Nordwind und Sonne, wer von ihnen ... Nordwind und Sonne, wer von ihnen ...


References

See QR-Code or paper for the link.

Download our Software!

WikiText is extremely hard to parse. Therefore we use MediaWiki’s HTML output and strip tables, boxes and other stuff...that is not (or unpredictably) read out in the spoken version.

Our tool uses MaryTTS [1] for sentence segmentation and tokenization, and for pronunciation normalization (adding some rules, in particular for formulæ and years). We also add the textual header that is read out before the actual content of the article.

WikiText is extremely hard to parse. Therefore we use MediaWiki’s HTML output and strip tables, boxes and other stuff...that is not (or unpredictably) read out in the spoken version.

Audio Alignment

We use a variation of SailAlign [2] as implemented in Sphinx 4 [3] with some modifications. SailAlign treats alignment as repeated and successively more restricted speech recognition using N-gram models derived from the text that is expected to be spoken. Correct stretches are used as landmarks to restrict re-recognition to tighter models.

SailAlign is very robust to text/audio mismatches which is a strong requirement given the modest data quality:

• text/audio mismatches (including text not read, audio not in text)
• mis-matching revision IDs (even leading to wrong articles)
• mis-normalization ("*" → "geboren" or "Sternchen"?)

Favor quality over quantity in the alignments: rather leave out timings for dubious cases (e.g. mis-normalized tokens) than providing a full alignment that is partially faulty.

Bootstrapping Acoustic Models

Alignment requires acoustic and pronunciation models.

→ Sphinx 5 PFM acoustic models with good results for English. For German, we used limited-quality acoustic models [4] and iteratively built new models trained on the aligned data found so far.

→ align more and more data as model quality improves.

We cross-checked the quality of the resulting models on the Kiel Corpus of Read Speech. We found that alignment quality continues to increase. We expect this effect to carry over to ASR.

Download our Software!

Our modular software consists of C and Java code bound together with Python and Shell and works on (at least) Linux and Windows. See QR-Code or paper for the link.

Utilization of the Corpus:

• speech recognition training
• improve accessibility / spoken article navigation
• top contributors speak >30 hours → train synthesis voices
• analyze prosody of reading (large amounts of diversely read text)
• analyze prosody of information structure (accessible through links, research on semantic Wikipedia, Dbpedia, ...)

Download the Corpus!

For each article, the corpus contains:

• audio file(s)
• original WikiText
• HTML generated by MediaWiki
• cleaned and normalized text
• alignment between text and audio
• meta-information (who, when, what)

Languages: German and English

We are actively working on more data and more languages.

→ find it at:

http://nats-www.informatik.uni-hamburg.de/SWC/

Our work, our corpus and this poster is CC-by-SA. Thanks to the Wikipedia contributors, as well as to Daimler-Benz Foundation for funding.

The Spoken Wikipedia project unites volunteer readers of Wikipedia articles. Hundreds of spoken articles in multiple languages are available to users who are – for one reason or another – unable or unwilling to consume the written version of the article.

We turn this speech resource into a time-aligned corpus, making it accessible for research and to foster new ways of interacting with the material.