PSSOH 2021, DOI: https://doi.org/10.5281/zenodo.6013449

Free Software in Speech Technology

Branislav Gerazov

Faculty of Electrical Engineering and Information Technologies
Ss Cyril and Methodius University in Skopje, Macedonia
gerazov@feit.ukim.edu.mk

Abstract

Speech Technologies are becoming increasingly important with the increased penetration of smart devices in our every day lives. They also are indispensable for allowing the digital and social inclusion of people with disabilities. Free software has always played an important role in speech technology research and real-world deployment in general. This is even more true today with the development of deep learning based systems. With this speech technology is continuing to catalyse the advancement and proliferation of free software and with that open and reproducible science.

Key words [speech technology, free software, open source software, speech synthesis, speech recognition, speaker recognition]

1 Speech Technology

Speech technology relates to the technologies designed to duplicate and respond to the human voice [1]. It includes a large range of different fields including: speech synthesis, speech recognition, and speaker recognition and verification. Speech technology allows:

1. speech based communication with electronic devices
2. conversion of speech to other forms of information and *vice versa*, such as: text, images, actions / movements, and neural activity.

Speech technology is important for the general public. It allows the creation and deployment of speech enabled devices and thus improves our daily interaction with technology. Some examples include handling a mobile phone while driving and handling smart devices in the household. These technologies are also a core element of the increasingly ubiquitous virtual assistants. Speech synthesis also enables the conversion of textual content into speech, facilitating easier consumption: audiobooks, news / web portals, and e-mail.

Speech technology is also important for people with disabilities. Speech synthesis empowers screen readers that allow the blind and visually impaired to use computers and smart devices. Assistive augmentative and alternative communication (AAC) devices use speech synthesis to give a voice to the deaf and dumb, and the speech impaired. Sign language devices can levy speech recognition to convert speech to sign language for the deaf.

The field of speech technology has always been well-supported by free software. Various speech toolkits have been released by research institutes and advanced by the community. Some of these offer commercial level performance. The paper lists a number of available free software packages for two speech technologies: text-to-speech (TTS) synthesis and automatic speech recognition (ASR). The list is by neither meant to be exhaustive or representative. It is mainly to be used for further exploration of the available free software packages in each of the subfields.

2 Free Software for Speech Synthesis

There are several paradigms in text-to-speech synthesis:

- articulatory synthesis – based on modelling the anatomy and physiology of the vocal tract,
- formant synthesis – based on the filter-source model for shaping the speech signal,
- concatenative synthesis – based on the concatenation of short audio recordings of natural speech,
- parametric synthesis – based on statistical and machine learning models of the speech signal dynamics,
- deep learning based synthesis – based on advanced neural network architectures for speech synthesis.

Four of these five paradigms, except formant synthesis, are covered with free software packages.

VocalTractLab [2] is an articulatory speech synthesizer. It also can be used as a tool to visualize and explore the mechanisms of speech production with regard to articulation, acoustics, and control. The project also has a development version on GitHub [3].

The early concatenative synthesis systems used short speech segments based on diphones, i.e. inter-phone transitions from the stable state of the first to the stable state of the succeeding phone. One of the free software synthesisers supporting diphone synthesis was MBROLA [4]. The Festival Speech Synthesis System supports both diphone and the more advanced unit selection synthesis that is based on larger segments, and was used in most commercial TTS systems before the wide adoption of deep learning models [5].

The Hidden Markov Model/Deep Neural Network (HMM/DNN) based Speech Synthesis System (HTS) [6] is the primary free software used for parametric speech syngvthesis, both in research, as well as commercially. RH Voice [7] that is built on top of HTS, is a free software product that is used as a screen-reader on Windows, Linux and Android by many users with disabilities.

Deep learning based synthesis is a very hot topic in speech research. In fact, most recent papers either include official implementations available as free software, or have been implemented by the community. One of the largest offerings of a variety of latest TTS models is Coqui-AI TTS [8]. A successor of Mozilla TTS, it offers over 20 pretrained language models, and has a vibrant and very active developer community. Another big package is the ESPnet: end-to-end speech processing toolkit [9]. It offers implementations of state-of-the-art synthesis models as well as pretrained models ready for use. Examples of single model implementations are: NVIDIA's Tacotron 2 And WaveGlow v1.10 For PyTorch [10], DeepVoice 3 [11], and Transformer-TTS [12].

3 Free Software for Speech Recognition

As in synthesis, there are several paradigms in automatic speech recognition (ASR):

- template based – the original ASR approach still viable for simple small-vocabulary single-word small-footprint systems. Based on comparison of input words to stored word templates,
- parametric - based on statistical and machine learning models, and
- deep learning - based on advanced neural network architectures.

All paradigms have free software support.

Template based ASR is straight-forward to implement using existing Python libraries, e.g. DTW: Dynamic Time Warping Python Module [13]. Some example systems available online are "DTW* applied to isolate word speech recognition" [14], and "Simple word recognition using dynamic time warping" [15].

Parametric ASR models can be divided in models based on Hidden Markov Models (HMMs) and those based on Weighted Finite State Transducers (WFSTs). The HMM Toolkit (HTK) [16] is one of the most used packages for ASR. Although it was primarily used for speech recognition, it was also applied to HMM based speech synthesis (HTS), character recognition and DNA sequencing. CMUSphinx [17] and PocketSphinx [18] were both based on HTK and were used in real-world applications and commercially. The Julius: Open-Source Large Vocabulary Continuous Speech Recognition Engine [19] is also a free software that was based on HTK and used for real-world ASR.

The Kaldi Speech Recognition Toolkit [20] is an advanced WFST based platform. It comprises high performing models and is actively developed by a big community. Kaldi is deployment ready for use in real-world applications. It has been used by other packages, such as the VOSK Speech Recognition Toolkit [21].

Deep Learning ASR, not unlike TTS, is a hot topic in speech research. Both Coqui-AI and ESPnet offer ASR support. Coqui Speech-to-text (STT) [22] is a fast, multi-platform, deep-learning toolkit. It boasts over 80 pretrained models for over 50 of the world languages, as well as state-of-the-art performance. Like it's TTS counterpart, it is a successor of Mozilla's Project DeepSpeech [23]. An up-and-coming framework that offers ASR is the SpeechBrain all-in-one speech toolkit[24].

4 Free Software for Speaker Recognition

Speaker recognition has also been offered in free software packages. Amongst them we can mention SPEAR: A Speaker Recognition Toolkit based on Bob [25], as well as the ALIZE Speaker Recognition Platform offering commercial grade speaker recognition [26]. SpeechBrain also includes speaker recognition and it's worth keeping an eye on in the future [24].

5 Conclusion

Free software remains to play an important part in speech technology research. Moreover, free software frameworks are becoming increasingly important for real-world deployment of speech technology, even in commercial speech systems. With this speech technology is one of the fields that is continuing to catalyse the advancement and proliferation of free software and with that open and reproducible science.

References

[1] Wikipedia: Speech technology https://en.wikipedia.org/wiki/Speech_technology
[2] VocalTractLab https://www.vocaltractlab.de/
[3] VocalTractLab GitHub https://github.com/TUD-STKS/VocalTractLab-dev
[4] MBROLA https://github.com/numediart/MBROLA
[5] Festival Speech Synthesis System https://www.cstr.ed.ac.uk/projects/festival/
[6] The Hidden Markov Model/Deep Neural Network (HMM/DNN) based Speech Synthesis System (HTS) http://hts.sp.nitech.ac.jp/
[7] RH Voice https://github.com/RHVoice/RHVoice
[8] Coqui-AI TTS https://github.com/coqui-ai/TTS
[9] ESPnet: end-to-end speech processing toolkit https://github.com/espnet/espnet
[10] NVIDIA's Tacotron 2 And WaveGlow v1.10 For PyTorch https://github.com/NVIDIA/

DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/Tacotron2
[11] DeepVoice 3 https://github.com/r9y9/deepvoice3_pytorch
[12] Transformer-TTS https://github.com/as-ideas/TransformerTTS
[13] DTW: Dynamic Time Warping Python Module https://github.com/pierre-rouanet/dtw
[14] DTW applied to isolate word speech recognition https://github.com/aishoot/DTWSpeech
[15] Simple word recognition using dynamic time warping https://github.com/crawles/dtw/blob/master/Speech_Recognition_DTW.ipynb
[16] The HMM Toolkit (HTK) https://htk.eng.cam.ac.uk/docs/cuhtk.shtml
[17] CMUSphinx https://cmusphinx.github.io/
[18] PocketSphinx http://www.speech.cs.cmu.edu/pocketsphinx/
[19] Julius: Open-Source Large Vocabulary Continuous Speech Recognition Engine https://github.com/julius-speech/julius
[20] The Kaldi Speech Recognition Toolkit https://github.com/kaldi-asr/kaldi
[21] VOSK Speech Recognition Toolkit https://github.com/alphacep/vosk
[22] Coqui Speech-to-text (STT) https://github.com/coqui-ai/STT
[23] Mozilla's Project DeepSpeech https://github.com/mozilla/DeepSpeech
[24] SpeechBrain all-in-one speech toolkit https://github.com/speechbrain/speechbrain
[25] SPEAR: A Speaker Recognition Toolkit based on Bob https://gitlab.idiap.ch/bob/bob.bio.spear
[26] ALIZE Speaker Recognition Platform https://github.com/ALIZE-Speaker-Recognition