PELT: An English language tutorial system for Polish speakers

Katarzyna Dziubalska-Kołaczyk, Anna Bogacka, Dawid Pietrala, Mikołaj Wypych, Grzegorz Krynicki

School of English & Center for Speech and Language Processing
Adam Mickiewicz University, Poznań, Poland
cslp@ifa.amu.edu.pl

Abstract

The Polish-English Literacy Tutor (PELT) is a multimodal multilingual tutorial system for foreign language learning (in this case English for adult Polish learners), and as such requires a specific speech recognition system dealing with highly accented, strongly variable second language speech. The aim of the paper is to present the challenges we encountered when preparing a new corpus of second language speech: phonetic characteristics of Polish English, corpus preparation and annotation, corpus statistics with the observed pronunciation problems of Polish speakers in English and the error-detector to be constructed. Solutions employed for PELT could be applied to accented foreigner speech recognition systems, e.g. English spoken by learners of various language backgrounds.

1. Introduction

In this paper we will present a work-in-progress report on the Polish-English Literacy Tutor (PELT) system for Polish learners of English.

The platform for developing the PELT system is the Colorado Literacy Tutor, developed by the Center for Spoken Language Research (CSLR), University of Colorado, Boulder, a comprehensive, scientifically-based reading program designed to teach children to read by interacting with a virtual tutor and through interactive books providing contextual feedback, reinforcement and individualized instruction. The technology developed for the Colorado Literacy Tutor involves automatic speech recognition, dialog systems and animated agents.

First, a tutorial system for native Polish was created. The corpus for the SONIC recognizer (CSLR) for native Polish contained recordings of selected sentences, recordings of 113 speakers (so far), annotations speech at sentence level, forced-time-aligned phonetic annotation and recognizer training. The most problematic challenge was the definition of visemes (lip shapes for phonemes) of three speakers, were video-recorded and in the first instance matched with the English visemes. SAMPA mappings of the English phonemes to visemes were adapted and used for phonetically similar Polish phonemes.

PELT is a particularly challenging tutorial development project, because it involves highly variable second language speech. Similar as the procedure is to collecting the PLT data, it requires more adaptations to the specific requirements of speech recognition in foreign, strongly accented speech, with a high degree of interpersonal variability. There are two major additional challenges: different characteristics of accented speech depending on the level of language proficiency of the learner, and aligning highly variable acoustic features to phonemes.

After comparing Polish and English phonology and describing Polglish (English with Polish accent) pronunciation, corpus collection procedures will be presented, i.e. both the design of the prompts as well as the criteria for the choice and proficiency evaluation of the subjects. Next, the system of annotation, based on the previously discussed features of Polglish, will be presented. The latter will be followed by corpus statistics. Finally, the necessary next steps towards the training of the Sonic recognizer on the collected corpus of Polglish data will be discussed.

2. Phonetic characteristics of Polish English

2.1. Polish vs. English phonology

Typologically, Polish differs from English by a number of essential phonological features. First of all, it is not stress-timed. As a result, vowels tend to maintain their quality and they may reduce to schwa (or be devoiced or deleted) only when phonostylistically conditioned. Other important rhythm-related differences concern: word stress – in Polish it is fixed on a penultimate syllable, and consonantal clusters – Polish is much richer in clusters in all word positions than English. Secondly, Polish is not weight-sensitive, neither in terms of vowel quality nor syllable weight. It also does not appreciate diphthongs. Thirdly, the segmental inventory of Polish is much nearer to the average balance between vowels and consonants (ca. 6 to over 20, according to [1]) than English.

As far as system adequacy is concerned, the inventory of Polish vowels is entirely different from the English one, while in consonants, there are some important systemic as well as distributional differences. For example, Polish lacks dental apical fricatives while it has dental laminal obstruents; the distribution of a velar nasal is restricted to homorganic pre-velar-stop contexts.

Finally, on a universal dimension, Polish is unmarked with reference to the process of word final obstruent devoicing, as well as interconsonantal voice agreement.

2.2. Polglish pronunciation: predictions

A Polish learner of English is predicted to have pronunciation problems stemming from all the above mentioned discrepancies between the Polish and English sound systems. The resulting errors will either be directly L1-induced (i.e. caused by the interference of the system-adequate features of Polish), or caused by the type-specific or universal processes.

An example of a typical L1-induced error is the substitution of some Polish dental or labio-dental obstruent (fricative or stop) for the English apical dental fricative.

The typological rhythm difference leads, among others, to the inability to reduce unstressed vowels as well as the difficulties in stress placement.

Word-final obstruent devoicing is probably the most notorious characteristics of Polglish, and predictably so, since this is a universal phonological process reinforced in Polish speakers by the system-adequacy.

The above are only illustrations of the predictable mispronunciations of English by Polish learners, since a systematic survey is not possible within the scope of this paper. For the needs of PELT the most representative selection of Polglish errors has been made, with a view to sensitivize the recognizer towards those features which most reliably distinguish particular levels of proficiency of the learners.

3. Corpus preparation and annotation

In order to prepare the recognizer for different characteristics of accented speech depending on the level of proficiency, many more speakers have to be recorded than in the case of native speech recognition. It is necessary to record learners of different levels of advancement. Since the general level of English does not closely correspond to pronunciation skills, we control the number of speakers declaring a given level of advancement in English. The speakers will be divided into proficiency groups by means of statistical tests performed on the number and quality of errors they make. The speech of any user beginning to use the program will be compared to the group characteristics and the users will thus receive training at the appropriate level.

Corpus collection has so far been based on sentences which had been used for recording native American English speakers. These prompts were designed to ensure maximum diversity of phonetic contexts when elicited from native speakers of English, and as a result they also contained the maximum range of contexts a foreign learner might have problems with.

The recording scenario so far included only read speech, as spontaneous speech at this stage of recognizer training would be too variable. Currently there are recordings of 116 speakers included in the corpus. Each speaker recorded 50 sentences, each set of sentences being different for each speaker. The speakers controlled the tempo of recordings themselves and were allowed to repeat a sentence if they wished to do so. 85 females and 31 males were recorded. Speakers' age ranged from 16 to 43, with the mean age 21,9 years and standard deviation 4,4 years. Speakers were controlled for the level of English: 24% were at the First Certificate in English level, 62% were at the Cambridge Advanced Certificate in English level, and 14% were at the Cambridge Proficiency Examination level. 71,6% declared to have been learning British English accent, 27,6% American English accent, and 0,9% were hesitant. Subjects were also asked to name geographical regions they came from and other foreign languages they spoke.

The entire PELT corpus: sentences, labeling files and technical specifications, is approximately 3,5 GB in size and contains a total of 6032 files, corresponding to 14h 37min 37sec of running speech.

The recordings were recorded, annotated and stored following EAGLES [4] and [5], IMDI [6], and OLAC [7] recommendations. The recordings were recorded using Edirol UA 25, one channel, 24-bit resolution, and 44100 sampling frequency. Recordings were performed in a quiet office in order to obtain realistic data for tutorial system environments. A dedicated user-friendly interface was added to a simple recorder based on MCIWin functions. Audio files were checked for misreadings, repetitions etc. as they were elicited, and if necessary the speaker was asked to repeat a sentence.

The PELT database was annotated by a group of students of English who completed a two year course in English phonetics. They were supposed to listen to the recordings, compare them to all its acceptable native readings and annotate the differences by means of a predefined tagging notation. "All acceptable native readings" were understood as all pronunciations accepted by educated native speakers of the standard variety of English identical to the variety declared by the subject in the interview that preceded the recording session, i.e. Received Pronunciation (RP) or General American (GA). We additionally assumed these "acceptable native readings" to be produced without disfluencies and noises. The taggers were instructed to refer to pronunciation dictionaries in the case of doubt what forms are acceptable.

For the recording protocol and annotations an XML format was used. It is anticipated that the learner corpus resource will be adapted for a wide range of teaching and speaker applications.

4. Corpus statistics

This quantitative summary reports on the analysis of 65 transcripts read and recorded by 65 subjects and tagged for errors, disfluencies and noises. The ongoing work is aimed at tagging all of the 116 transcripts corresponding to 116 speakers in the database as well as extending the speech database.

Departures from the transcript in the speech of the subjects were divided into phonetic and non-phonetic ones.

The list of phonetic errors (*Table 1*) was compiled on the basis of two empirical studies ([3] and one by Jarosław Weckwerth, private communication). The number and type of errors to be used in the annotation of PELT was, on the one hand, a result of a compromise between the predicted discrimination and classification power of the speech recognizer trained on the data, and the pedagogical usefulness of the tool in teaching English phonetics to Polish students on the other.

Table 1 presents the numbers and percentages of 10 error types grouped into 7 major categories (vowel errors types were lumped into a one category, other types errors into another category).

Table 1: Phonetic error type frequencies in PELT

	Error type	**Source of likely Polglish error (error in brackets)**	**Count**	**%**
CONSONANTS	/ŋ/	velar nasal (/ŋg/, /ŋk/, /n/) e.g. everything */ˈevrɪθɪŋk/	229	5,3
		/ŋ/+V with no /g/ (/ŋgV/) e.g. singer */ˈsɪŋgə/		
	voicing of consonants and voicing of consonant clusters	voiced /dɪs/ or /mɪs/ (/z/) e.g. this boy */ðɪz bɔɪ/	1361	31,7
		final voiced obstruent (devoicing) e.g. disguise */dɪsˈgaɪs/		
		voiced obstruent + /s/ or /s/ + voiced obstruent (regressive assimilation) e.g. absurd */əpˈsɜːd/		
	consonant clusters	/tʃt/, /dʒd/ etc. word-finally (schwa insertion) e.g. attached */əˈtætʃət/	14	0,3
	place of articulation	/θ/ → /f/, /s/ etc., /ð/ → /z/ etc. (except /ŋ/ → /n/) e.g. think *[sɪŋk]	506	11,8
	manner of articulation	/ʃ/ → /tʃ/ etc. e.g. cliché */ˈkliːtʃeɪ/	40	0,9
VOWELS	/ə/ or /ɜː/	schwa quality and/or quantity e.g. cater */keɪter/	1419	33,0
	mono phthongs	vowel quality error, vowel nasalisation e.g. fenced *[ˈfɛɯ̃ĩst]		
	di/tri phthongs	/eə/ or /ɪə/ (/j/ breaking, schwa) in RP e.g. tier */trʲə/		
		/ʊə/ (/w/ breaking, schwa, /u/) in RP e.g. poor */pʊʷə/		
OTHER	word stress	stress placement errors e.g. astronomy */æstrəˈnɒmɪ/	730	17,0
		secondary stress (reduced to unstressed) e.g. impartiality */ɪmpəʃɪˈælətɪ/		
	variety of English	inconsistence in the use of RP or GA e.g. after */ˈæftə/ instead of /ˈɑːftə/ or /ˈæftər/		
		total	4299	100

Non-phonetic departures from the acceptable native readings included word-level errors, disfluencies, restarts and noises.

Word-level errors (*Table 2*) included word deletion, word insertion, word order error, substitution of a transcript word by a different yet existent English word and misreadings – substitution of a transcript word by a different and non-existent word assuming this substitution was not motivated directly and solely by the phonetic difficulty of the transcript word but by not knowing what it means or just wrong reading of the transcript.

Table 2: Non-phonetic errors: word-level errors

Error	**Count**	**%**
deletion	305	24,7
insertion	297	24,0
word order	4	0,3
misreading	379	30,7
substitution	250	20,2
total	1235	100

Disfluencies included pauses, hesitated chunks and filled pauses. Hesitated chunks consisted of word(s) produced with hesitation, usually at a slower pace and possibly with pauses within words. The set of fillers and acknowledgements was adopted after [8].

Table 3: Disfluencies: pauses, hesitations, fillers and acknowledgements

Error	**Count**	**%**
pauses	174	50,1
hesitated chunks	138	39,8
filler (*um*, *hm*,...)	35	10,1
total	347	100

Restarts tagged in the corpus followed the notation presented in [9].

Table 5: Disfluencies: restarts

Error	**Count**
restarts	412

Noises tagged in the corpus included aside remarks, audible inhaling or exhaling, laughter, cough, throatclear, sniffing, steps, etc.

Table 6: Noises

Error	**Count**
Noises	193

5. Automatic error detector

The speech corpus presented in the paper is to be used as training data for automatic pronunciation errors detector. The goal of the detector is to automatically determine the type (and possibly intensity) of pronunciation errors occurring in English speech produced by Polish native speakers.

The pronunciation error typology presented in the previous sections will constitute the basis for the preparation of accompanying acoustic models and pronunciation models.

The problem of specialized models for a given type of pronunciation errors can be seen as a fine grained variant of accented speech recognition techniques described in [10] or [11].

The detector, given an acoustic observation sequence and an orthographic transcript is to evaluate the observation sequence using each of the acoustic and pronunciation models. The resulting scores for each model will allow to measure the intensity of pronunciation error by comparing the score of the error model to the score of the native English model. For the purpose of scoring comparable additional normalization factors need to be extracted from the acoustic and pronunciation models.

The implementation basis for the project is Sonic continuous speech recognition system developed at CSLR ([12]).

6. Summary

This paper presents work on preparing a speech recognizer to recognize highly variable, strongly accented Polglish. The underlying assumption has been that the input to the recognizer based on the comparison of English and Polish phonological systems and the annotation including errors made by Polglish speakers can help train the speech recognizer to recognize Polglish. This venture is supposed to be informative and valuable for both second language phonetics research by providing computational verification of linguistic hypotheses and for speech recognition by providing very challenging testing material.

7. Acknowledgements

We would like to thank Wojciech Dziejma, Anna Kubisz, Paula Orzechowska and Paulina Zydorowicz who collected most of the data and greatly helped in the annotation. Thanks go to Włodzimierz Sobkowiak for providing his PDI (Phonetic Difficulty Index) and the list of Polglish errors. Jarosław Weckwerth shared with us his observations on the degree of difficulty of particular features.

None of this would have happened without Ron Cole's encouragement and assistance, and generous sharing of the CSLR technology with our team.

8. References

[1] Maddieson, I. 1999. In search of universals. *ICPhS99.* vol. 3. 2521-2528.

[2] Sobkowiak, W. 2004. Phonetic Difficulty Index. In W. Sobkowiak and E. Waniek-Klimczak (eds.) Dydaktyka fonetyki języka obcego. Zeszyt Naukowy Instytutu Neofilologii Państwowej Wyższej Szkoły Zawodowej w Koninie nr 3. Konin: Wydawnictwo PWSZ w Koninie. 102-107.

[3] Sobkowiak, W. 2005. "Phonolapsological equivalence and similarity in the English lexicon". In F.Kiefer, G.Kiss & J.Pajzs (eds). 2005. Papers in computational lexicography, Proceedings of the 8th International Conference on Computational Lexicography (COMPLEX 2005), Budapest, Hungary, 16-18 June 2005. Budapest: Linguistics Institute. 200-212.

[4] Gibbon, D., R. Moore and R. Winski (eds.). 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

[5] Gibbon, D., I. Mertins and R. Moore. 2000. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation.* Dordrecht: Kluwer Academic Publishers.

[6] IMDI. http://www.mpi.nl/ISLE

[7] OLAC. http://www.language-archives.org/

[8] Heeman, Peter A. and James Allen. 1995. The Trains 93 Dialogues. TRAINS Technical Note 94-2

[9] Marie Meteer et al. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus.

[10] Ayako Ikeno, Bryan Pellom, Dan Cer, Ashley Thornton, Jason Brenier, Dan Jurafsky, Wayne Ward, William Byrne. 2003. "Issues in Recognition of Spanish-Accented Spontaneous English". In: ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, April.

[11] Kumpf, Karsten and Robin W. King. 1996. „Automatic Accent Classification of Foreign Accented Australian English Speech". In: Proceedings of ICSLP 1996 Vol. 3, Philadelphia, USA. pp 1740—1743.

[12] Pellom, Bryan. 2001. "SONIC: The University of Colorado Continuous Speech Recognizer". University of Colorado, Technical Report #TR-CSLR-2001-01, Boulder, Colorado.