www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

ARCHIVES OF ACOUSTICS
Arch. Acoust., **35**, 3, 309–329 (2010)

DOI: 10.2478/v10168-010-0027-z

The Use of Speech Technology in Foreign Language Pronunciation Training

Grażyna DEMENKO[(1),(2)], Agnieszka WAGNER[(1)], Natalia CYLWIK[(1)]

[(1)] *Adam Mickiewicz University*
Institute of Linguistics
Department of Phonetics
Al. Niepodległości 4, 61-874 Poznań, Poland
e-mail: {lin, wagner, nataliac}@amu.edu.pl

[(2)] *Poznań Supercomputing and Networking Center*
Z. Noskowskiego 12/14, 61-704 Poznań, Poland

(*received April 14, 2010; accepted July 20, 2010*)

In recent years the application of computer software to the learning process has been found to be an indisputably effective tool supporting the traditional teaching methods. Particular focus has been put on the application of techniques based on speech and language processing to the second language learning. Most of the commercial self-study programs, however, do not allow for introduction of an individualized learning course by the teacher and to concentrate on segmental features only. The paper discusses the use of speech technology in the training of foreign languages' pronunciation and prosody and defines pedagogical requirements for an effective training with CAPT systems. In this context, steps taken in the development of the intelligent tutoring system AzAR3.0 (German '*Automat for accent reduction*') in the scope of the Euronounce project (Cylwik *et al.*, 2008) are described with the focus on creation of the linguistic content. In response to the European Union's call for promoting less widely spoken languages, the project focuses on German as a target language for native speakers of Polish, Slovak, Czech, and Russian, and *vice versa*. The paper presents the design of the speech corpus for the purpose of the tutoring system and the analysis of pronunciation errors. The results of the latter provide information which is important for Automatic Speech Recognition (ASR) training on the one hand, and for automatic error detection and feedback generation on the other hand. In the end, Pitch Line software for implementation in the prosody visualization and training module of AzAR3.0 tutoring system is described.

Keywords: speech technology, pronunciation training, feedback, phonetic-acoustic databases, multimedia tools.

1. Introduction

The increasing use of speech technology can be especially seen in the area of foreign language education, which has led to the development of a new discipline known under the name of Computer-assisted language learning (CALL). The literature on CALL mentions a number of its potential advantages: elimination of time limitations and dependence on the teacher, possibility to work at the user's own tempo and to store his/her profile in order to monitor the progress, constant access to a number of additional materials such as visualizations, recordings and animations as well as elimination of stress related to the fact that the learner is being listened to by his/her classmates, the last of which is particularly important while acquiring L2 (second language) pronunciation and prosody. Computer-assisted language learning affords a creative approach in the field of language teaching, e.g. didactical constructivism in multimedia environment based on project-oriented collaboration of students and teachers. Additionally, the implementation of information technology in a multilingual language tutoring system augments the interest of young people who are familiar with computers and communication devices.

For many years, pronunciation was neglected in favor of grammar and vocabulary in language learning, because it was generally believed that training had no significant impact or even could have a detrimental effect on pronunciation (Krashen, Terrell, 1983; Scovel, 1988). However, the acknowledgement that a reasonably intelligible pronunciation constitutes an essential component of communicative competence and the results of later research which showed that an appropriate training could significantly improve pronunciation (e.g. Bongaerts, 1999) caused an increasing interest in pronunciation learning. One of the results was the development of the first CAPT (Computer-assisted Pronunciation Training) systems (for an overview see Subsec. 2.2).

Due to the inherent complexity, lack of knowledge in adequate prosody processing both for linguistic and technological needs and to the ensuing difficulty in their acquisition, intonation and other prosodic phenomena like rhythm and voice quality were ignored in language teaching for many years. There appear to be several reasons for the generally growing interest in intonation, which we have been witnessing in recent years. First, there have been important new advances in the theory of intonation, its functions and forms, aided by the growing accessibility of acoustic signal analysis, processing and interpretation. Second, the expansion of the analytical domains of traditional linguistics from sounds and words to larger units of inquiry such as phrases, discourse and interactions, has drawn attention to such subfields as pragmatics, discourse analysis, and conversation analysis. Third, applied linguistics has grown to give priority to the communicative function of prosody rather than its linguistic form.

The current research in the field of CALL focuses on a more effective integration of computer technology with the learning and teaching of languages and

on including the prosodic factor. The main goal of this paper is, therefore, to integrate both the segmental and suprasegmental aspects, especially in discourse and interaction, and to suggest a complex framework for studying foreign language pronunciation with AzAR3.0 (Jokisch *et al.*, 2005; Cylwik *et al.*, 2008) software.

The core of AzAR3.0 is a pronunciation trainer with integrated multilingual speech recognition and audio-visual feedback functionality. The system provides speech signal analysis of the user's speech input, which allows the user to compare his/her pronunciation with that of the tutor's, and to receive selective information on how to improve articulation.

AzAR3.0 includes PC-based and PDA-based learning programs, a web-based tool for providing the learning content ('authoring system') and a reference speech database for a given language pair combination (German as a target and Russian, Polish, Czech or Slovak as source languages and *vice versa*). In order to ensure effectiveness of the courseware, the system was designed according to strict pedagogical guidelines.

2. Challenges in computer-assisted language training

2.1. Pedagogical requirements for CAPT

Some of the factors affecting pronunciation learning cannot be manipulated, e.g. the first language and length of exposure to L2 whereas others can be controlled and thus, effectively used in CAPT systems. The latter include: input, output and feedback (Neri *et al.*, 2002b). First of all, the learner should have access to a large number of L2 speech samples from multiple speakers, to be able to improve perception of novel contrasts and to build models of general phonetic categories. Apart from the auditory channel, the visual one should be available (visualizations of the articulatory movements), as the combination of the two channels improves the production and perception of L2 contrasts. It is important that the input is meaningful to the learner and that the learning context is relevant and realistic (e.g. by including scenarios resembling real life situations) in order to stimulate the learner's motivation.

The second factor, i.e. the output, refers to speech production. Output is essential for pronunciation improvement – the strong accent of some long-term residents suggests that mere exposure to L2 (input) is not sufficient for that purpose. The production of their own output gives the learners a chance to test their hypotheses on the L2 sounds. By comparing their output with the input model, learners can form correct L2 representations. The exercises for speech production should be varied, realistic and engaging so as to boost learners' motivation. They should not be limited to listening and repeating minimal pairs or isolated sentences, because the controlled practice does necessarily lead to an improvement in a real conversation. Interaction with the system, e.g. in role-plays with char-

acters or interactive dialogues, is a more effective method, but it can be applied only if ASR technology is specially tuned for the recognition of non-native speech is available. What could also be useful for the learner from the pedagogical point of view, is the possibility to self-monitor the progress.

The third essential factor in pronunciation training is the feedback. Explicit and detailed feedback is necessary to make the learner aware of the discrepancies between his/her production (output) and that of the model speaker (input). Interferences from L1 which can be considered as the main source of pronunciation errors can sometimes prevent learners from perceiving such discrepancies. As shown in (Lyster, 1998), immediate feedback is essential – so, ideally, it should be provided by means of ASR technology. It should involve the aural and visual channels and provide assessment of these errors which are crucial for intelligibility, and thus it should take into account both the segmental and suprasegmental features. However, beside scoring learners' production (ideally, on the basis of some hierarchy of mispronunciations with regard to intelligibility), the feedback should provide information on the type and location of an error and instruct the learner how to correct it.

The tutoring system AzAR 3.0 presented in this paper meets the pedagogical requirements defined here. It provides the learner with an extensive input which consists of speech produced by two model speakers per language (reference database, see Subsec. 3.2). Apart from the aural channel, AzAR offers the visual mode including animated visualization of the vocal tract (lip area and articulators movements) as well as a formants graph for specific phones and a display of the speech signal under the transcribed and phonemically segmented reference utterances. For the output, a comprehensive curriculum was provided. It was created on the basis of extensive analyses (Subsec. 3.3) of non-native speech databases (see Subsec. 3.2) and thus it concentrates on these segmental and suprasegmental aspects, which are crucial for intelligibility and are typical and common among L2 learners with specific L1. In the end, the application of ASR technology in AzAR3.0 makes it possible to generate individualized and immediate corrective feedback (details are given in Sec. 4).

2.2. Audio and visual training

The application of multimedia tools for audiovisual feedback to detect deviations from standard articulation in the target language has shown an especially high effectiveness in PC-based learning pronunciation systems. Although prosody visualization seems to be more complex, speech analysis has been used for teaching L2 intonation patterns since 1970s, e.g. (Abberton, Fourcin, 1975; de Bot, Mailfert, 1982). The main principle is that the sound waveform or pitch contour of the student's utterance are visually displayed alongside those of the teacher's. An example of a program that displays visual pitch curves is a product of Kay Elemetrics called Visi-Pitch, that has been available for a number of

years for DOS-based personal computers. With Visi-Pitch, the students are able to see both the reference speaker's and simultaneously their own intonational curve. A highly valued system in which clear and intelligible feedback is provided on intonation, stress and rhythm, is *BetterAccentTutor* developed by Kommissarchik (2000) for teaching American English prosody to non-native speakers of English.

The main shortcomings of hardware and software used currently for prosody training can be summarized as follows:

- Technical aspects:
  - Low speech signal energy;
  - No extrapolation for voiceless sounds;
  - Not entirely correct/reliable F0 extraction;
  - Lack of voice quality visualization.
- Methodological shortcomings:
  - Lack of user-friendliness, i.e. learners do not know how to interpret displays and to evaluate results;
  - Examples and exercises are focused on word and sentence-level intonation;
  - Lack of integration of such prosodic features as tone, duration, loudness;
  - Lack of voice quality analysis – even when a learner can produce individual sound segments which are very similar to those produced by the teacher, they may still sound 'wrong' due to overall voice quality.

It should also be noted that the importance of auditory and/or visual feedback with regard to prosody is difficult to assess, because computer programs providing feedback require the learners to monitor and evaluate themselves critically. Apart from visual display, no further feedback is provided and there is a lack of objective assessment.

2.3. Speech technology in language learning

Although the majority of recent studies have demonstrated the effectiveness of audio and visual training in improving learners' perception, several problems are still unsolved.

One of the problems found in some of the earlier software programs was the lack of feedback processing, for instance the learner's pitch could be measured and instantly displayed but interruptions in the intonation contour during voiceless parts of the utterance and the inclusion of perceptually irrelevant pitch variations made it difficult for the learner to interpret the feedback.

Speech synthesis is not widely used in CALL systems. As it often sounds artificial, at present most developers seem to prefer recordings of natural voices. However, there have been attempts to investigate the possibility to use synthetic stimuli in language teaching. The potential for using speech synthesis in CALL

applications lies in the text-to-speech synthesis, and especially in integrating speech synthesis with visual models of the face, mouth and vocal tract.

A number of studies, e.g. (Dalby, Kewley–Port, 1999; Anderson, Kewley–Port, 1995), on the other hand, showed the usefulness of automatic speech recognition in pronunciation training. Nevertheless, ASR systems perform poorly when confronted with non-native speakers (Morgan, 2004) and various methods have been proposed to enhance their performance with non-native speech, e.g. (Goronzy, 2002; Bouselmi *et al.*, 2007). The effectiveness of CALL systems based on speech recognition is not only determined by the capabilities of the speech recognizer but also by (a) – the type of feedback and teaching method and (b) – the inclusion of repair strategies to safeguard against the recognizer's error. Unfortunately, at present speech recognition systems are poor at handling information contained in the speaker's prosody. The limitations of the technology imply that the learner's utterances have to be predictable and that the detection of errors is only possible with a limited degree of detail, which makes it difficult to give the learner corrective feedback. However, speech recognition systems can be used to measure the speed at which the learners speak and the rate of speech has been shown to correlate with L2 learners' proficiency.

Advanced PC-based learning systems (e.g. Pronunciation Power, American Sounds, Phonics Tutor, Eyespeak) include (verify http://www.learningvillage.com/html/guide.html or https://calico.org/CALICO%20Review/): 1) speech analyzing windows or frames, 2) Internet-based features like email answering, online help and chat sessions with human tutors, 3) animated visualizations of the articulatory mechanics, video clips showing jaw, lip and tongue movements and waveform patterns of sound samples. Users are able to record sound files and to acoustically compare a graphical representation of their utterances to those of the instructor. A few systems (e.g. Fonix iSpeak 3.0, Pro-Nunciation) include synthesized speech or TTS solutions. During the last decade, speech recognition technology has been implemented into innovative interactive systems like Istra and Pronto (Dalby, Kewley–Port, 1999) and in the European research project Interactive Spoken Language Education ISLE (http://nats-www.informatik.uni-hamburg.de/~_\isle/). The ISLE system targets German and Italian learners of English and aims at providing feedback focusing mainly on word-level errors, i.e. it points to mispronunciation of specific sounds and lexical-stress errors. While this feedback design seems satisfactory, the system yields poor performance results. Non-native speech databases have been created, and some researchers have investigated the possibility to improve the performance of speech recognizers and to try various probabilistic models to produce pronunciation scores from the phonetic alignments generated by HMM-based acoustic models (Teixeira *et al.*, 2000). In the FLUENCY project (Eskenazi, Hansma, 1998), a speech recognizer was used to detect foreign speakers' pronunciation errors for L2 training. The research also involved prosody and the correlation between pronunciation and prosody errors was investigated. However, neither the information on the

placement of the intonation errors nor suggestions on how to improve the intonation were provided, leaving the comparison task to the users. A well-known software, Tell Me More of Auralog, improved error detection and feedback for pronunciation practice by pointing out erroneous phonemes and showing a 3D animation to visualize the 'standard' articulation. Through ASR, the computer recognizes the student's utterance and moves on to an appropriate conversational exchange. A similar method is used by U.S. Army researchers and by the developers of *TraciTalk* to design game-like programs to teach L2 (Neri *et al.*, 2002a). In this case, the student orally asks the computer to perform a task such as 'put the book on the table'. If the computer understands the utterance, it will perform the action required by the student. This type of feedback is very effective in reinforcing the correct pronunciation behavior. However, all these programs are unable to offer help if a student cannot make him/herself intelligible because, for instance, he/she cannot correctly pronounce a sound. Additionally, their technology for suprasegmentals (concerning only intonational aspects) is very limited.

3. Towards optimal technology for L2 pronunciation training

3.1. Euronounce project

Intelligent Language Tutoring System with Multimodal Feedback Functions (acronym Euronounce) is a project within the framework of European Commission's Lifelong Learning Programme, which aims at creating L2 pronunciation and prosody teaching software. The Euronounce project was preceded by two earlier projects carried out by the Euronounce coordinator, TU Dresden, between 2004 and 2007. As a result, an audio-visual systems AzAR and AzAR2.0, aimed at teaching Russians German pronunciation, were created (Jokisch *et al.*, 2005). Following the baseline developed in these projects, the Euronounce project aims at creating software for German as a source language (L1) and Polish, Slovak, Czech as target languages (L2) and *vice versa*, beside segmental adding of suprasegmental exercises. In accordance with the current emphasis on communicative and socio-cultural competence, more attention is paid to discourse-level communication and to cross-cultural differences in pitch. In order to achieve these goals, a new version of AzAR (AzAR3.0), developed in the scope of the Euronounce project, will also improve the underlying speech and visualization technology, e.g. by a more-in-depth and more specific analysis of prosodic features.

3.2. Speech databases and speaker selection

It seems clear that in order for pronunciation tutors to be successful not only target, but also source language needs to be taken into account (Neri *et al.*,

2002a; Eskenazi, 1999). It is understandable if we keep in mind that most errors result from L1 and L2 interference and consist primarily in transferring allophonic and phonotactic rules from our mother tongue to the target language and replacing L2 phonemes with their most similar L1 counterparts (Flege, 1995). Taking only L2 into account is one of the main flaws of ASR-based pronunciation tutors as they mostly fail to recognize non-native speech. For that reason, in the development of AzAR 3.0 for the language pairs Polish-German and German-Polish (as well as for other languages), three speech databases were created:

1. Reference database – target language speech by professional speakers of the target language;
2. Non-native speech database – target language speech by non-native speakers;
3. Source-language accent database – source language speech by source language native speakers.

The ***reference database*** consists of a set of reference utterances for a given L1–L2 pair, uttered by two native speakers (one male and one female), which serve as template utterances for exercises which are designed in the way that allows practicing production as well as perception at the phonemic and prosodic level in isolated words, simple phrases, complex phrases and continuous speech.

The ***non-native database*** consists of recordings of non-native speech produced by 36 speakers per language pair (18 speakers per language direction). The speakers were recruited from among students of the target language, with a different degree of proficiency specified according to Common European Framework of Reference for Languages, i.e. levels A1-A2, B1-B2, C1-C2. The database was balanced with respect to sex and proficiency of the speakers. The proficiency ratings were based partly on the information provided by the students in a questionnaire which also included self-judgment of their language skills.

The non-native database contains recordings of six tests including different types of texts:

Accent test – is a collection of 125 sentences containing those Polish (and other target languages) sounds and phonetic phenomena – which are considered as difficult from the point of view of a German learner, e.g. Polish [x] in words such as 'ich' (Eng. 'their') which Germans might pronounce as [C].

Dialectological test – 124 sentences containing words with alternative pronunciations according to the dialect spoken and a full range of Polish phonemes in different contexts, word and sentence positions, vowels in minimal pairs, e.g. tik:tak (Eng. 'tick':'yes'), consonants in oppositions voiceless vs. voiced, e.g. pić:bić (Eng. drink:hit), vowels in stressed and/vs. unstressed positions, e.g. ma:mama (Eng. 'has':'mum'), etc.

Spontaneous speech test – addressed only to more advanced students. It consists of four simple tasks such as finishing a sentence, e.g. 'My hobby is...' and

explaining the meaning of a proverb or idiomatic expression, commonly known both in Poland and Germany. The purpose of the test was to collect speech samples for the final assessment of the learners' proficiency level.

Continuous speech test – three passages (72 sentences altogether) taken from stories by H.Ch. Andersen and Grimm Brothers, two of which were addressed to upper-intermediate and advanced students only.

Phondat corpus – it contains three sets of phonetically rich and balanced sentences (341 altogether) for the purposes of ASR training and testing, and collecting mispronunciations of consonant clusters. Polish is an exceptionally consonantal language allowing for sequences of even four or five consonants in a word, e.g. drgnąć (Eng. 'to quiver'), which is often a source of pronunciation errors of the foreigners.

Prosody test – a set of 59 sentences designed in such a way so as to collect evidence of those prosodic errors that are most easily detectable and most crucial for comprehension, e.g. erroneous stress placement or non-native-like vowel duration. The purpose of the test was to investigate the realization of prosodic/intonational features by advanced L2 learners and L1 interferences in the domain of prosody. Details concerning the structure of the test are given in (Cylwik *et al.*, 2009).

Except for the spontaneous speech test, the recorded speech material was automatically transcribed and aligned on the phoneme level. Speech samples in Polish were annotated using a modified version of Polish SAMPA (Demenko *et al.*, 2003) and non-native German speech data were annotated according to German SAMPA (www.bas.uni-muenchen.de/Bas/BasSAMPA).

The ***source-language accent database*** has been collected for the "general" speech recognizer training. It consists of at least 50 hours of speech provided by more than 100 speakers.

The recordings for the speech databases were conducted using the WiGE rec software, in a simple studio with low noise and reverberation. Basic quality requirements were: sampling frequency 44.1 kHz, minimal resolution 16 bit, minimal SNR of 35 dB. The recordings for the reference speech database were conducted in a professional studio to ensure highest quality.

3.3. Statistical and linguistic analysis of non-native Polish and German speech databases

The analyses were carried out on the basis of a subset of Polish and German non-native databases. The whole speech material was annotated at the segmental and suprasegmental (lexical and sentence stress) level. Guidelines for the labeling of pronunciation errors (classified as substitutions, insertions, deletions) and other phenomena (e.g., different types of acoustic and non-linguistic interferences, stress shifting, word-internal pause insertions) occurring in the non-native speech,

were defined. Firstly, a trained phonetician – a native speaker of the target language – verified and corrected the canonical transcription which was then used to produce a phonetic segmentation. Subsequently, the annotator identified the portions of the signal perceived as mispronounced and marked deviations from the canonical pronunciation and accentuation. As expected, the analyses of the annotated speech material from the accent and prosody tests provided by fifteen Polish and nine German native speakers (representing four different proficiency levels) showed that the number of pronunciation errors decreased with the proficiency level of the learner: beginners made significantly more errors (42.67% of all pronunciation errors) than, intermediate (31.35%) or advanced learners, who contributed the least to the overall number of pronunciation errors (26%, all results normalized with respect to the number of learners in each group). The distribution of different types of errors (substitutions, insertions and deletions) is similar in different proficiency groups (Fig. 1).

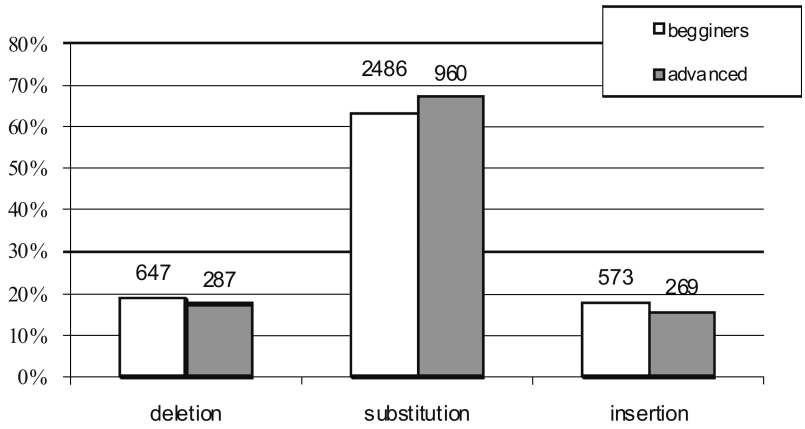

Fig. 1. Frequency and distribution of pronunciation errors in different proficiency groups.

As regards pronunciation, German learners of Polish have most difficulty in:

a) pronunciation of Polish palatal and non-palatal fricatives (/s′/, /z′/, /S/) and affricates (/d^z/, /d^z′/, /t^s/, /t^s′/) especially if they occur in clusters,
b) pronunciation of palatal consonants (/n′/, /s′/, /z′/, /t^s′/, /d^z′/) which are most often realized as their non-palatal counterparts (/n/, /s/, /z/, /t^s/, /d^z/),
c) realization of the voiced – voiceless contrast which is the main basis of the phonological opposition between consonants in Polish, but not in German where the lenis – fortis contrast plays the major role; as a result, German learners tend to devoice phonologically voiced consonants or to realize them similarly to German lenis consonants (perceptually salient partial devoicing),

d) pronunciation of Polish graphemes [ą] and [ę] which are mapped to sequences consisting of the vowel /o/ and /e/ followed by nasalized /w/ (in case of [ą]) or /j/ (in case of [ę]) or /n/, /m/, /n′/ depending on the following context; German speakers pronounce [ą] and [ę] as monophthongs: nasal /E∼/ or /O∼/ or use Polish GTP rules, but in inappropriate contexts,
e) pronunciation of sequences of vowels followed by semi-vowels (/j/, /w/): they are substituted by German diphthongs (e.g. /aj/ replaced by /aI/) or tense long vowels (e.g. /ej/ pronounced as /e:/),
f) avoiding reduction of final unstressed vowels (there is no /@/ in Polish).

The most frequent pronunciation errors found in the Polish – German corpus (recordings of Polish speakers reading speech material in German) concern:

a) vowel production: about 78% of tense long vowels were mispronounced by the Polish learners (see Fig. 2 below) – realization of short vowels (both lax and tense) was significantly less problematic as they are more similar to Polish vowels with respect to both acoustic and articulatory features,
b) pronunciation of German diphthongs /aI/, /OY/, /aU/ which are realized as sequences of a vowel followed by a glide (/aj/, /oj/, /aw/),
c) production of consonants absent in Polish: /h/, /C/ which were most often pronounced as /x/,
d) realization of /N/ without velar burst in words such as *fängt, jüngerer, jungen*.

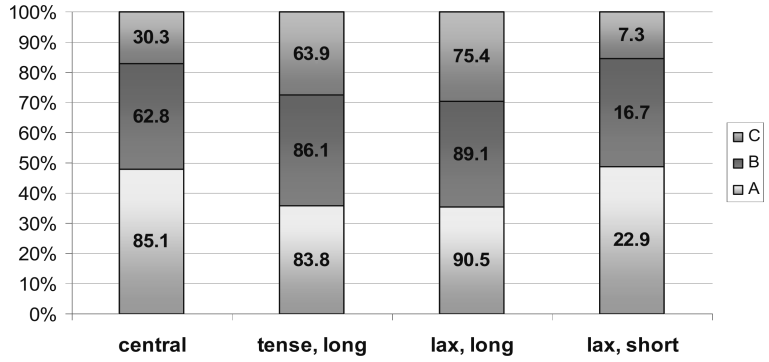

Fig. 2. Frequency and distribution of pronunciation errors in different proficiency groups.

The results discussed here were used to create a curriculum of production and perception exercises for a comprehensive L2 pronunciation and prosody training in AzAR. Apart from that, the knowledge of typical errors made by L2 learners was used to define mispronunciation hypotheses which are used to detect and assess pronunciation errors in AzAR.

4. Feedback system

4.1. System performance

Lack of proper (or any) feedback is often named as the most serious flaw in educational software, e.g. (Dalby, Kewley–Port, 1999; Engwall *et al.*, 2004). Good software should not only assess the degree of pronunciation correctness but also instruct on how to improve it, show where exactly the error has been made, e.g. which phone has been produced erroneously and offers feedback that is easy to interpret. To answer these needs, the AzAR3.0 software provides a multimodal feedback – it includes visual and audio modules in the form of curriculum recordings by a reference voice and the visualization of the speech signal under the transcribed and phonemically segmented reference utterances. The Euronounce project combines the objective features of an ASR system with the knowledge of human experts. The system cannot detect all errors which the human teacher could find and, in contrast to other approaches, does not try to do so. Nevertheless, it is able to reliably detect errors which the human expert has identified for a given corpus of test utterances.

The software uses HMM-based speech recognition and speech signal analysis on the learner's input which makes the visual and aural comparison of the user's own performance with that of the reference voice possible. Even though ANN and SVN could give better recognition results at the phone level, currently HMM-based classifier was used because of the reliable methodology of building such classifiers that was developed in previous projects. In the future it is planned to test the ANN and SVN classifiers.

From a technical point of view, designing a voice-interactive pronunciation tutor goes beyond the state of the art required by commercial dictation systems. While the grammar and vocabulary of a pronunciation tutor are relatively simple, the underlying speech processing technology tends to be quite complex since it must be customized to recognize the halting speech of language learners. Acoustic models are generalized so as to accept and recognize correctly a wide range of different accents and pronunciations. In general terms, the procedure consists of building native pronunciation models and then measuring the non-native responses against the native models. For the need of the preliminary evaluation of Polish acoustic speech models, 116 h of orthographically annotated speech produced by 321 speakers were used. The speakers were split into 5 cross-validation sets in such a way that the number of speakers and the number of sentences were approximately equal across different cross-validation 'folds'. Each set serves as a testing corpus once, with the other four used for building of the models. The stochastic acoustic speech models for Polish were trained using HTK tools (Young *et al.*, 1997). The number of Gaussian mixtures in each state was experimentally set to 24. Polish phonetic alphabet consisting of 39 phonemes (Demenko *et al.*, 2003) was used as the default setup. For the present experiment, it was decided to divide *each* of the six Polish vowels (/i/,/y/,/e/,/a/,/o/,/u/) into two

separate monophones, one representing lexically stressed instances and the other representing their unstressed counterpart. The list of contextual questions was accordingly modified to distinguish stressed and unstressed phonemes. The extension of the above method consisted not only in dividing vowel models, but also distinguishing 'sonorants' placed within stressed syllables from those located inside unstressed syllables. The results of the acoustic modeling (without any linguistic models) are summarized in Table 1.

Table 1. Acoustic modeling results for different train and test speech-rate classes. 'mu' denotes mean word-level accuracy [%], 'sig' denotes a standard deviation across 5 cross-validation folds.

Setup	Number of phonemes	mu	sig
Standard	39	45.82	1.27
With stressed vowels	45	49.10	1.18
With stressed vowels and sonorants	52	49.62	1.52

The difference between standard (39 phonemes) and vowel-distinguishing setup is statistically highly significant (a matched-pairs difference is 3.28% with a standard deviation of 0.65%). Further extension of the phonetic alphabet, with two monophones for each sonorant, does not give any significant boost when tested by the one-tailed t-test (a mean fold difference of 0.52% with standard deviation 0.96% gives a P-value 14.7%).

For the auditory feedback, a German phoneme recognizer was trained and adapted to the specific German speech data (recorded in the reference database). The alignment to generate the spoken transcription for the educational speech material was carried out by a mixed German and Polish phoneme recognizer, considering a lexicon with several alternatives for each word.

The best-scored phone sequence wins over the alternative sequences. The dynamic matching between canonical and aligned phone sequence provides a warping function of temporal information and information about the best-matching areas of the speech signal. The system marks the potential area of wrong pronunciation in the utterance and suggests suitable exercises to reduce the speaker-specific pronunciation errors. All uttered phones are marked using color scale from red for mispronounced phones to green for those pronounced correctly. The user can listen to and play back the reference voice as well as see the speech signal for a given utterance, record and listen to his/her own utterance and see the speech signal for this utterance, and finally get feedback on his/her own pronunciation. An additional visual mode includes animated visualization of the vocal tract (lip area and articulators movements) and a formants graph for particular phones. A typical AzAR template for an exemplary minimal pair is showed in Fig. 3 and Fig. 4. From top to bottom, the panel containing the text to be produced is shown together with the formants/articulation graph, the oscillo-

gram/spectrogram of user's and reference speaker's utterances: below them the transcription and segmentation panel can be seen.

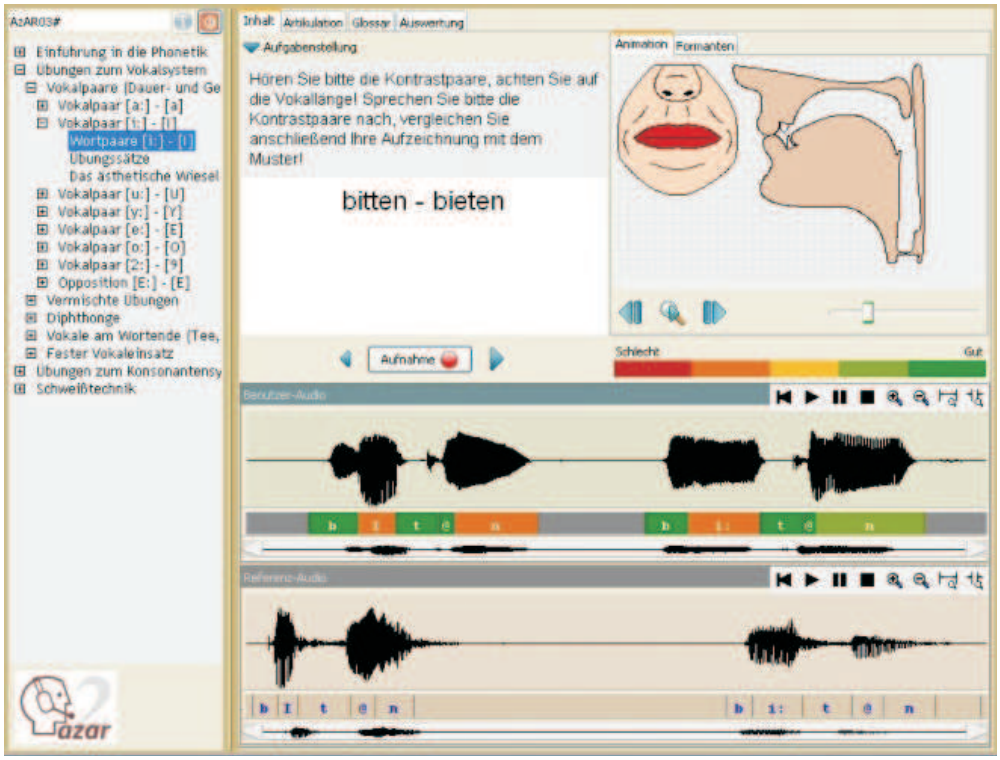

Fig. 3. AzAR template for pronunciation assessment of an exemplary minimal pair "bitten – bieten" (DE). In the right top corner animated visualization of the vocal tract (lips area and articulators movements) are displayed.

Pronunciation quality is visualized by colors. The segments that appeared problematic in this example (marked in a light and mid-grey) are German vowels /i:/, /I/, /a/, /E/, /6/ and consonants: /r/, /N/ – all of them have different articulation from the corresponding Polish phonemes (additionally, there is no /6/ in Polish), so substitutions can be expected.

All parts of the display are easily customizable by the learners to fit their individual needs.

For pronunciation training, also traditional instruction is being recommended (Chun, 1998) since visuals can be too difficult for the user to interpret and listening drill is not enough when one keeps in mind that L2 learners tend to perceptually associate foreign sounds with more familiar L1 sounds. Therefore, beside audio-visual feedback, the AzAR3.0 software also includes a text tutorial on articulatory and basic acoustic phonetics with glossary, phonemes description and classification, anatomic information, etc.

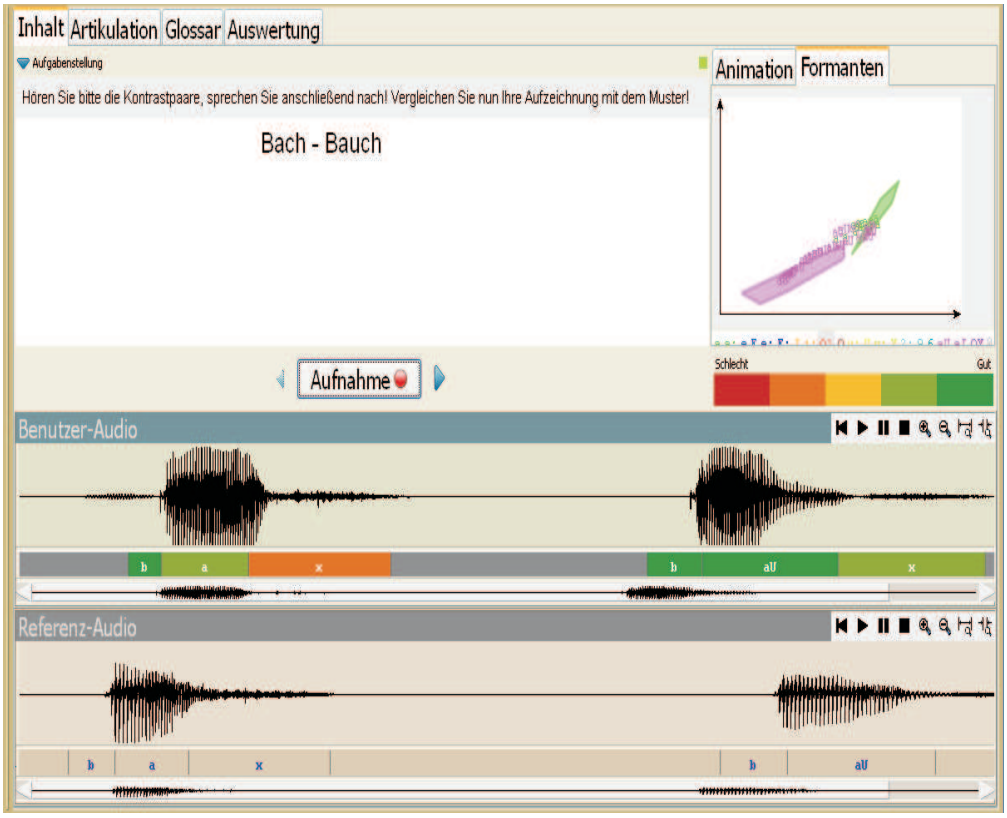

Fig. 4. AzAR template for the pronunciation assessment of a word pair "Bach – Bauch". In the right top corner a formants graph for particular phones is displayed.

4.2. Suprasegmental feedback

In spoken conversation, intonation and stress information not only help the listeners to identify phrase boundaries and word emphasis, but also the pragmatic thrust of the utterance (e.g. interrogative vs. declarative). One of the main acoustic correlates of stress and intonation is the fundamental frequency (F0); other acoustic features are loudness, duration, and tempo. In order to provide an effective feedback on prosody, training software should visualize the "relevant" intonation pattern of a given utterance as realized by an L2 student and a native speaker. Apart from that, it should draw attention to acoustic features involved in the realization of intonation. For example, the software could (a) instruct learners to compare the steepness of their falling or rising pitch movement to that of the native speaker, and/or (b) provide a quantitative measurement of the actual pitch slopes of both the native speaker and the learner. An effective feedback of this kind requires implementation of some kind of pitch stylization and normalization. The Pitch Line program (Demenko, Wagner,

2007) designed for approximation and parameterization of intonation contours responds to these needs and therefore it could be useful for the AzAR3.0 environment.

The stylization method adopted in Pitch Line is based on the assumption that intonational tunes can be regarded as *strings of events* (pitch accents, boundary tones) associated with the segmental structure of the utterance. The events are modeled as rising, falling or rising-falling pitch movements. They are delimited by target points in the contour (F0 minima and maxima) which define their start, peak and end; some of the targets are effectively corresponding to phonological tones (H, L). The parts of the F0 contour corresponding to the events are approximated by the functions given below (γ denotes the degree of curvature):

$$0 < x < 1 \qquad y = x^{\gamma},$$
$$1 < x < 2 \qquad y = 2 - (2 - x)^{\gamma}.$$

The stretches of contour between subsequent events are called *connections* and are approximated by straight lines. In Pitch Line the approximation is carried out semi-automatically: the choice of the approximation function, i.e. R-rising, F-falling, or C-connection (cf. Taylor, 2000) and the alignment of the function with the segmental string depend on the human labeler and are decided upon by clicking in the appropriate location on the approximation panel. It is assumed that the start and end of the approximation functions have to be aligned with some segmental landmark located on the pre-accented, accented or post-accented syllable. During the approximation, the normalized mean square error can be controlled: it is displayed on the approximation panel. Similarly to other stylization methods (e.g. Momel or PaIntE), the Pitch Line represents intonation contours on a fundamental frequency scale. F0 provides useful information about the acoustic properties of the speech signal, but it is not the most accurate representation of the intonation contour as it is perceived by human listeners. Therefore, in the future it is planned to apply a more perception-oriented scale in semitones (ST) and to compare the effect of different representations of intonation on the effectiveness of prosody training.

At the output, the Pitch Line provides a file containing the values of the stylized F0 curve and another file with parameters describing the events: slope (describing the steepness of the F0 curve), Fp (F0 value at the point of the alignment of the approximation function), amplitude of the pitch movement and shape coefficient of the curve.

Figure 5 illustrates the editing window of Pitch Line. The upper panel contains the waveform; the mid-panel shows SAMPA transcription of the utterance: *ocenił w mig sytuację* (Eng. 'he judged the situation in a flash'). The bottom panel presents the original F0 contour (dotted line), the stylized contour (solid line), approximation functions (R, F, C) used for stylization of the intonation events and NMSE. The vertical lines show approximate phoneme boundaries.

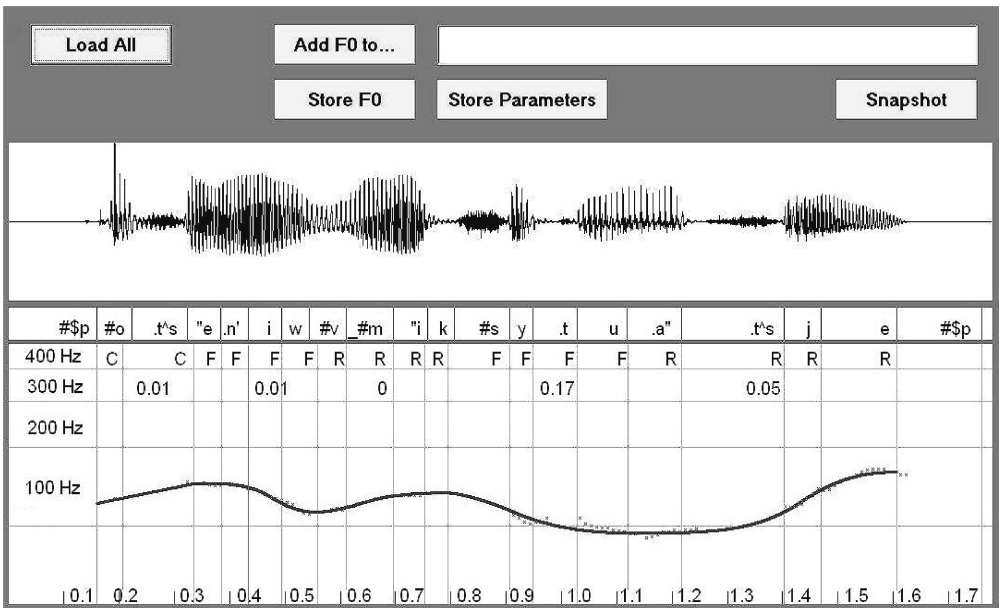

Fig. 5. The editing window and stylization of an intonation contour with the Pitch Line program.

The potential application of Pitch Line to prosody training in AzAR 3.0 could be useful for several reasons. First of all, the stylized contours represent the macroprosodic component of F0 curves which carries linguistically relevant information, but is perceptually equivalent to raw F0 contours consisting additionally of a microprosodic component caused by the nature of the individual phonematic segments of the utterance. The perceptual equivalence of the original and Pitch Line-stylized pitch contours was shown in a perception study (for details see: Demenko, Wagner, 2007): the general impression of the listeners was that the phrases resynthesized with the stylized F0 contours sounded very natural. Apart from that, a quantitative evaluation was carried out – it consisted in measuring the NMSE value between the original and stylized F0 contours of 1000 phrases from a novel passage read by two speakers (male and female). The average NMSE value for the two speakers was 0.003, which indicates that the proposed method provides an accurate approximation of F0 contours.

Secondly, at the output the program provides quantitative information which can be used to instruct the learner how to improve his or her prosodic realizations (e.g. "try to move the peak of the accent higher"). From this quantitative information, a more abstract qualitative representation can be derived and used in the evaluation of the learner's realization (e.g. realization of a low tone instead of a high one is considered to be a more severe error than realization of a different slope of a high pitch accent) and feedback generation. In the learning process it is important to provide the student with a linguistically relevant representation and Pitch Line could be used for that purpose.

5. Disscussion

The tutoring system AzAR 3.0 developed in the scope of the Euronounce project was designed according to specific pedagogical guidelines. They identify the input, output and feedback as the essential factors for pronunciation improvement. The AzAR3.0 functionality provides several audio-visual modes of user feedback, e.g. showing animated articulatory organs to correct wrong movements of tongue, lips, etc. or playing back reference utterances, but the core function is marking mispronounced phones within the spoken utterance using a colored scale from red ("bad") to green ("good"). The marking of mispronounced parts of the user's utterances is based on different phonetic-phonologic and prosodic distance measures – identifying typical cross-lingual influences of the source language on the target language, such as:

- Confusion of specific phoneme classes;
- Wrong phoneme duration;
- Articulation mistakes e.g. voicing of voiceless phonemes.

Fig. 6. The AzAR 2 demonstrator.

AzAR 3.0 is an "extended" version of the system developed in previous projects – it adopts speech processing components (such as HMM-based ASR) of AzAR 2.0. AzAR 3.0 includes improved audio and visual algorithms (like phoneme recognition or region of interest recognition) and quality assessment methods, developed on the basis of the analyses of large multilingual speech databases and offers training of prosodic features.

The AzAR2.0 system (Fig. 6) was tested and optimized for Russians learning German and runs on PC and PDA. Preliminary qualitative evaluations were conducted to gauge the effectiveness of using AZAR2.0 system to teach Polish speakers German pronunciation. Fifteen native speakers of Polish who were intermediate learners of German, were asked to subjectively evaluate the effectiveness

of a 2-hour pronunciation training using the AzAR2.0. Thirteen subjects considered it a good tool to learn pronunciation individually and 2 users expressed willingness to use it in consultation with a teacher. ASR and prosodic module for Polish language were subjected to detailed testing. The present results concerning ASR for educational needs should be regarded as a preliminary verification of the development of acoustic models for Polish. Although they are promising, further experiments are indispensable to improve the obtained acoustic models, especially for accented syllables and to provide an outcome that would be practically useful in the designed speech recognition system.

As far as the use of the Pitch Line software in AZAR3.0 is concerned, the work is in progress to develop automatic pitch target detection so that a fully automatic stylization is possible. However, for multilingual implementation of prosody module in AZAR3.0 and in order to address the multiple levels of communicative competence such as *grammatical*, *attitudinal*, *discourse*, or *sociolinguistic*, it is necessary to distinguish the intonational features with regard to four aspects of pitch change (T'Hart *et al.*, 1990): its direction, range, speed and alignment with a particular syllable in the utterance and to link them to tonal categories which constitute a higher-level representation of the utterance's intonation.

In order to further optimize AzAR3.0 for both the pronunciation and prosody training, the software will be tested in real learning environment.

Acknowledgments

This project has been funded with support from the European Commission within the Lifelong Learning Programme (project 135379-LLP-1-2007-1-DE-KA2-KA2MP). This publication reflects the views of the authors only, and the Commission cannot be held responsible for any use which may be made of the information contained therein. The project homepage is located at:

http://www.euronounce.net.

This research was partially supported by the Polish Ministry of Scientific Research and Information Technology, projects Nos. R00 035 02 and OR00006707 http://www.man.poznan.pl/online/pl/projekty/105/Laboratorium_Jezyka_i_Mowy.html.

References

1. Abberton E., Fourcin A.J. (1975), *Visual feedback and the acquisition of intonation*, [in:] *Foundations of Language Development*, Lenneberg E.H. and Lenneberg E. [Eds.], pp. 157–165, Academic Press, New York.
2. Anderson S., Kewley–Port D. (1995), *Evaluation of speech recognizers for speech training applications*, IEEE Proceedings on speech and audio processing, **3**, (4), 229–241.
3. Bongaerts T. (1999), *Ultimate attainment in L2 pronunciation: The case of very advanced late learners*, [in:] *The Critical period Hypothesis and Second language Acquisition*, Birdsong D. [Ed.], Mahwah, NJ: Lawrence Erlbaum.

4. de Bot K., Mailfert K. (1982), *The teaching of intonation: Fundamental research and classroom applications*, TESOL Quarterly, **16**, 71–77.

5. Bouselmi G., Fohr D., Illina I. (2007), *Combined Acoustic and Pronunciation Modelling for Non-Native Speech Recognition*, Proceedings of INTERSPEECH, pp. 1449–1452, Antwerp.

6. Chun D.M. (1998), *Signal analysis software for teaching discourse intonation*, Language Learning & Technology, **2**, (1), 61–77.

7. Cylwik N., Demenko G., Jokisch O., Jäckel R., Rusko M., Hoffmann R., Ronzhin A., Hirschfeld D., Koloska U., Hanisch L. (2008), *The use of CALL in acquiring foreign language pronunciation and prosody – general specifications for Euronounce Project*, Proceedings of Speech Analysis, Synthesis and Recognition (SASR), Piechowice.

8. Cylwik N., Wagner A., Demenko G. (2009), *The EURONOUNCE corpus of non-native Polish for ASR-based Pronunciation Tutoring System*, Proceedings of SLaTE Workshop on Speech and Language Technology in Education, Wroxall Abbey Estate, Warwickshire.

9. Dalby J., Kewley–Port D. (1999), *Explicit Pronunciation Training Using Automatic Speech Recognition Technology*, Calico Journal, **16**, (3), 425–445.

10. Demenko G., Wagner A. (2007), *Prosody annotation for unit selection text-to-speech synthesis*, Archives of Acoustics, **32**, (1), 25–40.

11. Demenko G., Wypych M., Baranowska E. (2003), *Implementation of Grapheme-To-Phoneme Rules and Extended SAMPA Alphabet In Polish*, Speech and Language Technology, **7**, 79–95.

12. Engwall O., Wik P., Beskow J., Granström G. (2004), *Design strategies for a virtual language tutor*, Proceedings of 8th ICSLP, pp. 1693–1696, Jeju Island.

13. Eskenazi M., Hansma S. (1998), *The Fluency pronunciation trainer*, Proceedings of Speech Technology in Language Learning, pp. 77–81, Marholmen.

14. Eskenazi M. (1999), *Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype*, Language Learning & Technology, **2**, 2, 62–76.

15. Flege J.E. (1995), *Second-language speech learning: Findings and problems*, [in:] *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, Strange W. [Ed.], pp. 233–273, Timonium, MD: York Press.

16. Goronzy S. (2002), *Robust Adaptation to Non-native Accents in Automatic Speech Recognition*, Springer Verlag.

17. t'Hart J., Collier R., Cohen A. (1990), *A Perceptual Study of Intonation*, Cambridge University Press, Cambridge.

18. Jokisch O., Koloska U., Hirschfeld D., Hoffmann R. (2005), *Pronunciation learning and foreign accent reduction by an audiovisual feedback system*, Proceedings of 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII), pp. 419–425, Beijing.

19. Kommissarchik J., Kommissarchik E. (2000), *Better Accent Tutor – Analysis and visualization of speech Prosody*, Proceedings of InSTIL 2000, pp. 86–89, Dundee.

20. Krashen S.D., Terrell T.D. (1983), *The Natural Approach: Language Acquisition in the Classroom*, Pergamon Press, Oxford.

21. Lyster R. (1998), *Negotiation of Form, Recasts, and Explicit Correction in relation to error types and learner repair in immersion classrooms*, Language Learning, **48**, 183-218.

22. Morgan J. (2004), *Making a Speech Recognizer Tolerate Non-Native Speech Through Gaussian Mixture Merging*, Proceedings of InSTIL/ICALL 2004, pp. 213–216, Venice.

23. Neri A., Cucchiarini C., Strick H. (2002a), *Feedback in Computer Assisted Pronunciation Training: When technology meets pedagogy*, Proceedings of 10th Int. CALL Conference on "CALL professionals and the future of CALL research", pp. 179–188, Antwerp.

24. Neri A., Cucchiarini C., Strick H., Boves L. (2002b), *The Pedagogy-Technology Interface in Computer Assisted Pronunciation Training*, Computer Assisted Language Learning, **15**, 5, 441–467.

25. Taylor P. (2000), *Analysis and synthesis of intonation using the tilt model*, J. Acoust. Soc. Am., **107**, 3, 1697–1714.

26. Teixeira C., Franco H., Shriberg E., Precoda K., Sönmez K. (2000), *Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners*, Proceedings of 6th ICSLP, pp. 187–190, Beijing.

27. Scovel T. (1988), *A Time to Speak. A Psycholinguistic Inquiry into the Critical Period for Human Speech*, Newbury House Publishers, Cambridge.

28. Young S., Odell J. , Ollason D., Valtchev V., Woodland P. (1997), *The HTK Book* (for HTK Version 2.1), 1997.