Pronunciation training in L2 Polish using a CAPT system AzAR

Agnieszka Wagner, Adam Mickiewicz University, Poznań, Poland

ABSTRACT

The role of CAPT (computer-assisted pronunciation training) systems can be especially important in the practice of speaking skills. Oral proficiency is fundamental for all language learners, but it requires more time than a traditional language classroom can offer. The progress in automatic speech recognition (ASR) has brought new opportunities for applying this technology in CAPT. Although some limitations are still present, a number of studies have shown that corrective feedback provided by ASR can be effective in improving pronunciation skills. This paper describes the development of a CAPT system AzAR for the language pair L1 German/L2 Polish. The system performs speech signal analysis on the users' speech input, which allows for the comparison of the user's and tutor's pronunciation, detects and evaluates mispronunciations and provides multimodal feedback. In the paper the development of the linguistic content, the results of pronunciation analyses and their application, and the functionalities of the software are presented. In the end, the application of AzAR to individual pronunciation training is discussed and a critical evaluation of the system is presented.

Keywords: CAPT; pronunciation and prosody training in L2; speech technology

1. Introduction

Computer-assisted language learning (CALL) and computer-assisted pronunciation training (CAPT) play an increasingly important role in second language (L2) education (Eshani – Knodt 1998, Cho et al. 2007, White – Gillard 2011). For many years, in L2 learning, grammar and vocabulary were favoured over pronunciation training. Today it is acknowledged that language learners have to acquire *various* skills to function properly in the foreign language, but very often some skills, particularly speaking and pronunciation, are not practiced sufficiently due to traditional classroom constraints (e.g. limited in-class time) or psychological factors (e.g. learning at self-paced speed, high level of anxiety while producing speech in a foreign language). In this context, CALL and CAPT systems can be beneficial to L2 learners as they provide a private, stress-free environment in which students can access virtually

unlimited input, practice at their own pace and, through the integration of automatic speech recognition (ASR) receive individualized, instantaneous feedback. Compared to classroom instruction, CALL and CAPT systems make it possible to address individual problems and to present the student only with certain tasks specifically related to the student's needs and skills. Such student profiles can be stored by the system in a log-file so that it is possible to monitor problems and improvements individually.

The advantages of CALL and CAPT have motivated research and development of numerous systems over the last 20 years. Consequently, at present there are quite many CALL and CAPT products available on the market (Eskenazi – Hansma 1998, Franco et al. 2000, Mak et al. 2003, Wang et al. 2009, Strik et al. 2013). However, at a closer look they are not entirely satisfactory for language teachers and individual learners, due to limitations in the input, feedback and output (see section 2).

This paper describes the development of a CAPT system AzAR for the language pair L1 German/L2 Polish. It concentrates on the linguistic input to the software, and to a lesser extent, on its technological features and its application to pronunciation training in L2 Polish. The paper also presents a critical evaluation of the software in regard to state-of-the-art pedagogical guidelines for pronunciation teaching (see section 2).

2. Review of literature

At least since the nineties, it is well known that a reasonably intelligible pronunciation is an essential component of communicative competence. As numerous studies showed tailor-made training may result in highly intelligible or even native-like pronunciation (Bongaerts et al. 1997, Derwing et al. 1998). On the other hand pronunciation difficulties can place speaker at a professional or social disadvantage (Morley 1991) and perceived accentedness can lead to significantly lower status ratings (Brennan – Brennan 1981). In order to provide an effective pronunciation using CAPT it is necessary to consider factors that affect L2 pronunciation most severely and to specify pedagogical guidelines for the teaching of pronunciation. Since it is impossible or very difficult to control age of L2 learning and amount of native language (L1) use which have significant effect on the degree of foreign accent (Piske et al. 2001), these guidelines should take into account those variables that can be monitored, namely *input*, *output* and *feedback* (Neri et al. 2002).

As regards *input*, it should be interactive, multimodal (audio-visual e.g. film fragments and radio interviews) and provided in a considerable amount by different native speakers. The input should be meaningful to the learner including simulations of real-life situations (e.g. in interactive dialogues or role-

plays). To encourage speech production (the *output*) exercises should not be limited to 'listen and repeat' of isolated words, sentences or minimal pairs – they should be realistic, varied and engaging to stimulate interaction between the learner and the system. The software should provide detailed information on the way the target speech sounds are to be produced by explaining (in a multimodal way) position of the articulators. For this purpose systems should offer 3D representations or computer animations of the lips and oral cavity accompanied by a written explanation. As regards *feedback*, the system should provide learners with immediate, comprehensible and individualized information on their performance. The feedback should be based on ASR technology specially tuned for non-native speech recognition (Cucchiarini et al. 2002, Mak et al. 2003, Jokisch et al. 2005). It is necessary that the system can detect *and evaluate* pronunciation errors, and provide constructive and meaningful explanation concerning mispronunciations, as well as clear instruction on how to improve the performance. The feedback system should be programmed to select and address only those errors which are *frequent*, *persistent*, *perceptually relevant* and which can be *reliably detected* (Neri et al. 2002).

3. Study: Developing a CAPT system for L2 Polish

An example of an ASR-based CAPT system is AzAR developed in two projects in the years 2005-2010. Initially, the system was created for the language pair L1 Russian/L2 German (Jokisch et al. 2005) and later it was extended to Polish, as both target and source language (Demenko et al. 2009), and Slovak and Czech. The system uses expert's knowledge on typical errors made by L2 learners caused by interference with their native language phonology and phonetics. AzAR architecture separates the structure from the content, which enables adaptation of the system to a new language or set of exercises.

In the following sections linguistic content created for the version of the software for L2 Polish and features of the ASR-based feedback system are presented (sections 3.1-3.5), and application of the software to pronunciation training in L2 Polish is discussed (section 3.6). In the end, a critical evaluation of the software is provided (section 3.7).

3.1. Speech databases

Two speech databases were collected: *development database* containing non-native Polish speech and *reference database* containing recordings of the model speakers (one professional actress and one actor). Apart from that, *source language databases* containing native German and Polish speech (both

developed in previous projects) were used for the purpose of ASR training and testing.

The development database was created for the purpose of ASR adaptation, studying foreign accent and learning processes, and collecting evidence of mispronunciations typical of Polish learners with L1 German on which the curriculum and the feedback system were based. It consists of recordings of sentences containing phonetic phenomena that can pose difficulty for speakers with L1 German and words that have alternative pronunciations according to the dialect spoken, three sets of phonetically rich and balanced sentences, three phonetically rich passages and spontaneous speech (for details see Cylwik et al. 2009).

18 German and 19 Polish speakers were recorded (ca. 54 hours of speech). The non-native speakers of Polish represented different proficiency levels (A1-C2) and were recruited from among students studying temporarily in Poznan and attending a Polish course and lecturers of the Faculty of Modern Languages and Literature at Adam Mickiewicz University. The distribution of proficiency level and gender is balanced among the speakers.

The reference database consists of recordings of text material implemented into the tutoring system as curriculum for German learners of Polish and includes isolated words, minimal pairs, and sentences.

All the recordings were conducted in a high-standard studio with no perceivable reverberation or background noise. The whole speech material was automatically transcribed and aligned at phoneme, syllable and word level and manually corrected. The non-native speech material from the development database was also annotated with respect to deviations from the canonical pronunciation: substitutions, insertions and deletions.

In the labelling of mispronunciations the labellers could select from among the phoneme sets of L1 and L2 or from the inventory of intermediate phonemes – a set of labels describing approximations to Polish or German phonemes and diacritics available in the IPA alphabet. The labelling was carried out in the speech analysis software *Praat*.

3.2. The analysis

The analysis of speech production in non-native Polish aimed at investigation of phonetic interferences between German and Polish and identification of pronunciation errors which are frequent, persistent, perceptually relevant and which can be reliably detected by the ASR technology. The analyses were based on speech material (1433 utterances) from 9 German learners of Polish. The results showed that the number of pronunciation errors decreased with the proficiency level of the learner, but at the same time distribution of

substitutions, deletions and insertions was similar in all proficiency groups (Figure 1).

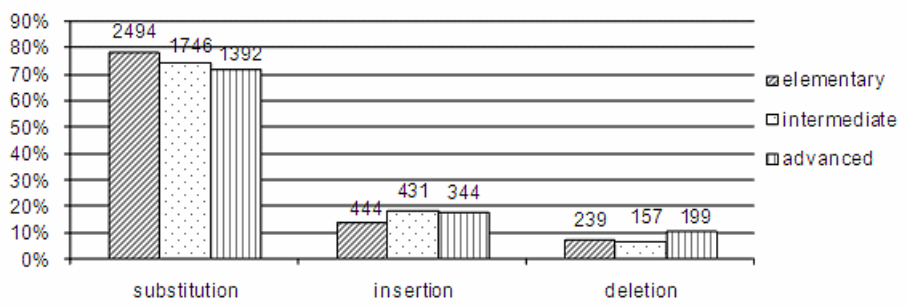

Figure 1. Frequency and distribution of pronunciation errors in different proficiency groups

The results indicated that German learners of Polish had most difficulty in the realization of:

- fricatives and affricates (especially palatal) which do not exist in German e.g. /ɕ/, /ʑ/, /tɕ/, /dʑ/, /dz/
- Polish graphemes ą and ę
- sequences of vowels followed by glides (/j/, /w/) which are pronounced as German diphthongs (e.g. /aj/ – /aI/) or tense long vowels (e.g. /ej/ – /e:/)
- voiced/voiceless contrast
- final unstressed vowels (which became reduced)
- consonant clusters

As in previous studies on non-native pronunciation (Flege et al. 1995, Wells 2000) three sources of errors were identified: 1) substitutions or deletions of sounds and consonant clusters that do not exist in German or are too difficult to pronounce, 2) carry-over of pronunciation regularities from L1 German, and 3) overgeneralization of L2 Polish regularities.

The results of the analyses were applied to create a comprehensive curriculum for pronunciation training in L2 Polish (see section 3.3) and to formulate mispronunciation hypotheses which constitute an integral part of the acoustic feedback generation system (section 3.4).

3.3. Curriculum

The curriculum consists of more than 30 exercises (674 utterances) and is divided into segmental and suprasegmental section. The segmental section includes exercises on vowel contrasts, glides, consonant contrasts and clusters, facultative and obligatory assimilations and gemination. More advanced exercises include optional palatalization and assimilation and difficult consonant clusters. The suprasegmental section consists of exercises on regular and irregular word stress, prosodic foot, clitics and sentence-level intonation. Each exercise (a minimal pair or a sentence) in the curriculum was provided with a recording of the model pronunciation realized by two professional actors.

3.4. Mispronunciation hypotheses

For each training unit in the curriculum mispronunciation hypotheses were explicitly defined as lists of possible erroneous realizations in relation to the canonical pronunciation of the target word. Whenever more than one hypothesis could be defined, they were ranked according to their effect on speech intelligibility from the least to the most severe ones (Table 1).

Table 1. Example of the annotation of mispronunciation hypotheses (SAMPA transcription)

text	canonical pronunciation	mispronunciation hypotheses
pęk	p e N k	p e-E~ N- k; p e-w~ N- k
pąk	p o N k	p o-O~ N- k; p o-w~ N- k

3.5. Features of the feedback system

During the training the learner selects specific unit of the curriculum, listens to the model utterance realized by the reference voice and records her own production of this utterance. The system displays a waveform of the model utterance and of that realized by the learner (Figure 2). The learner can listen to her own realization and to compare it with that of the reference voice.

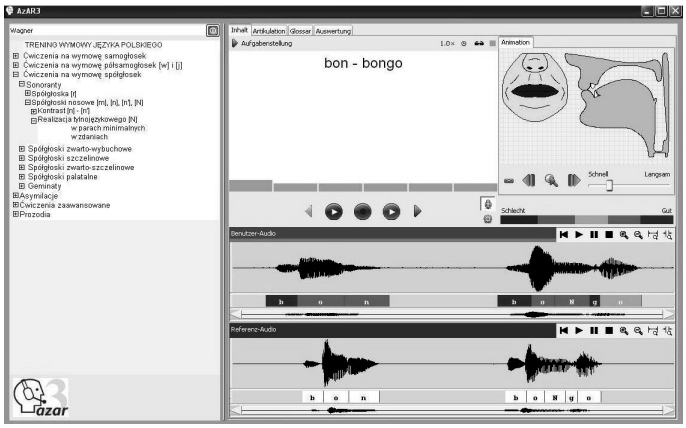

Figure 2. Training session in AzAR (/n/-/N/ contrast in minimal pairs)

The software uses Hidden Markov Model-based speech recognition and speech signal analysis on the learner's input which makes it possible to compare (visually and auditively) user's own performance with that of the reference voice's. Automatic error detection is performed on the phonemic level: all uttered phones are marked using a colour scale to give the learner immediate information on the overall output quality (Figure 3). An additional visual mode includes animated display of the vocal tract (lips area and articulatory movements).

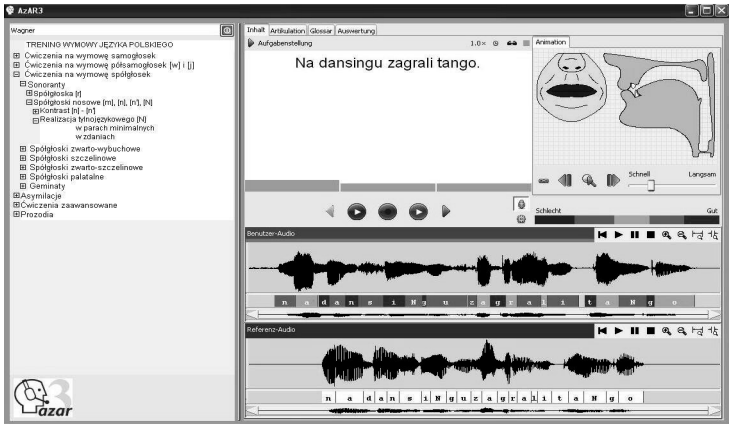

Figure 3. Training session in AzAR (/n/-/N/ contrast in sentence)

AzAR allows keeping track of the learner's performance, so that identification of the features that should be practiced is possible and the learner's progress can be monitored. Additionally, a comprehensive tutorial in phonetics is provided which, among other things, explains how to interpret the acoustic displays. For each exercise in the curriculum a passage containing information on the classification, features and articulation of the phone is provided, as well as a sagittal slice of the vocal tract during the phone production and pictures of the lip area and tongue position.

3.6. Pronunciation training in L2 Polish with AzAR

AzAR for the target language Polish has been applied to individual training of Polish pronunciation and prosody by foreign students attending Polish Phonetics course at AMU. Since the time for pronunciation training and instruction offered during the course is limited, it is indispensable that students practice individually – they are provided with the software and are encouraged to train with it at home to achieve higher oral proficiency. Such training gives them the possibility to practice at their own pace, in a stress-free environment and to concentrate on specific pronunciation problems. The students' opinions on the system were collected and positive feedback was reported.

3.7. Critical evaluation of the software

Based on the observations made during the implementation phase it can be said that the current version of AzAR for L2 Polish should be regarded as a prototype, because it has some serious pedagogical and technological limitations. First of all, the *input* provided to the learner should be extended by including more than just two native Polish speakers, because exposition to L2 speech by multiple speakers enhances perception of novel contrasts and consequently, leads to improved production. The input could also be more varied and engaging in order improve learner's motivation by providing real-life materials such as radio interviews or short film episodes, and by including interactive exercises e.g. dialogues or role plays. As regards *output*, at present, pronunciation training in AzAR is limited to "listen and repeat" drills of isolated words, sentences and minimal pairs. In the future, the system should offer more exercises aiming at learning aspects related to production of connected speech (e.g. dialogues or role plays) and addressing different individual cognitive styles (Neri et al. 2002). As concerns prosody training, the detection and evaluation of learner's production should be based not only on the acoustic-phonetic features (i.e. comparison of the target and learner's F0 – fundamental frequency – contours and durations using dynamic time warping

technique), but it also needs to take into consideration communicative and linguistic functions prosodic parameters (F0, duration and intensity). For this purpose some higher-level representation of prosody that links acoustic-phonetic realization with linguistic function could be used. As regards the ASR-based feedback, the system yields rather poor performance results: preliminary tests indicated high percentage of false alarms (correct realizations indicated as incorrect by the system) and missed mispronunciations (incorrect realizations evaluated as correct). Consequently, in the future formal tests should be carried out that take into account various aspects of the ASR system performance. Their results will constitute the starting point for further optimization of the ASR system using more, carefully transcribed native and non-native speech data.

4. Conclusions

The pronunciation tutoring software AzAR for L2 Polish is a knowledge-based system that uses ASR technology and meets strict pedagogical requirements whose lack is one of the major drawbacks of the existing CAPT applications (Neri et al. 2002). Even though the current version of the software has some serious limitations in scope of the input, output and feedback, it gives the learner access to meaningful training materials and can be helpful to develop sensitivity to phonetic contrasts in L2 Polish, and to build awareness of articulatory phenomena. In the future, formal evaluation of the system performance is necessary to provide basis for optimization of the applied ASR technology. Robust and reliable error detection and evaluation are prerequisites for further extension of the input and output provided to the learner.

References

Bongaerts, Theo – Chantal van Summeren – Brigitte Planken – Erik Schils
1997 "Age and ultimate attainment in the pronunciation of a foreign language", *Studies in second language acquisition* 19 (4): 447-465.

Brennan, Eileen M. – John S. Brennan
1981 "Accent scaling and language attitudes: Reactions to Mexican American English speech", *Language and Speech* 24(3): 207-221.

Cho, Kwansun – John G. Harris – Ratree Wayland
2007 "Effectiveness of a robust computer assisted pronunciation training tool", *The Journal of the Acoustical Society of America* 122: 3017.

Cucchiarini, Catia – Helmer Strik – Lou Boves
2002 "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech", *The Journal of the Acoustical Society of America* 111: 2862.
Cylwik, Natalia – Agnieszka Wagner – Grażyna Demenko
2009 "The EURONOUNCE corpus of non-native Polish for ASR-based Pronunciation Tutoring System", in: Name Surname (ed.), PAGES.
Surname, Name (ed.)
2009 *Proceedings of SLaTE workshop on speech and language technology in education Wroxall Abbey Estate, Warwickshire*. Location: Publisher.
Demenko, Grażyna – Agnieszka Wagner – Natalia Cylwik – Oliver Jokisch
2009 "An audiovisual feedback system for acquiring L2 pronunciation and L2 prosody", in: Name Surname (ed.), PAGES.
Surname, Name (ed.)
2009 *Proceedings of 2nd ISCA workshop on speech and language technology in education*, *SLaTE*. Location: Publisher.
Derwing, Tracey M. – Murray J. Munro – Grace Wiebe
1998 "Evidence in favour of a broad framework for pronunciation instruction", *Language Learning* 48 (3): 393-410.
Ehsani, Farzad – Eva Knodt
1998 "Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm", *Language Learning and Technology* 2 (1): 45-60.
Eskenazi, Maxine – S. Hansma
1998 "The fluency pronunciation trainer", Name Surname (ed.), PAGES.
Surname, Name (ed.)
1998 *Proceedings of the STiLL workshop*. Location: Publisher.
Flege, James E.– Murray J. Munro – Ian R. MacKay
1995 "Factors affecting strength of perceived foreign accent in a second language", *The Journal of the Acoustical Society of America* 97: 3125???
Franco, Horacio – Victor Abrash – Kristin Precoda – Harry Bratt – Kristin Precoda – Ramana Rao – John Butzberger – Romain Rossier – Federico Cesari
2000 "The SRI EduSpeakTM system: Recognition and pronunciation scoring for language learning", in: Name Surname (ed.), 123-128.

Surname, Name (ed.)
2000 *Proceedings of InSTILL 2000.* Location: Publisher.

Jokisch, Oliver – Uwe Koloska – Diane Hirschfeld – Rüdiger Hoffmann
2005 "Pronunciation learning and foreign accent reduction by an audiovisual feedback system", in: Name Surname (ed.), 419-425.

Surname, Name (ed.)
2005 *Affective Computing and Intelligent Interaction*. Heidelberg: Springer Berlin.

Mak, Brian – Manhung Siu – Mimi Ng – Yik-Cheung Tam – Yu-Chung Chany – Kin-Wah Chan – Ka-Yee Leung – Simon Ho – Fong-Ho Chong – Jimmy Wong – Jacqueline Lo
2003 "PLASER: pronunciation learning via automatic speech recognition", in: Name Surname (ed.), 23-29.

Surname, Name (ed.)
2003 *Proceedings of HLT-NAACL*. Location: Publisher.

Morley, Joan
1991 "The pronunciation component in teaching English to speakers of other languages", *Tesol Quarterly* 25 (3): 481-520.

Neri, Ambra – Catia Cucchiarini – Helmer Strik – Lou Boves
2002 "The pedagogy-technology interface in computer assisted pronunciation training", *Computer Assisted Language Learning* 15 (5): 441-467.

Piske, Thorsten – Ian R. MacKay – James E. Flege
2001 "Factors affecting degree of foreign accent in an L2: A review", *Journal of phonetics* 29 (2): 191-215.

Strik, Helmer – Joost van Doremalen – Jozef Colpaert – Catia Cucchiarini
2013 "Development and Integration of Speech technology into Courseware for language learning: the DISCO project", in: Name Surname (ed.), 323-338.

Surname, Name (ed.)
2013 *Essential Speech and Language Technology for Dutch*. Heidelberg: Springer Berlin.

Wang, Hongcui – Christopher J. Waple – Tatsuya Kawahara
2009 "Computer Assisted Language Learning system based on dynamic question generation and error prediction for automatic speech recognition", *Speech Communication* 51 (10): 995-1005.

Wells, John C.
2000 "Overcoming phonetic interference", *English Phonetics*, *Journal of the English Phonetic Society of Japan* 3 (2000): 9-21.
White, Erin – Sharlett Gillard
2011 "Technology-based literacy instruction for English language learners", *Journal of College Teaching and* Learning *(TLC)* 8 (6): 1-6.