



SARMATA 2.0 Automatic Polish Language Speech Recognition System

Bartosz Ziółko^{1,2}, Tomasz Jadczyk^{1,2}, Dawid Skurczok^{1,2}, Piotr Żelasko¹, Jakub Gałka^{1,2},
Tomasz Pędzimąż^{1,2}, Ireneusz Gawlik¹, Szymon Pałka¹

¹AGH University of Science and Technology,
Department of Computer Science, Electronics and Telecommunications,
al. Mickiewicza 30, Kraków, Poland, www.dsp.agh.edu.pl

²Techmo, Kraków, Poland, techmo.pl

{bziolko, jadczyk, skurczok, pzelasko, jgalka, pedzimaz, igawlik, pszymon}@agh.edu.pl

Abstract

A speech recognition system for the Polish language is described. The presentation will focus on an adjustment of the Kaldi toolkit for Polish, our own grapheme to phoneme conversion tool and a corpus of Polish we collected. The approaches to commercial applications will also be described.

Index Terms: ASR, corpus, phonetic transcription, Polish

1. Introduction

Automatic Speech Recognition (ASR) systems are becoming increasingly more popular, even for languages with fewer native speakers (around 60 million worldwide for Polish). An example of recent successes of the industry is the Polish ASR delivered by Google, which is used in voice searching in Android applications. Our approach to ASR system migrated through the years from the use of the HTK system [1], through our own implementation of discrete wavelet transforms and k-NN classifier [2] followed by GMM and HMM classifier with standard parameterization, to the application of the Kaldi toolkit [3] for Polish. Only a handful of the world's languages, such as English, benefit from resources such as a wide selection of thousands of hours long speech corpora or representative text corpora that are necessary in speech recognition development. Our corpus consists of around 40 hours of annotated speech in different recording conditions and of different quality. It is not the largest corpus of Polish speech, but probably the largest of those that are available for licensing. During training of our ASR system other corpora are also used, such as: CORPORA [4], Globalphone [5] and Luna [6].

Another crucial element of the system is OrtFon 2.0. This is a very successful licensed program utilizing a phonetic transcription system based on existing knowledge of Polish phonetics. Its implementation is much more effective than the previous system. It also has another advantage - ease and user-friendly introduction of additional linguistic data.

2. AGH Corpus

Our choice of the methods of data collection as well as decisions on the statistical profile of the corpus were mainly dictated by the need for a large number of speakers and large amount of recording data. We focused primarily on building a large and well-annotated training corpus [7] rather than collecting a complete set of various dialects, ages or topics. We designed our

corpus to be dedicated to ASR training and tests, and therefore provided all required metadata only for those tasks. This involves annotation of the start and end time of each utterance, identifier and gender of a given speaker and identifier of the subcorpus from which a given recording originates. Corpus is composed of 4 main parts: colloquial speech, voice interface commands, TTS data and telephonic speech. All utterances are annotated in MLF files.

In total, the AGH corpus has 42 hours and 9 minutes of recordings, uttered by 391 speakers. About 55% of speakers are male and 45% are female. Men (about 22 hours) and women (about 20 hours) contribute almost equally to the corpus in terms of total recording time. The majority of speakers are 20-35 years old. All recordings are stored in mono WAVE files with 16 kHz sampling rate and 16 bit precision. Some of the annotations were automatically checked for orthographic correctness using OpenSJP Polish dictionary [8] and manually corrected. The post-processing of parts of the corpus included creation of a list of words which are foreign or phonetically ambiguous and preparation of manual transcription for them, so that OrtFon 2.0 is able to deal with these issues.

Telephonic speech is the part of our corpus that is most relevant to our current research, as we will describe in section 4. Because of a lack of good quality telephonic speech corpora for Polish, we had to gather the data ourselves. For this purpose, we designed a call recorder program that uses Voice over Internet Protocol (VoIP). In the first version of this program, the speakers called the recorder using either a regular telephone or a cell phone, and read out a list of words (or sentences), which was given to them earlier, and was also known by the call recorder. The speakers were asked to start reading a sentence each time after they hear a signal in the phone. Successive signals were played after a timeout, which was adapted to the length of phrase to be read. This allowed us to automatically annotate the recordings, but the downside was that every speaker had to conform to the tempo dictated by the device. It resulted in some recordings being cut prematurely - these were detected in tests performed with ASR, where they achieved low scores, and were removed from the corpus after confirming that they were beyond repair. Recently we developed a new call recorder to address these problems. It is integrated with an ergonomical web service, which allows users to navigate through sentences and record them in their own, comfortable tempo. We are planning to use it on wider scale in order to gather a large, representative corpus of Polish.

3. Ortographic to phonetic transcription

Grapheme to phoneme transcription is widely used in Text-To-Speech [9, 10] and ASR systems [11, 2]. There are two general methods of performing such transcription automatically: rule-based and data-driven. Rule-based methods are implemented in expert systems that are based on context decisions provided by linguists.

In cases of some very irregular languages, such as English, the pronunciation dictionaries are difficult to build. In others, including Polish, there are rules allowing programming. As simplistic as it may seem, the fact that there are thousands of rules makes the resulting conversion program difficult to set up appropriately. For example, some of the rules may overlap and correct precedence must be established. One of our aims in designing a new transcription tool was to create both a fast and flexible system while preserving an easy and intuitive interface. Many existing tools for language processing are written using some scripting language. It significantly degrades speed, efficiency and creates problems during integration with systems created in compiled languages, especially in an embedded system. Moreover, some high-security environments, such as internal police systems, prohibit the use of any scripting language.

Our tool is designed to work as a part of an ASR system. Such systems require a high speed of data processing, especially for real time applications. Most of ASR systems are created using C or C++ (e.g. HTK or Kaldi). Some of them use other languages, but only for research and initial algorithm design (e.g. CMUSphinx [12] utilizes both C and Java). Due to this fact, our design choice was to implement OrtFon 2.0 in C++.

4. Robust speech recognition

Our ASR system aims to work in a difficult environment. It is designed to recognize short phrases or single word commands in Interactive Voice Response (IVR) systems over telephone. In such scenarios total speech input is usually shorter than 10s, so any attempt to use speaker adaptation or normalization techniques faces the problem of overfitting to a small amount of data. Moreover, various telephone channels and codecs modify signals in very different ways. Sometimes the channel type or codec is known from the VoIP service provider, however it may be misleading because the information provided concerns only properties of the last section of the route. We also observed that some GSM codecs cut out high frequency Polish fricatives at the beginning of the speech signal.

Recently, we started using Kaldi [3] which shows most promising results. As input for training we use signals recorded from telephone lines, which are part of our corpus described in section 2. Additionally we modify these recordings with Vocal Tract Length Normalization (VTLN) using different normalization coefficients when utilizing DBN.

Our system supports context free grammars in form of Speech Recognition Grammar Specification (SRGS). It is a W3C standard, and is widely used in commercial systems. Decoding words with grammar allows the decoder to prune phrases that do not belong to a given formal language. Parsing is a complex process and every path in decoding lattice requires its own parser. This motivated us to introduce some optimizations.

First, we use only grammar recognizers in every path. They do not produce parsing trees, but they can discard wrong

word sequences while decoding. Additionally, every recognizer shares some internal state with other recognizers that have the same words at the beginning of the decoding path. It helps to reduce the memory overhead. Finally, in order to generate the parsing tree we use a full parser on the best word path at the end of decoding. We decided to use Earley recognizer because it accepts any kind of context free grammar. The users of a speech recognition system generally create these grammars, so we cannot assume any restriction in its form. Using a top down parser also has the advantage that in every step we can list all words which may appear as next in the decoding path.

5. Acknowledgements

The research was funded by the National Science Centre allocated on the basis of the decision DEC-2011/03/D/ST6/00914 and by National Centre for Research and Development LIDER/37/69/L-3/11/ NCBR/2012 grant. We thank Raul Cruz for help with proofreading of the paper.

6. References

- [1] B. Ziółko, S. Manandhar, R. C. Wilson, M. Ziółko, and J. Gałka, "Application of HTK to the Polish language," *Proceedings of IEEE International Conference on Audio, Language and Image Processing, Shanghai*, 2008.
- [2] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, and M. Mąsior, "Automatic speech recognition system dedicated for Polish," *Proceedings of Interspeech, Florence*, 2011.
- [3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.
- [4] S. Grochowski, "CORPORA - speech database for Polish diphones," *Proceedings of Eurospeech 1997*, 1997.
- [5] N. T. Vu, F. Kraus, and T. Schultz, "Multilingual a-stabil: a new confidence score for multilingual unsupervised training," *Proceedings of IEEE Workshop on Spoken Language Technology, SLT 2010, Berkeley*, 2010.
- [6] M. Marciniak, *Anotowany korpus dialogów telefonicznych (Eng. Annotated corpus of telephone dialogues)*. Exit, Warszawa, 2010.
- [7] P. Żelasko, B. Ziółko, T. Jadczyk, and D. Skurzok, "AGH corpus of Polish speech," *Language Resources and Evaluation*, vol. early access, 2015.
- [8] "Open source online dictionary of the Polish language," 2014, <http://sjp.pl/open>.
- [9] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," *The Third ESCA Workshop in Speech Synthesis*, 1998.
- [10] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96*, vol. 1, pp. 373 – 376, 1996.
- [11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book*. UK: Cambridge University Engineering Department, 2005.
- [12] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 1, pp. 35 – 45, January 1990.