Polish Read Speech Corpus for Speech Tools and Services

Koržinek Danijel
Polish-Japanese Academy of Information Technology, Warsaw, Poland
danijel@pja.edu.pl

Marasek Krzysztof
Polish-Japanese Academy of Information Technology, Warsaw, Poland
kmarasek@pja.edu.pl

Brocki Łukasz
Polish-Japanese Academy of Information Technology, Warsaw, Poland
lucas@pja.edu.pl

Abstract

This paper describes the speech processing activities conducted at the Polish consortium of the CLARIN project. The purpose of this segment of the project was to develop specific tools that would allow for automatic and semi-automatic processing of large quantities of acoustic speech data. The tools include the following: grapheme-to-phoneme conversion, speech-to-text alignment, voice activity detection, speaker diarization, keyword spotting and automatic speech transliteration. Furthermore, in order to develop these tools, a large high-quality studio speech corpus was recorded and released under an open license, to encourage development in the area of Polish speech research. All the tools and resources were released on the the Polish CLARIN website. This paper discusses the current limitations and future plans of the project.

1 Introduction

Much of the data used in Humanities and Social Sciences (HSS) research is stored in the form of audio recordings. Examples of this include radio and television programmes, interviews, public speeches (e.g. parliament, public events), lectures, movies, read literary works and other recordings of speech. This data contains valuable information from many aspects of HSS research. This encompasses both the linguistic (with the emphasis on vocabulary and pronunciation) and sociological (emphasis on speaker) points of view. During our project, we have met many scientists that have shown interest in processing either already available data and corpora, or would like to process recordings (e.g. interviews) they intended to make in the future.

The main issue with processing acoustic data is that it is more expensive and time consuming than, for example, traditional, textual data. It demands both the know-how and lots of effort to achieve comparable results. That is why it is often overlooked by researchers who either do not have the time or the funding to deal with such things. Our primary goal was to create free and accessible solutions for people from the HSS community.

Similar efforts in other consortia already exist, like WebMAUS (Kisler et al., 2016) speech segmentation services in LMU, AVATech (Lenkiewicz et al., 2012) by Max Planck Institute and Fraunhofer Institute which provide video and audio processing services including speech segmentation, VAD and speaker diarization, and TTNWW (CLARIN-NL, 2013) which includes speech transcription services for Dutch. It is worth noting that many of them (although not all) are language dependent, requiring a re-implementation of these services in individual countries.

This paper will first describe the tools and corpora created during the project. Next, it will describe a few of the existing and planned applications of the tools and services. Finally, it will describe the plans of development for the upcoming years.

2 Speech Tools

One of the earliest decision during the project was to release all the tools in the form of web services, rather than downloadable applications. There are many advantages to this: ease of use (no installation required), better support, stable environment and performance. However, a few disadvantages as well: more effort required from the consortium, increased response time if many people use the platform, issues with releasing sensitive data. Most of these have been addressed individually and by releasing the source code of the tools for ambitious individuals.

The main website located at `http://mowa.clarin-pl.eu` was divided into three sections: speech corpora downloads, grapheme-to-phoneme (G2P) conversion and the rest of the speech processing tools. The reason for removing the G2P from the rest of the tools is because it uses a different set of modalities (i.e. text-to-text) from the rest of the tools (audio + text [opt.] to text).

2.1 Grapheme-to-phoneme conversion

This tool allows converting any text written in the ortographic (i.e. written) into its phonetic (i.e. spoken) form. It is one of the primary steps in any process that involves speech data but may also serve as a tool outside of the acoustic speech processing context.

The tool is created using a rule-based system. It accepts any form of text, although it does not perform text normalization (it does not expand numbers, dates or abbreviations automatically). The system is completely rule based and contains a list of exceptions for names, foreign and other atypical words. A statistical system, based on the Sequitur (Bisani and Ney, 2008) tool, is also available but due to available data, it does not outperform the rule based system in any way. The tool can generate both word lists (with multiple pronunciations) and a canonical transcription of the text.

2.2 Speech-to-text alignment

As one of the most useful tools available, it allows aligning a sequence of words to the provided audio recording of speech. This can be understood simply as automatically generating a set of time-codes, when both the audio and its transcription are known. It is a very useful tool because it can be used to easily look up specific events in large sets of recordings. It also makes it possible to perform simple statistics related to the duration of individual speech events.

The tool was created, based on the SailAlign (Katsamanis et al., 2011) concept, in order to work efficiently with long audio files. The engine is constructed around the Kaldi toolkit (Povey et al., 2011), just like most of the tools on the site, but the main work-flow is managed using a set of libraries written in Java.

2.3 Voice activity detection

Voice activity detection (VAD) is frequently found as a pre-processing step of many speech processing tools. Its purpose is to isolate portions of audio that contain speech from those that contain other types of acoustic events, like silence, noise or music. Apart from the aforementioned use as a pre-processing step, it can also be useful as an indexing tool for large quantities of audio. This tool is completely language and domain independent, although it may fail with very noisy data.

This tool was constructed using an Artificial Neural Network based classifier that performs VAD in an online manner. The non-speech data is further analyzed using an SVM classifier to try and classify types of noise.

The VAD component was used extensively during previous project and was already known to perform reasonably well for real-world data. A simple experiment confirmed a fairly high level of recall ($\sim$99%) with a not so good precision ($\sim$58%), which was a conscious decision (preferring no to lose any speech, while sometimes accepting non-speech falsely). The classification component did much worse, depending on the class, and needs more work if it is to be used in the future.

2.4 Speaker diarization

This tool is used to segment a large audio file into portions spoken by individual speakers. There are several types of speaker related segmentation strategies that can be performed: speaker change detection recognizes only the segments where different speakers are talking, speaker diarization additionally annotates which segments belong to the same speaker and speaker identification recognizes exactly who the speaker talking in each segment is (e.g. by name). Our tool only does the second algorithm.

It is mostly useful for adaptation of various tools and models to individual speakers but some researchers have mentioned that they would like to use it for other types of analyses that require speaker segmentation. Our tool is based around the LIUM (Meignier and Merlin, 2010) toolkit and just like the previous one, it is completely language independent. Other toolkits were also tested during the project (for example SHoUT(Huijbregts, 2006)) but LIUM seemed to perform best on real-world data. Some toolkits were not tested and it is possible something else may be used in the future.

2.5 Keyword spotting

Often times, an accurate transcription of audio material is not necessary because we are only interested in individual words appearing in the text. Keyword spotting (KWS) is a process that takes an audio file with a list of words that need to be found and generates a list of keyword occurrences with their location in the audio file.

Our system was based around the Kaldi toolkit (Povey et al., 2011) but it was also expanded to support an open vocabulary scenario. Given the limited language model vocabulary size, it would be impossible to predict all the words that people may be looking for. Therefore, our system uses a combination of words and syllables, so when a word out of vocabulary needs to be found, its syllable representation is used instead.

A small corpus was prepared to test this component and the overall precision was very high ($>\sim$95%) with the recall being reasonably high for known words ($\sim$82%) and low for words that were OOV ($\sim$20%). It seems that the syllable model worked well sometimes but still needs improvement to deal with OOVs.

2.6 Automatic speech transliteration

This tool uses an Automatic Speech Recognition (ASR) system (based on the Kaldi toolkit (Povey et al., 2011)) to generate a probable orthographic transliteration of audio recording of Polish speech. Initially, this tool was not planned for inclusion in the project but due to overwhelming interest, it was nonetheless added in order to gauge its usefulness in the future. The current system uses our Euronews model for recognizing broadcast news (Marasek et al., 2014) and in order for it to be useful for other types of recordings, it has to be adapted to the proper domain.

3 Speech Corpus

In order to produce most of the tools mentioned in the previous section, a large set of good quality recordings is required. This is usually expensive to produce and even if such data is available for purchase from third-parties, it is usually very expensive and unobtainable by most researchers. Prior to our work, there was no free, high quality, large-vocabulary audio corpus of Polish speech. Our goal was to create such a corpus and release it on an open license, both for commercial and non-commercial use.

The corpus was recorded in a studio environment using two microphones: a high-quality studio microphone and a typical consumer audio headset. The corpus consists of 317 speakers recorded in 554 sessions, where each session consists of 20 read sentences and 10 phonetically rich words. The size of the audio portion of the corpus amounts to around 56 hours, with transcriptions containing 356674 words from a vocabulary of size 46361. In addition to the studio corpus, a smaller corpus of telephony quality was also recorded. It contains 114 sessions, amounting to around 13 hours of recorded speech.

Both the studio and telephone quality corpora were released in two forms. The first one is the EMU database (Cassidy and Harrington, 2001), which allows for easy lookup of data and even some statistics thanks to the integration with the R platform. Unfortunately, the current version of the system relies on downloading the rather sizable corpus locally onto the computer. The second form is a baseline speech recognition system created using the Kaldi toolkit. It is thoroughly documented and allows researchers to download and build their own speech processing tools, if they are so inclined.

4 Applications

A couple of projects have already utilized our tools and resources for their own uses. Our speech alignemnt tool was used by a consortium partner in order to further annotate the corpora on their Spokes platform (Pezik, 2015). The studio speech corpus was used in a paper by a Czech research team (Nouza et al., 2015). We have also managed to cooperate with a team from the Institute of Applied Linguistics at the Warsaw University on their project titled “Respeaking - the process, competences and quality” (project code NCN - OPUS6 -2013/11/B/HS2/02762). Finally, one of the most interested groups were researchers of sociology interested in automatic transliteration of sociological interviews. We managed to receive several hours of recordings by a group of researchers from the The Cardinal Wyszyński University in Warsaw. Some preliminary results show promise but more work is needed to achieve success.

We intend to open several new areas of applications in the future. The new project will concentrate mostly around these three domains: parliamentary speeches, historical early and mid-20th century news segments and improved systems for the transliteration of sociological interviews.

5 Future plans

With the project being prolonged for the next two years, several improvements are planned. The main focus will be on creating working speech recognition solutions for the aforementioned domains. To achieve this, certain tools, like the G2P conversion including text normalization and possibly other modules, like speaker diarization and VAD, will have to be improved. The biggest improvements, however, will lie in the speech recognition engine, itself. Many experiments are planned, including various adaptation techniques, Deep Neural Network for acoustic modeling (Vu et al., 2014), Recurrent Neural Networks for language modeling (Mikolov et al., 2013) and possibly end-to-end LSTM based systems (Miao et al., 2015).

No new corpora will be recorded, although lots of data will have to be collected, in order to adapt the tools to their respective domains. It is unclear whether all of the data will be released for other researchers, due to legal concerns. Our primary intention will be to improve the services available on our website and to provide the trained models and tools for free, for others to use as they deem necessary.

References

[Bisani and Ney2008] Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

[Cassidy and Harrington2001] Steve Cassidy and Jonathan Harrington. 2001. Multi-level annotation in the emu speech database management system. *Speech Communication*, 33(1):61–77.

[CLARIN-NL2013] CLARIN-NL. 2013. Ttnww - tst tools voor het nederlands als webservices in een workflow. `https://portal.clarin.nl/node/1964`. [Online; accessed 2016-09-27].

[Huijbregts2006] Marijn Huijbregts. 2006. Shout speech recognition toolkit.

[Katsamanis et al.2011] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. 2011. Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.

[Kisler et al.2016] Thomas Kisler, Uwe Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl, and Nina Prner. 2016. Bas speech science web services - an update of current developments. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[Lenkiewicz et al.2012] Przemyslaw Lenkiewicz, Eric Auer, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpe. 2012. Avatechautomated annotation through audio and video analysis. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 209–214. European Language Resources Association.

[Marasek et al.2014] Krzysztof Marasek, Krzysztof Wołk, Danijel Koržinek, Łukasz Brocki, and Ryszard Gubrynowicz. 2014. Spoken language translation for polish. *Forum Acousticum*.

[Meignier and Merlin2010] Sylvain Meignier and Teva Merlin. 2010. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010.

[Miao et al.2015] Yajie Miao, Mohammad Gowayyed, and Florian Metze. 2015. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Nouza et al.2015] Jan Nouza, Petr Cerva, and Radek Safarik. 2015. Cross-lingual adaptation of broadcast transcription system to polish language using public data sources. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland*, pages 181–185.

[Pezik2015] Piotr Pezik. 2015. Spokes-a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, number 116 in Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press.

[Povey et al.2011] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

[Vu et al.2014] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard. 2014. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643. IEEE.