DuStt – a Speech-to-Text Engine for Dutch ⋆,⋆⋆

Willem Röpke, Roxana Rădulescu, Kyriakos Efthymiadis, and Ann Nowé

Artificial Intelligence Research Group, Vrije Universiteit Brussel, Belgium
{Willem.Rokpe,Roxana.Radulescu,Kyriakos.Efthymiadis,Ann.Nowe}@vub.be

Abstract. We develop and demonstrate a speech-to-text engine for Dutch, starting from the open-source project DeepSpeech and using the Corpus Gesproken Nederlands. The DuStt engine provides models targeted towards Dutch, Flemish or speakers from both Belgium and The Netherlands. Users can upload or record their own input as well as load pre-recorded samples and obtain a transcription on the spot. The demonstration is video available at: https://youtu.be/DtTK0uo5W7s.

Keywords: Speech-to-Text · Corpus Gesproken Nederlands · DeepSpeech

1 The DuStt Engine

Speech-to-Text (STT) engines recognize and transcribe spoken language into text. This transcription can be used to complete a multitude of tasks, such as parsing voice commands or providing automatic subtitles. The performance of STT models has been steadily increasing in the last decade, due to advances in deep neural networks and newly developed architectures.

Currently developed methods are usually targeted towards English and all the state-of-the-art results are also bench-marked on a wide range of English datasets. The DuStt Engine is an initiative to attract more attention towards datasets and models curated for the Dutch language.

Architecture In order to train our models, we selected the architecture provided by the DeepSpeech[1][1] open-source project, developed by Mozilla, as a starting point. We then adjusted the network and its parameters (e.g., hidden layer size, learning rate, batch sizes) for our dataset.

Corpus The dataset we explored for building our STT Dutch engine is the Corpus Gesproken Nederlands[2] [2], with a total of 900 hours of spoken Dutch, amounting to a vocabulary of over 9 million words. In total, the speech data contains a split of 76, 23% and 23,77% of Dutch and Flemish audio files respectively. Because the length of the audio files was too long, the first pre-processing step was to split the files into smaller chunks, each averaging around 6 seconds. A

⋆ Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

⋆⋆ This work was carried out by the first author during his bachelor project [4].

[1] https://github.com/mozilla/DeepSpeech

[2] https://ivdnt.org/downloads/tstc-corpus-gesproken-nederlands

CEUR-WS.org/Vol-2491/demo57.pdf

second issue we encountered concerned noisy or wrong transcriptions and overlapping timestamps for the provided annotations. In order to handle these issues, we have eliminated the components that incorporate face-to-face speech, a noisy setting even for humans. Moreover, we have also decided to eliminate the files that had overlapping timestamps for the transcriptions, as it was impossible to tell how to properly assign the annotations without a laborious manual process.

Graphical Interface The DuStt Engine provides an interface (Figure 1) that allows users to load pre-trained neural models targeted either for Dutch, Flemish or for both type of speakers. Furthermore, users can upload or record on the spot an audio sample or load an existing one and obtain a transcription for it.

Performance The performance obtained for the trained models averages around the 23-30% range for WER (word error rate) and the 14-20% range for CER (character error rate).

We plan to further extend the DuStt engine with models trained using different frameworks (e.g., PyTorch-Kaldi [3]) and also improve the quality of the data, to obtain higher-performing models for Dutch.

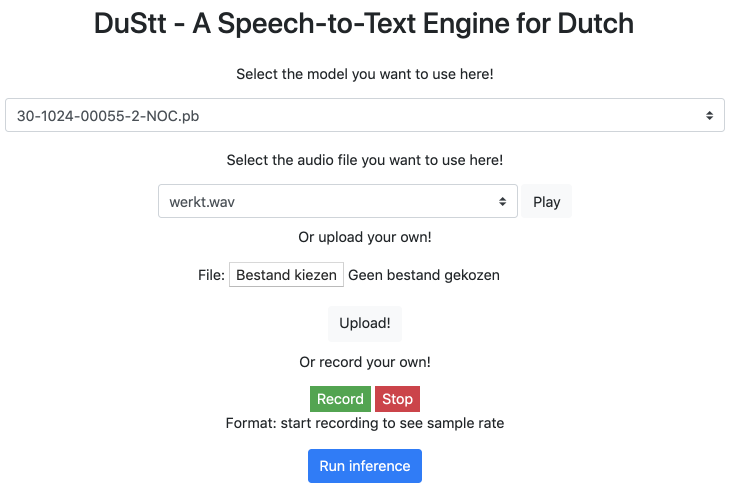

Fig. 1. The graphical interface of the DuStt Engine

References

1. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
2. Oostdijk, N.: The Spoken Dutch Corpus. Overview and First Evaluation. In: LREC (2000)
3. Ravanelli, M., Parcollet, T., Bengio, Y.: The pytorch-kaldi speech recognition toolkit. In: In Proc. of ICASSP (2019)
4. Röpke, W.: Building a Speech-to-Text Engine for Dutch. Bachelor thesis, Vrije Universiteit Brussel (2019), `https://ai.vub.ac.be/files/Ropke_Bachelor_thesis_1819.pdf`