quality assurance, we exclude these entries from the portion of the dictionary we want to validate and optionally process them with alternate alignment models specialized for these modes of pronunciation.

6. In a small number of exceptional situations we might be lucky in the sense that an alignment success or failure happens to be right, but for the wrong reason. An example of this – from English, not Icelandic – is the abbreviation "AI" and its phonemic transcription /eI aI/. The transcription is correct and this pairing aligns cleanly under a sensible but simple alignment model for ordinary English words (since the English letter "a" is often pronounced /eI/ and the letter "i" is often pronounced /aI/). However, the entry itself is an acronym or letter sequence, not an ordinary word, so the success of the alignment is more or less a lucky accident.

True positives point to deficiencies of the data which can only be revealed by informative alignment models. Non-parametric models, by contrast, take all training data at face value and produce alignments for all training pairs. (Non-parametric models can fail on rare occasions on unseen test data, but this is entirely due to unseen pairings. These types of failures do not provide a useful signal for data quality issues.)

This suggests that the goal of our modeling work then ought to be to maximize the usefulness of alignment failures as a signal for data problems. This is achieved by minimizing type I and type III errors, which maximizes true positive alignment failures. Our iterative model refinement starts with a small model with very few parameters (alignable letter/phoneme pairs). We use this model to align the development data, sample a small number of alignment failures, and decide for each entry:

- If the entry is not an ordinary native word (it might be an acronym, foreign word, etc.), we exclude it from the model building data. This reduces type III errors.

- Otherwise, the entry is an ordinary native word and we inspect its phonemic transcription. If the transcription is intuitively correct, we enrich the model by adding alignable letter/phoneme pairs to make the entry align. This reduces type I errors.

- The remaining alignment failures are true positives, signaling genuine problems in the data. We manually correct the transcriptions, which is the main objective of this effort.

With the updated richer model, we re-align the data and sample a new set of alignment failures. This process stops naturally when all model building data – typically the entire pronunciation dictionary – align cleanly without alignment failures. This does not guarantee that the dictionary is without flaws: latent problems can remain due to type II errors, due to data that were excluded because they do not fit the assumptions of the generative process, or because the chosen generative alignment process is not powerful enough to make stronger statements about the data. We can control for type II errors by pruning the model back (for example, by removing parameters with zero or low expected counts). We can verify the quality of excluded data with specialized models – for example, checking the quality of acronyms is easy and only requires a small dictionary of letter names.

4. Results

We applied the above method to the existing Pronunciation Dictionary for Icelandic described above. We manually specified an alignment model and iteratively refined it until it reached approximately 400 parameters, including certain letter sequences of length zero to five and phoneme sequences of length one to five. A complete model over all combinations of letter and phoneme strings of the specified lengths would have had nearly 27.4×10^{15} parameters, of which all but a few hundred are not meaningful. We used the alignment model to align the entire dictionary, consisting of nearly 60,000 word forms. Our constrained model revealed many data quality issues in the existing dictionary. Many of these are due to subtle but easy-to-make mistakes. Because transcription mistakes are overall quite rare, the problems that were uncovered would have been very difficult and/or time-consuming to detect by proofreading without computer assistance. Our model-driven approach not only helped us focus on the data quality problems, but also yielded an alignment model that can be used as the first step in building pronunciation models.

Examples of data quality problems that were detected and fixed include:

- Inspection of alignment failures resulting from processing the entire 60,000-word dictionary revealed 450 entries with data quality problems. These were then manually reviewed and fixed. The range of mistakes that were detected using constrained alignments include typos in the transcriptions that resulted in non-existent phonemes, typos involving substitution of phonemes that can be easily mistyped but cannot be substituted for each other (such as /b/ and /g/), omission of phonemes or whole syllables in the transcription, extraneous phonemes in the transcription, transcription that matches entries before or after the current entry, etc.

- In a few cases the spelling of words was fixed. This happened when the phonemic transcription of a word in the dictionary appeared plausible, but the spelling contained obvious mistakes, such as missing or extraneous letters.

Below are several examples of observed alignment failures that required corrections to the dictionary:

- Transcription of “stefnuskránni” missing the final /nI/:

s	t	e	f	n	u	s	k	r	á	**nn**	**i**
s	d	E	b	n	Y	s	g	r	au:		

- Transcription of “orgelsins” (genitive, definite) contains extraneous phonemes /eig/:

o	r	g	e	l			s	i	n	s
O	r	g	E	l	**ei**	**g**	s	I	n	s

- Transcription of “öfluga” uses /g/ as the pronunciation of the letter “f”, instead of the expected /b/:

ö	**f**	l	u	g	a
9	**g**	l	Y	G	a

All of these examples pass quality checks that look at orthography or transcription in isolation. In all cases, the spelling appears to be free of typos and the phonemic transcription is not obviously ill-formed. It is only by trying to match up the letters and phonemes that problems begin to become noticeable. The constraints in our informative alignment model are crucial as well: without them the impossible alignment of letter “f” to phoneme /g/ in “öfluga” would go undetected, as would be the case in a nonparametric model.

In addition to using alignment checking for quality assurance, all entries in the entire dictionary were classified as belonging into three categories: (1) native and native-like words, defined theoretically as those that follow Icelandic spelling rules and operationally as those which pass our alignment constraints cleanly; (2) words that are pronounced character by character such as “ADSL”; (3) everything else, including foreign and exceptional words.

Our modified dictionary is a key ingredient of a commercial speech recognition system for Icelandic, which has been publicly available since 2012 (Damiba, 2012). The alignment model developed as part of the cleanup of the pronunciation dictionary forms the basis of the statistical pronunciation model in our Icelandic speech recognizer. In addition, the classification of words has proved useful for building pronunciation models for this ASR system. When designing pronunciation models, it is important not to mix different modes of pronunciation within the same low-level mode. For example, a model for native Icelandic words should not be trained on words like “ADSL”, where the relationship between letters and phonemes is different from what it is in ordinary words. Our method allows us to identify only the high-quality native Icelandic words, which pass the constrained alignment checks, and to use this subset as training data for high-quality pronunciation models.

Our method is generic and applicable to many languages and writing systems. The method and associated tools are being used to monitor the quality of pronunciation resources in English, French, Italian, Spanish, German, Portuguese, and Korean, among other languages. The model refinement process proved straightforward and the number of parameters of the models typically reflects the complexity or depth of the respective writing systems. The following table shows the sizes of alignment models for a few languages:

	parameters
Spanish	78
Italian	81
Portuguese	163
Korean	174
German	183
French	214
Icelandic	406

In all cases, loanwords and exceptional words with unusual alignment properties were excluded. The ranking of model sizes corresponds to the intuitive orthographic depth of the various languages. Spanish is a typical example of a shallow orthographic system, whereas Icelandic writing proved considerably more complex.

5. Conclusions

We have described a general method for verifying the quality of pronunciation dictionaries. By computing constrained alignments between orthographic and corresponding phonemic forms, we can flag impossible or very infrequent correspondences of letters and sounds. We applied our method to an existing pronunciation dictionary for Icelandic, resulting in a modified dictionary which passes a large (but undoubtedly incomplete) set of quality checks. Our modified dictionary is a key ingredient of a commercial speech recognition system for Icelandic. We are planning to contribute our modifications back to the original liberally licensed dictionary resource, with the expectation that it will enhance the state of Icelandic language technology.

6. Acknowledgements

Many thanks to Bertrand Damiba, Eiríkur Rögnvaldsson, Kristinn Guðjónsson, Ravi Rajakumar, Trausti Kristjánsson, and the anonymous reviewers.

7. References

Damiba, B. (2012). Voice Search arrives in 13 new languages. *Official Android Blog*, 2012-08-17.

Ristad, E. S. and Yianilos, P. N. (1998). Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Rögnvaldsson, E. (2003). Phonetic transcription guideline: Icelandic. Technical report, ScanSoft Inc.

Rögnvaldsson, E. (2004). The Icelandic speech recognition project Hjal. In Holmboe, H., editor, *Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000–2004*, pages 239–242. Museum Tusculanums Forlag, København.