Towards an acoustically-validated phonetic corpus of spoken Swedish.

Jim O'Regan
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
joregan@kth.se

Jens Edlund
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
edlund@speech.kth.se

Abstract

In this document, we describe ongoing work towards the creation of a phonetically transcribed corpus of spoken Swedish, with aims towards creating a pronunciation dictionary that takes into account dialectal variation. Using speeches from the Swedish Riksdag (Parliament), we use the output of a phonetic recognition system to validate pronunciations for a variety of Swedish dialects.

1 Introduction

"Today, every reputable dictionary makes at least some use of corpus evidence" (Hanks, 2020). This is true not only in the case of monolingual dictionaries, in the selection and ordering of word senses, but also in bilingual dictionaries (e.g., Ó Mianáin and Convery (2014)). Where this is not true, however, is in the case of the pronunciation dictionary.

Although dedicated pronunciation dictionaries intended for human readers are somewhat rare, they are available for use in several areas of speech technology, for example, in text-to-speech (TTS) and computer-assisted pronunciation training (CAPT).

Like other types of dictionary, pronunciation dictionaries are difficult and expensive to create manually, and where they exist, they typically tend to be restricted to one or two standard dialects.

Languages with an alphabetic script typically follow a known set of rules, and grapheme-to-phoneme (G2P) conversion can automate the creation of pronunciation dictionaries, though how effectively depends on the orthography of the language in question, and dictionaries for technological purposes typically require pronunciations for words that do not follow the rules of the language proper, such as foreign names.

These rules can be augmented in stages to handle dialectal variation, assimilation to adjoining words, adjustments in speaking rate, etc., but the interactions between these rules can lead to a large number of possible candidate pronunciations, not all of them realistic.

Although Swedish has relatively few truly exceptional pronunciations, it has a "deep" orthography: for example, with some loanwords it can be necessary to know the donor language to predict the pronunciation. Many words can have multiple valid pronunciations which depend on speaking rate or register; as with most other languages, the pronunciations of words in connected speech change depending on the adjoining words.

Swedish is part of a dialect continuum with Danish and Norwegian, and there are Swedish dialects that share features with both of these languages. The Stockholm dialect is typically considered the standard dialect of Swedish, and existing pronunciation dictionaries tend to target it exclusively.

Our aim in this work is to construct a corpus of phonetically annotated spoken Swedish, representing multiple dialects, and with provenance in the form of identifiers pointing to the source of each speech, with timestamps for each spoken word.

Our initial aim is to create pronunciation dictionaries for use in text-to-speech, but the corpus itself is of general interest for the study of Swedish in particular, and we aim for it to be of interest for speech science more generally.

2 Data

Our data comes from the recorded speeches of the Swedish Riksdag (Parliament). The official transcripts comprise the official record of debates in Riksdag, to which are supplemented recordings of the sessions. Parliamentary speech, at least in the Swedish case, exists at a point between read

and spontaneous speech: the official transcripts are prepared from a pre-filed version of the speech to be read, but the speaker may diverge from this script in reaction to other events, or based on their own speaking style.

Representatives to Riksdag come from all parts of Sweden, and tend overwhelmingly to speak using their own dialects. Riksdag recordings are therefore a valuable resource for Swedish dialects, with the notable exception of Finland Swedish.

3 Method

Because the official transcripts are not always an exact representation of the speech that was delivered, we used speech recognition on the recordings of each speech to find matching sentences, using the model described by Malmsten et al. (2022), based on wav2vec 2 (Baevski et al., 2020).

In addition to simple matches, we ran a number of filters in a number of stages on the unmatched portions to account for a number of phenomena, such as pairs of sentences being joined, denormalisation (numbers, abbreviations, etc.), alternate spellings, predictable recognition errors, and so on.

To create the phonetic transcriptions, we fine-tuned the same model on the phonetic transcriptions of the Waxholm dataset (Bertenstam et al., 1995), which we aligned to the validated sentences using timestamps.

The resulting pairs of word and transcription are then compared against a reference pronunciation, possibly generated and/or expanded using a variety of rule sets to account for assimilation, speaking rate, and dialect.

4 Ongoing work

Our work to date has concentrated on the recordings made from 2010 to 2020; almost 50 years of further recordings are available. The older recordings have been digitised, both from audio and video recordings.

Much of our current focus is on extracting more sentences from the data we have, to maximise our use of the remaining data. A great number of sentences in the data differ only in the placement of a phrase: for example, what appears at the end of a sentence in the transcript might have been read at the start of the same sentence. We are investigating the use of dependency parsing to locate such cases, as they ought to appear as different serialisations of the same tree.

A second phonetic recognition pass is being planned using Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). One limitation of the model we used is that the large size of the context window can sometimes cause short sounds to be discarded. Older methods, such as those used in MFA, use a shorter window and are less prone to discarding such sounds. MFA also provides phoneme-level timestamps. Swedish is a language with pitch accent, and while tools exist to extract pitch information, having the ability to align the pitch contour to the vowels of the word will greatly help in making determinations of which accent applies to a word.

Acknowledgments

The authors thank the researchers from KBLabb at the National Library for providing us with a version of their speech recognition model that we used to create the phonetic recognition model. The corpus will be made more widely accessible through the Swedish Research Council funded national infrastructure Nationella Språkbanken and Swe-Clarin (2017-00626).

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proc. NeurIPS 34*, NIPS '20. Curran Associates Inc.

Johan Bertenstam, Blomberg Mats, Rolf Carlson, Kjell Elenius, Björn Granström, Joakim Gustafson, Sheri Hunnicutt, Jesper Högberg, Roger Lindell, Lennart Neovius, Lennart Nord, Antonio de Serpa-Leitao, and Nikko Ström. 1995. Spoken dialogue data collected in the Waxholm project. *STL-QPSR*, pages 50–73.

Patrick Hanks. 2020. *English Dictionaries and Corpus Linguistics*, Cambridge Companions to Literature, page 219–239. Cambridge University Press.

Martin Malmsten, Chris Haffenden, and Love Börjeson. 2022. Hearing voices at the National Library–a speech corpus and acoustic model for the Swedish language. In *Proc. of Fonetik 2022*, Stockholm. KTH Royal Institute of Technology.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Pádraig Ó Mianáin and Cathal Convery. 2014. New English–Irish Dictionary. *Dictionaries: Journal of the Dictionary Society of North America*, 35:318–333.