Towards quality transcriptions of large, limited domain archive data

Jim O'Regan, Jens Edlund
Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
joregan@kth.se, edlund@speech.kth.se

Abstract

Despite the proliferation of speech data in public archives, research in areas such as speech science and, more broadly, (digital) humanities and social science is often hampered by the lack of search facilities. Recent advances in automatic speech recognition (ASR) help, but custom ASR systems, specifically adapted to the characteristics of an archive, can still outperform more general solutions.

In this work, we explore the use of general-purpose ASR to obtain training data for archival transcription. We use Swedish parliamentary data as a near-ideal example in which existing partial transcripts can be used to support the process, and aim to apply these methods to less ideal archives, such as broadcast news, where the only associated text resources are weakly related materials such as contemporary news from print media.

Results indicate that while fine-tuning on the intersection of outputs from two ASR systems does not surpass state-of-the-art performance, it does show promise for further refinement.

Introduction

Parliamentary recordings are interesting for many reasons, not least from a speech science and speech technology perspective. In addition to their well-established use for training speech-to-text systems, or *automatic speech recognition* (ASR), they are rich in unique characteristics that are interesting in their own right. Because they cover a large time span, they can serve as a primary source for diachronic investigations. They also serve as a bridge between read and spontaneous speech: although speakers typically begin with a prepared script, they often deviate in response to other speeches or events during the same session. And as parliamentary speakers come from all regions of a country, they contain a wide range of dialects. Furthermore, the diversity of topics discussed leads to both detached discourse, speeches brimming with affect, and displays of sentiment for rhetorical effect.

We present an investigation that utilises ASR to iteratively improve automatic transcriptions of Swedish parliamentary recordings with several parallel objectives. Beyond creating accurate transcriptions, we seek a methodology that allows us to use general-purpose ASR to efficiently obtain task-specific training data for archival speech materials. We focus on parliamentary speech with an eye towards other limited domain archives that share similar characteristics. For example, broadcast news also involves a finite number of speakers under similar speaking conditions.

The development of bespoke ASR systems we aim for will bridge gaps in the public record, broadening the scope of research outside speech technology (or "AI") and speech science, for example in digital humanities.

Our intention is that this work be used as a yardstick for other planned projects, both for Swedish and the other languages of Sweden, as it represents an approximation of ideal conditions. The experiences here will determine our

approach towards projects where conditions are less favourable and serve towards directing our efforts where they are best placed, not least within the scope of the creation and maintenance of national research infrastructure for speech technology and speech science in Språkbanken Tal.

Background

Parliamentary data in ASR training

The availability of transcripts makes parliamentary recordings attractive for the development of ASR systems: for many languages, it is the only source of transcribed speech in sufficient quantity of speech and diversity of speakers to train an ASR system (e.g., Solberg & Ortiz, 2022). In many cases, however, they are limited to scheduled speeches that were filed in advance, with no guarantee that the transcript reflects where a speaker went off script, nor of interactions outside of that script.

Research access to speech archives

The current work is carried out in the context of two large-scale projects that investigate speech archives: *SweTerror* (Edlund et al., 2021) and *Increasing the availability of the audiovisual collections of the Swedish National Library*. Both projects address the gaps in the records, aiming for an improved representation of the speech archives.

The focus on parliamentary data is partially due to the goals of the former project, partially because access to other sources of archival speech such as broadcast speech is limited by obstacles that have proven difficult to reconcile.

In addition, the parliamentary audio is accompanied by a large amount of transcribed text; in other domains this will not be available, or available only indirectly and in part. In broadcast news, for example, we can assume that the news as broadcast on television will broadly match that as reported in print to enough of an extent that we can extract at least direct quotations.

Swedish ASR

Two systems stand out as the current state-of-the-art for freely available Swedish ASR: Whisper (Radford et al., 2022) and a wav2vec2-based system (Baevski et al., 2020) using a model pre-trained on over 10 000 hours of Swedish speech (Malmsten et al., 2022), VoxRex[1].

Whisper has received attention partly because of its accuracy, its multilingual capabilities, its ability to translate to English, and its ability to generate capitalised and punctuated output. When capitalisation and punctation are removed, however, the VoxRex model, on the other hand, is still the most accurate model available for Swedish.

Method

The extensive data available in Riksdagens Öppna Data points to two key questions:

Intersection of ASR Systems. Can the intersection of outputs from two ASR systems provide a sufficient basis for developing a bespoke ASR system tailored to our specific needs?

Alignment with print news. Is the potential improvement from integrating print news with the parliamentary recordings substantial enough to justify the effort involved?

In addition to these technological questions, we consider the question of how much effort will be involved in obtaining human transcriptions for a human-in-the-loop approach.

[1] https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish

Data acquisition

Our data is drawn from the proceedings of the Swedish parliament (Riksdag) over an almost 10-year period, from September 2012 to January 2022. The public API was queried for items containing video, and the API results for each video was stored. The API response contains metadata, such as the date of the session, the URI of the video, and, in many cases, broadly timed information about individual speeches within the session, containing the name of the speaker, their political party, and a transcript of the speech.

A further decade of material is available through the API, recorded under similar conditions, which we have also downloaded but have not yet processed. Recordings are available for an earlier 50 years of debates, albeit with decreasing amounts of accompanying text. This material, which is not available through the API, is currently being collected with the help of Riksdagen and the Swedish National Library and prepared for use in our continued work.

The downloaded data produced by the Swedish parliament in the last decade alone amount to 5 925 hours of video recordings: manually transcribing this is infeasible. (Partial) transcripts are sometimes available, but of the 9 372 video files in our data, 2 864 came without any transcription at all.

Automatic transcription

The videos were automatically transcribed using both Whisper and VoxRex. Whisper, for all its merits, has drawbacks, likely due to the use of YouTube subtitles in training. In one sample file, which began with 5 minutes of silence, it repeatedly output the phrase "Tack till mina supporters via www.patreon.com" ("Thanks to my supporters via www.patreon.com"); and while Whisper was, in at least one case, able to override the language setting for a video in English, in at least one other case, it translated the English audio to Swedish—an impressive feat, as it was only directly trained to translate *to* English.

The texts were aligned using scripts based on those used in the creation of Librispeech (Panayotov et al., 2015), adapted to operate on the Riksdag API. This will form part of a pipeline for continuous processing of new videos as they are made available through the API, and the source code[2] is available under the Apache Public License.

Two separate sets of alignments were computed: one based on the output of VoxRex and the official transcripts, one based on the intersection of both ASR systems. We also realigned the output of VoxRex and the official transcripts, by first passing the transcripts through a text normalisation system: we ran a new ASR pass employing biased 3-gram language models derived from the normalised text and aligned it to the normalised text.

For word-level timings, the output of VoxRex was used in both cases. Word-level timings are important both for further processing the data, and for the ability to search the data itself. We want researchers to have access not merely to the fact that a word was spoken, but also to how it was spoken, which necessitates access to the actual speech. Timings from Whisper were disregarded, as the prediction model Whisper uses to estimate timings results in unpredictable granularity (some files have timings with sub-second timings, others whole seconds only), and utterance (sentence) level timings only.

Test and validation sets

For testing and validation, a set of speakers were selected at random, balanced for gender: 8 men and women

[2] https://github.com/jimregan/sync-asr

for each set. Two continuous segments of speech lasting no less than two minutes was selected for each speaker.

The segments were professionally transcribed independently by three transcribers. The first transcriber prepared two sets of transcriptions, for both two-minute segments, while the other transcribers prepared only the first set of segments.

In addition to the text transcription, these subsets are additionally annotated with markings to represent other acoustic events—such as lip smacks, breaths, and coughs—with false starts and other partially articulated words transcribed to the extent possible and marked. No express guidelines were given regarding numbers: all three annotators elected to spell out some, while using digits for others.

As part of the transcription process, a record was kept of the amount of time spent on each annotation. The times are summarised in Table 1; the average time per minute of recorded audio, taking transcription and quality assurance into account, is 9.27 minutes—roughly 10 times real time.

The test and validation sets were not designed for ASR as it is commonly understood and have phonetically-motivated transcriptions for some high frequency Swedish words that better reflect their quickly spoken pronunciations. Word Error Rate (WER) calculated on the outputs of conventional ASR systems—all of those mentioned in this work—can be expected to incur a penalty of up to 3.62, but as this can be expected to affect all systems equally, we have not attempted to adjust for it, although the vast majority of adjustments required consists of removing high-granularity information and could reliably be done automatically.

Table 1: Times spent by annotator on transcription and quality control of approx. 32 minutes of audio (64 in the case of A1). All times given in minutes.

ID	Transcription	Quality assurance
A1	532	95
A2	202	42
A3	264	52

Text normalisation

Text normalisation was performed using NVIDIA's NeMo text processing library (Zhang et al., 2021), to which we contributed support for Swedish. The system is based on (Ebden & Sproat, 2014).

Training

Hyperparameters were kept across all training instances as our concern was the effect of data selection on model quality.

Fine-tuning was performed using 8 NVIDIA GeForce RTX 3090 GPUs, and took approximately 5.8 hours to reach 12 000 steps. The models were allowed to continue further, but none improved WER beyond that point.

The base model used was the KB-wav2vec model described in Malmsten et al., (2022)—the same base model as VoxRex—which the authors kindly shared. This model is pretrained on 10 000 hours of Swedish, selected to represent a broad range of accents.

The base model was fine-tuned by adding a layer on top of the model, trained to classify into the characters of the Swedish alphabet along with two punctuation characters necessary for orthographic conventions, and a space character. The classifier was trained using Connectionist Temporal Classification (CTC) loss (Graves et al., 2006). Fine tuning was performed using the `fairseq` toolkit.

All resulting models have been made available via the Huggingface hub, where full training metrics can be viewed. The configuration file for each model is included in the repository.

Results

The results of our experiments are summarised in Table 2, along with results from state-of-the-art systems, for comparison.

VoxRex outperforms our models by quite a margin—this is unsurprising, as theirs was fine-tuned on a larger dataset that was constructed to be phonetically diverse, with a range of dialects.

The most surprising result is that doubling the size of the training data for the system using the intersection of two ASR systems did not have any effect. This may be explained by the intersection being the lowest common denominator of two systems. We have 1 186 hours in total of similar data: the increase in the scale of the data may be enough by itself to boost the performance of the system.

The relative increase gained by matching ASR output with the official transcript suggests that the intersection of the ASR systems included common errors. The VoxRex model already produces high-quality output, and the mismatches are quite often due to divergences from the official transcript.

The relatively small improvement from adding a biased language model and normalisation is not disheartening, as the text normalisation system we used is intended for more generic use, while parliamentary texts contain some very specific items, such as legal code references, that require extra support. It is worth considering that the size of the training data for this model was slightly less than half that of the others; the slight improvement did not seem worth it by itself to justify processing more data, but moving to a more flexible system that can produce more than a single possible normalisation does seem worthwhile.

Table 2: WER results

Model	WER
Whisper small	25.47
Whisper large v2	14.38
VoxRex	12.87
100h, whisper + wav2vec[3]	16.38
200h, whisper + wav2vec[4]	16.38
100h, wav2vec/transcripts, no LM[5]	14.44
49h, wav2vec/transcripts, LM, norm[6]	13.98

Discussion

Because of the requirement of contiguous segments, the test and validation sets overwhelmingly contain audio that commences at the start of a speech: they typically begin with a phrase like “tack herr/fru talman” (“thanks mister/madam speaker”). The words “herr” and “fru” are relatively uncommon in modern Swedish, and are typically misrecognised by both VoxRex and Whisper: VoxRex, because of their similarity in pronunciation to other, more common words; Whisper, because it typically omits them: many of the speeches appear with subtitles on the Riksdag Youtube channel, and follow transcription conventions, which include starting all speeches with “Talman!”, no matter what was actually spoken.

A factor in favour of our 49-hour model is the use of text normalisation: although only one possible representation was generated, it was risk-free, as the string of digits would not otherwise have matched the output of VoxRex. We have extended our work on normalisation to the audio-based normalisation described by Bakhturina et al., (2022), which allows multiple possible outputs to be generated before being filtered by the acoustic evidence, but it was not completed in time to be used in this work.

The output of this audio-based normalisation can also be used as a

[3] https://huggingface.co/jimregan/wav2vec2-swedish-riksdag-100h

[4] https://huggingface.co/jimregan/wav2vec2-swedish-riksdag-200h

[5] https://huggingface.co/jimregan/wav2vec2-swedish-riksdag-100h-transcripts-nolm

[6] https://huggingface.co/jimregan/wav2vec2-swedish-riksdag-49h-transcripts-lm

means of automatically creating a relatively reliable corpus, for use with more current neural network-based methods of text normalisation (Sproat & Jaitly, 2017), as there currently does not exist any suitable corpus for Swedish (Tånnander & Edlund, 2022).

Conclusions

We have examined the creation of custom ASR systems for transcribing archival data to support speech-centric research in various fields. Due to its availability, as well as the availability of partial transcripts, we concentrated on parliamentary data, and assessed the viability of using the intersection of outputs from two ASR systems as training data. While the current attempt does not surpass state-of-the-art performance, it comes close, so we view these results as promising although further refinement is required.

Acknowledgements

The work presented is funded by Riksbankens Jubileumsfond (In19-0144:1_RJ) and the Swedish Research Council (2020-05052_VR). The results will be made more widely available through the Swedish research infrastructure Språkbanken (2023-00161_VR).

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS 2020*, Vancouver.

Bakhturina, E., Zhang, Y., & Ginsburg, B. (2022). Shallow fusion of weighted finite-state transducer and language model for text normalization. *Interspeech 2022*, 491–495.

Ebden, P., & Sproat, R. (2014). The Kestrel TTS text normalization system. *Natural Language Engineering*, *21*(3), 333–353.

Edlund, J., Brodén, D., Fridlund, M., Lindhé, C., Olsson, L.-J., Ängsal, M. P., & Öhberg, P. (2021). A multimodal digital humanities study of terrorism in Swedish politics. In *Intelligent Systems and Applications (IntelliSys 2021)* (Vol. 295, pp. 435–449). Springer, Cham.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *ICML 2006*, 369–376.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., & Guðnason, J. (2017). Building an ASR corpus using Althingi's parliamentary speeches. *Interspeech 2017*, 2163–2167.

Malmsten, M., Haffenden, C., & Börjeson, L. (2022, June). Hearing voices at the National Library: A speech corpus and acoustic model for the Swedish language. *Fonetik 2022*, Stockholm.

Mansikkaniemi, A., Smit, P., & Kurimo, M. (2017). Automatic construction of the Finnish parliament speech corpus. *Interspeech 2017*, 3762–3766.

Nouza, J., Červa, P., & Žďánský, J. (2022). Lexicon-based vs. Lexicon-free ASR for Norwegian parliament speech transcription. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech, and dialogue* (pp. 401–409). Springer.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *ICASSP 2015*, 5206–5210.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*.

Solberg, P. E., & Ortiz, P. (2022). The Norwegian parliamentary speech corpus. *LREC 2022*, 1003–1008.

Sproat, R., & Jaitly, N. (2017). *RNN approaches to text normalization: A challenge*. arXiv.

Tånnander, C., & Edlund, J. (2022, November). Towards a Swedish test set for speech-oriented text normalisation. *SLTC 2022*, Stockholm.

Zhang, Y., Bakhturina, E., & Ginsburg, B. (2021). NeMo (inverse) text normalization: From development to production. *Interspeech 2021*, 4857–4859.