# Application of HTK to the Polish Language

Bartosz Zi´ołko1, Suresh Manandhar1, Richard C. Wilson1, Mariusz Zi´ołko2, Jakub Gałka2
*1Department of Computer Science, University of York*
*Heslington, YO10 5DD, York, UK*
*2Department of Electronics, AGH University of Science and Technology*
*al. Mickiewicza 30, 30-059 Krak´ow, Poland*
*{bziolko,suresh,wilson}@cs.york.ac.uk, {ziolko,jgalka}@agh.edu.pl*

## Abstract

A speech recognition system based on HTK for Polish is presented. It was trained on 365 utterances, all spoken by 26 males. The features of Polish with respect to speech recognition are described. Some aspects of speech recognition differ in comparison to English. Errors in recognition were analysed in details in an attempt to find reasons and scenarios of wrong recognitions.

## 1. Introduction

There is little research in automatic speech recognition (ASR) system for Polish and there is no system which would work on sentence level for a relatively rich dictionary. Polish differs from the languages most commonly used in speech recognition like English, Japanese and Chinese in the same way as all Slavic languages. They are highly inflective and non-positional. These disadvantages are compensated by a specific feature of Polish language. The relation between phonemes and the transcription is more distinct.

We used HTK, developed at the Machine Intelligence Laboratory of the Cambridge University Engineering Department. It uses Hidden Markov Models (HMM) [1] as the recognition engine. While this solution seems to work well, it is necessary to add extra tools on grammar and semantic level to enlarge dictionary and retain good recognition results.

Errors in speech recognition have many different reasons [2].

Some of them appear because of phonetic similarities. We want to find other possible reasons for errors. Results are presented with deep analysis of wrong recognitions and types of substitutions. This knowledge may help in future speech recognition system design and in preparing data for corpora and model training. There are three general types of errors: random, systematic and gross. Random (or indeterminate) errors are caused by uncontrollable fluctuations of voice that affect parameterisation and experimental results. Systematic (or determinate) errors are instrumental, methodological, or personal mistakes causing lopsided data, which is consistently deviated in one direction from the true value. The detection of such errors is most important as then a model has to be altered. Gross errors are caused by experimenter carelessness or equipment failure which is quite unlikely here as we used a profesionally recorded data which were already used by other researchers.

In this paper we focus on HTK results for Polish and on detailed analysis of possible sources of errors. Section 2 describes aspects of Polish which are important for ASR in comparison with English. Section 3 provides information on ASR system and corpus we used. Section 4 describes triphone statistics for Polish which we collected. Section 5 contains results and description of all error patterns we found. We sum up the paper with conclusions.

## 2. ASR characterization of polish

To create Polish continuous speech recognition system, experiences gained during research on other languages can be used. English is the most elaborated language to build ASR system. As Polish and English are languages of the same Indo-European group, we focus on existing solutions for English ASR. Anyway there are some differences between these languages which have a larger or smaller impact on ASR. These differences result in some variations in algorithms and methods. We searched for differences which seem to be important in ASR.

English has a large number of homophones. What is more, many combinations of different words have

978-1-4244-1724-7/08/$25.00 ©2008 IEEE

similar pronunciation. Polish has much fewer homophones.

Pronunciation of vowels in English is very similar. If a vowel is not stressed it is usually pronounced as ɜ or ɪ. What is more, both of these phonemes have quite similar sounds and spectra. It means that unstressed vowels are almost indistinguishable in English. This contrasts with Polish.

Modern English has emerged as a mixture of about thirty languages. It results in quite simple general rules (which were necessary for a language to be widely accepted by different people) together with many irregularities (as a kind of residues), especially in pronunciation. Modern Polish bases strongly on Latin. In contrary to English, it resulted in complicated grammar rules and morphology but has quite few irregularities in pronunciation.

English is a positional language, while Polish is an inflective one. The meaning of a word in English depends strongly on the position of a word in a sentence. A position has a secondary importance in polish; an exact meaning of a word depends mainly on morphology. For example in English sentences ‘Mike hit Andrew’ and ‘Andrew hit Mike’ have quite different meanings. In Polish (using Polish similar names) ‘Michał uderzył Andrzeja’, ‘Michał Andrzeja uderzył’, ‘Andrzeja Michał uderzył’ and ‘Andrzeja uderzył Michał’ are all acceptable and mean almost the same. However all apart from the first one stress some part of information and sound quite strange without a special context of other sentences. To change subject and object we have to use another morphological ending ‘Andrzej uderzył Michała’. This fact means that the usage of language syntax models is very difficult for Polish and possibly not as crucial as for English. On the other hand, analyzing morphology seems to be crucial in case of ASR for Polish.

In English, conjugation and declension are relatively simple and adjectives do not need any type of agreement. In Polish there are groups of different ways of conjugation and declension. Each verb has typically different forms for each combination of gender (there are three basic genders in Polish, however, linguists distinguish 8 categories), person and singular or plural number. Each noun has 7 forms (cases) depending on the position and relation with other words in the sentence. Adjectives and numbers are agreed with the nouns they describe. There is no general rule of word morphological agreement, like adding ’s’ or ’es’ in English. Different groups of words have their own types of endings. The fact is that a single word in Polish may have several dozens of different flectional forms and even several hundreds of derived forms topically correlated (i.e. most verbs have almost 200 forms including conjugation of participle, perfect and imperfect forms). This fact causes a full dictionary of Polish language for ASR very complicated. Its size may cause delays of ASR system.

English is well known to have a vast vocabulary. It might be due to a large number of dialects and existing several versions of English situated all around the world. Polish dictionary seems to be smaller in this aspect.

Polish has a few phonemes which are rare in other languages and do not exist in English. They sound clearly different than other phonemes. Being more particular they have much higher frequency and sound to non-Polish speakers almost like rustles or hums. These phonemes are very easily detectable and as such can be additionally used as a kind of boundaries between blocks of other phonemes.

## 3. Hidden markov model toolkit (HTK)

The HTK [3] is a toolkit which use hidden Markov models (HMM), for ASR system. Research into speech synthesis, character recognition and DNA sequencing are its other applications. We used version 3.3 in our research. HTK consists of many modules and tools. All of them are available in C source form. The HTK provides facilities for speech analysis, HMM training, testing and results analysis. The system fits hypothesis of every recognition with one element from the dictionary, provided by a user. It compares possible phonetic transcriptions of words. The toolkit supports HMMs using both continuous density mixture Gaussians and discrete distributions.

Our system has been trained on part of a set called CORPORA [4]. Speech was recorded in an office with the working computer in the background. The database contains 365 utterances (spoken alphabet, digits, names and short sentences), each spoken by several females, males and kids (45 persons), giving 16425 utterances in total. One set spoken by male and one by female were hand segmented.

The rest were segmented by a dynamic programming algorithm using a model trained on hand segmented ones. The optimization was used to fit borders using existing hand segmentation of the same utterance spoken by two different people. Experiments were conducted on speech files with the sampling frequency f0 = 16 kHz. This gives sampling period t0 = 62.5 μs. All available utterances for 26 male speakers were used for training, considering all of them as single words in HTK model. We created the decision tree to find context making the largest difference to the acoustics and which should distinguish clusters using rules of phonology and phonetics in Polish [5] to create tied-state triphones. The Mel-frequency cepstral coefficients (MFCC) [6, 7] were calculated for parameterization. Twelve MFCCs plus an energy with first and second derivatives were used, giving a standard set of thirty nine elements. We used 25 [ms] windows for audio framing and 0.97 preemphasis filtering. Segments were windowed using Hamming method. Thirty seven different phonemes were distinguished using a phonetic transcription provided with the corpus.

## 4. Triphone statistics

A list of all possible triphones in a given language or dictionary is necessary to use HTK in its most effective configuration. We provided our research by analyzing triphones in Polish to create a general ASR system for Polish. It is not straightforward to obtain phonetic information from an orthographic text-data [8, 9]. Transcription of text into phonetic data has to be applied first [10]. We used PolPhone [11] software for this aim. The SAMPA extended phonetic alphabet was applied with 39 symbols and pronunciation rules typical for cities Krak´ow and Pozna´n. We altered the PolPhone phonetic alphabet to a 37 symbol version which is used in the CORPORA [4] and currently recognized as a SAMPA standard for Polish. We reduced the number of symbols by changing phoneme *c* to *k* and phoneme *J* to *g*. We also replaced *w~* to *o~* and *j~* to *e~*. These changes were done to adapt to an official standard version of SAMPA. It is frequently used, in example, in the audio corpus with transcription [4]. We preferred this one rather than an extended SAMPA used in PolPhone, which is going to be suggested as a new standard. For programming reasons, we used our own single letter only symbols, corresponding to SAMPA symbols instead of typical ones, to distinguish phonemes easier while analyzing received phonetic transcriptions. Statistics can be now simply calculated by counting number of occurrences of each phoneme, phoneme pair, and phoneme triple in analyzed text, where each phoneme is just one symbol. The analysis of the whole corpus took 3 weeks using PolPhone and scripts written in Matlab.

One of the key uses for this data is ASR system. This is the reason for quite specific choice of analyzed texts. Data for statistics were collected mainly from transcriptions of parliament meetings.

Total numbers of 148,016,538 phonemes were analyzed. They were grouped in 38 categories (including space). 1,095 different diphones and 14,970 different triphones were found. It has to be mentioned that all combinations like *#*, where * is any phoneme and # is space, were removed as we do not treat these triples as triphones. The reason for it is that first phoneme * and the second one are actually in 2 different words but we are interested in triphone statistics inside words. This list seems to be not fully representable because of text choice, specifically vast amount of parliament transcriptions, which caused probably some anomalies. Assuming 38 different phonemes (including space) and subtracting mentioned *#* combinations there are 53,503 possible triples. We found 14,970 different triphones which gives a conclusion that almost 28% of possible combinations were actually discovered as triphones existing in Polish. An average length of words in phonemes can be estimated as 6.22 due to space (noted as #) appearing with frequency 16.09.

Besides the frequency of triphones occurring, we are also interested in distributions of different frequencies. We expected to receive a very different distribution as very large amount of text was analyzed. We hoped to have very few triphones with occurrences smaller than 3 and deduce that they are not real triphones but errors due to foreign names etc. in the corpus. Even though we added extra text to the corpus several times, the distribution did not change much at all. We noted around 1600 triphones which occurred just once, 800 with occurrence 2, 500 with 3, 300 to 400 for 4 to 6 occurrences, 200 for 7 to 9, and up to 100 for 10 or more, every time after we analyzed extra text. Such phenomena is nothing unexpected in natural language processing on a level of words or above, where amount of analyzed text do not change statistics (considering reasonable large amounts). The open question is if we would find distribution we expected if we analyzed much bigger corpus or there is no limit in number of triphones lower than number of possible combinations. The new trigrams come from unusual Polish word combinations, slang and other variations of dictionary words, onomatopoeic words, and foreign words, errors in phonisation and typos in the text corpus. Still it is possible that the large numbers of triphones with very small occurrence are non-Polish triphones which should be excluded. In our further works we will assume that from statistical point of view it is not important, especially when smoothing operation is applied in order to eliminate disturbances caused by lack of text data [1].

## 5. Experimental results

The system was trained on 9490 utterances, 365 for each of 26 male speakers. The orthographic dictionary contains 365 elements, but due to differences in pronunciation between different speakers, the final version of the dictionary, using phonetic transcriptions, contains 1030 positions.

We started recognition evaluation using data of the only male speaker who was not used in training (Table 1). 6 out of 365 utterances were substituted giving correctness 98.36 %. Audio files of females, boys and girls were also recognized to check correlation between parameterisation of different age and gender. We received correctnesses 79.73%, 95.34% and 92.05% for adult female speakers. Child male speakers were recognized with correctnesses 60.55%, 95.07% and 75.62%. We noted correctness's 88.22% and 84.11% for girls. All non-adult male speakers gave clearly worse results, however there is no obvious difference between degradation in results related to age or gender. Even girl speakers, for which both age and gender differed from the training speakers, were recognized with the similar number of errors.

**Table 1. Word recognition correctness for different speakers (the model was trained on adult male speakers only)**

| speaker | age | gender | substitutions | correctness |
|---|---|---|---|---|
| AO1M1 | adult | male | 6 | 98.36 |
| AF1K1 | adult | female | 74 | 79.73 |
| BC1K1 | adult | female | 17 | 95.34 |
| BW1K1 | adult | female | 29 | 92.05 |
| AK1C1 | child | male | 144 | 60.55 |
| AK2C1 | child | male | 89 | 75.62 |
| CK1C1 | child | male | 18 | 95.07 |
| LK1D1 | child | female | 43 | 88.22 |
| ZK1D1 | child | female | 58 | 84.11 |

**Table 2. Errors in different types of utterances (for all speakers)**

| type | errors | being recog. | % of errors |
|---|---|---|---|
| sentences | 2 | 1026 | 0 |
| digits | 21 | 90 | 23 |
| alphabet | 130 | 297 | 44 |
| names and commands | 312 | 1872 | 17 |

**Table 3. Names which appeared most commonly as wrong recognitions**

| name | no. | name | no. | name | no. |
|---|---|---|---|---|---|
| Lucjan | 14 | Alina | 3 | Aniela | 2 |
| Marian | 8 | Bożena | 3 | Apolonia | 2 |
| Urszula | 8 | Diana | 3 | Benon | 2 |
| Daniel | 7 | Emilia | 3 | Celina | 2 |
| Joanna | 7 | Helena | 3 | Damian | 2 |
| Mariola | 7 | Ireneusz | 3 | Danuta | 2 |
| Beata | 5 | Rudolf | 3 | Izabela | 2 |
| Karolina | 5 | Julita | 2 | Maria | 2 |
| Marzena | 5 | Karol | 2 | Marta | 2 |
| Romuald | 5 | Lech | 2 | Oleńka | 2 |
| Alicja | 4 | Leonard | 2 | Renata | 2 |
| Anna | 4 | Leszek | 2 | Urban | 2 |
| Halina | 4 | Łucja | 2 | Zenon | 2 |
| Julian | 4 | Łucjan | 2 | Zofia | 2 |

**Table 4. Errors in pronounced alphabet**

| letter | errors | letter | errors | letter | errors |
|---|---|---|---|---|---|
| en | 9 | ce | 5 | a | 2 |
| em | 8 | e | 5 | es | 2 |
| er | 8 | ka | 5 | żet | 2 |
| pe | 8 | be | 4 | eł | 1 |
| će | 7 | de | 4 | ku | 1 |
| ą | 6 | ge | 4 | u | 1 |
| eń | 6 | i | 4 | wu | 1 |
| te | 6 | o | 4 | el | 1 |
| y | 6 | zet | 3 | ę | 1 |
| esz | 6 | eś | 3 | ef | 1 |
| źet | 6 | | | | |

Types of errors were carefully analyzed. First we checked percentage of correctly and wrongly recognized utterances depending on the type of utterances (Table 2). It can be clearly seen that smaller units are much more difficult to recognize: 44 % for one syllable units (spoken letters of alphabet), 23% and 17% for single words and almost no errors for sentences, even though we evaluated the system also on speakers of gender and age which were not used during the training. It suggests that recognition based on MFCC parameterization only is not sufficient. The context has to be used for allowing HMM models work correctly (or much better parameterization if possible).

All sentences were treated as single words during training and testing. The recognition of sentences is on an exceptional level, especially considering, that we used many speakers of gender and age not used during training procedure. The only two wrong recognition are quite bizarre. In both cases the correct transcription and wrong recognition are phonetically very different and very easily distinguishable for human listener.

There are several interesting detailed observations in patterns of wrong recognitions. Only one name was recognized as a sentence and quite few were recognized as spoken letters. The majority of wrong hypothesis were words which means the model deal property with length of utterances.

The very interesting fact is that even if names are recognized wrongly their gender is still correct most of the time. 79 female names were recognized as other female names, with only 17 female names recognized as male names. Some explanation might be that the majority of female names in Polish ends with 'a'. However, such phonological similarity is probably not strong enough for this effect. It is difficult to explain fully this phenomenon. The similar pattern was found in case of male names. 50 male names were wrongly recognized as other male names and only 14 male names were recognized as female names.

There are some pairs of phonologically similar names like Lucjan (l uh ts y aa n) and Lucian (w uh ts iy aa n) or Mariola (m aa r ih o l aa) and Marian (m aa r y aa n) which where quite commonly mistaken with each other. However, most of wrong recognitions seem to have no explanation like this. What is more some wrong detections with large phonological differences appear quite frequently. For example Barbara (b aa r b aa r aa) was recognized wrongly three times and all of them were as Marzena (m aa * eh n aa), where * is a special Polish phoneme similar to ’ge’. It has to be stressed that many pairs of very similar words were recognized quite correctly, like name Maria (m aa r y aa) was only twice recognized as Marian and Marian as Maria just once. We can conclude that phonological simmilarities can cause wrong detections but they seem to be not a major source of faults.

Table 3 shows names which were used as wrong hypotheses. There is an interesting tendency that these words were mostly correctly recognized when the audio with their content was analysed. It suggests that some utterences are generally more probable than others for the recognition of the whole set. We can say that they are represented more strongly in the language models. In a similar way names which were wrongly recognized rarely appear in Table 3 as they are weakly represented. It has to be stressed that all utterances were used 26 times (Table 3) during the training. The best example of this behaviour is a name Łucjan, which was recognized for virtually all test speakers as Lucjan. The name Lucjan was always correctly recognized. What is more Lucjan was provided as a hypothesis for several other names, including Jan (y aa n) which was recognized as Lucjan in case of two different speakers. In this example the name Lucjan was provided as a recognized word 23 times (including correct ones) and Lucian twice, in both cases incorrectly.

Table 4 presents wrongly recognized letters of alphabet. We already mentioned that this group is most likely to contain errors because these elements are very short and the HMM model cannot use all its advantages. We can also observe that sonants (n, m, r) tend to be the most difficult for recognition. Letters a, ha, jot were recognized correctly for all speakers.

## 6. Conclisions

Wrong recognized parts of speech are usually removed on the next step of speech recognition, i.e. by syntactic or semantic corrections. They are also very often ignored in analysis of systems. The effects considered in this paper appeared sporadically but we decided to analyze them carefully to improve our system. Nevertheless the detailed analysis enables us to find the reason of their occurrences. Some of these are difficult to explain at this level; however, we noted very interesting and strong tendencies. Sometimes it is possible to make some improvements in the part of speech recognition system which is responsible for the mapping of the orthographic transcription to the phonetic ones. The statistics of phonemes, diaphones and triphones were collected for Polish using a large corpus of mainly spoken formal language. The statistics are available on request by an email.

## 7. Acknowledge

We would like to thank Stefan Grocholewski and Institute of Computer Science, Pozna´n University of Technology for providing a corpus of spoken Polish CORPORA.

## 8. References

[1] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] S. Greenberg, S. Chang, and J. Hollenback, “An introduction to the diagnostic evaluation of switchboard corpus automatic speech recognition systems,” *Proceedings of NIST Speech Transcription Workshop*, 2000.

[3] S. Young, G. Evermann, M. Gales, Th. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book*, Cambridge University Engineering Department, UK, 2005.

[4] S. Grocholewski, “Zało˙zenia akustycznej bazy danych dlaje¸zyka polskiego na no´sniku cd rom (eng. Assumptions of acoustic database for Polish language),” *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*, 1995.

[5] M. Ke¸pi´nski, *Kontekstowe zwia¸zki cech w sygnale mowy polskie j(eng. Context feature relations in Polish speech signal), PhD Thesis*, AGH University of Science and Technology, Krak´ow, 2005.

[6] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357– 366, 1980.

[7] S. Young, “Large vocabulary continuous speech recognition: a review,” *IEEE Signal Processing Magazine*, vol. 13(5), pp. 45–57, 1996.

[8] J.N Holmes, I.G. Mattingley, and J.N. Shearme, “Speech synthesis by rule,” *Language and Speech*, vol. 7, pp. 127–143, 1964.

[9] D. Ostaszewska and J. Tambor, *Fonetyka i fonologiawsp´ołczesnego je¸zyka polskiego (eng. Phonetics and phonology of modern Polish language)*,

PWN, 2000.

[10] D. Oliver, *Polish Text to Speech Synthesis, MSc. Thesis in Speech and Language Processing*, Edinburgh University, Edinburgh, 1998.

[11] G. Demenko, M. Wypych, and E. Baranowska, “Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis,” *Speech and Language Technology, PTFon, Pozna´n*, vol. 7, no. 17, 2003.