ISBN 978-83-60810-22-4
ISSN 1896-7094

Applying Speech and Language Technology to Foreign Language Education

Grażyna Demenko, Natalia Cylwik, Agnieszka Wagner
Adam Mickiewicz University, Institute of Linguistics, Department of Phonetics
Email: {lin, nataliac, wagner}@amu.edu.pl

***Abstract*—In recent years modern techniques involving speech processing have been gaining increasing interest among researchers and companies involved in the integration of new technologies into second language (L2) tutoring systems. At the same time, pronunciation and prosody have finally gained due attention among L2 teachers and learners. The paper describes technical and linguistic specifications for the EURONOUNCE project whose aim is to create software which will integrate non-native speech analysis and recognition with a primary goal of detecting L2 learners' pronunciation and prosodic errors and offering multimodal feedback. The software is aimed at specific language pairs, namely Polish, Russian, Czech and Slovak learners of German and vice versa. Beside information concerning the collection, structure and annotation of the multilingual speech corpora the article outlines the feedback system as well as the Pitch Line program which can be implemented in the prosody training module of the Euronounce tutoring system.**

I. Introduction

DUE TO increasing technological advance the world has been facing in recent decades there have been more and more attempts at integrating technology into traditionally non-technological areas such as education, which allowed for the development of Computer-Assisted Language Learning (CALL) systems making use of modern techniques, i.e. visual and audio aids to teach foreign languages. Along with the development of Speech and Language Technology the systems became more and more advanced applying speech processing, mainly speech analysis and recognition to the teaching software, which resulted particularly useful in systems aimed at teaching pronunciation, so called CAPT (Computer-assisted Pronunciation Training) systems. The advantages are indisputable: possibility to learn at home or in other places and at the learners' own pace, access to additional learning material like recordings, animations, visualizations, pictures, possibility to store evidence of the progress, and perhaps the most important of all, elimination of stress resulting form the fact that we are being heard by other classmates. Although yet a decade ago the reliability of such systems was repeatedly questioned [1], in the last years there have been a number of reports on large effectiveness of pronunciation software based on automatic speech recognition (ASR) and automatic error detection [2]-[5]. Unfortunately, the existing self-study programs rarely include prosodic features, which is the result of a long-standing negligence prosody has suffered in language teaching due to poor knowledge of its functions, forms and acquisition. This trend has recently changed and we have been witnessing a growing interest in teaching suprasegmentals for numerous reasons: due to new advances in the theory of intonation, the growing accessibility of acoustic signal analysis, processing and interpretation and due to a shift of focus to pragmatics, discourse and conversation analysis and to communicative function of prosody, which replaced traditional approach giving priority to smaller units: sounds and words, and linguistic forms of prosody. The Euronounce project follows this new tendency as it aims at creating an intelligent language tutoring system with multimodal feedback functions which would integrate both segmental and suprasegmental aspects of speech. This paper provides technical and linguistic specifications for the Euronounce project with a particular focus on the application of computational techniques based on speech and language processing to second language learning.

II. Challenges in Computer-assisted Language Learning

A. Audio and visual training

The advantages of multimedia tools in education are multiple. The use of tools for audiovisual feedback to detect deviations from standard articulation in the target language have shown especially high effectiveness in PC-based pronunciation learning systems. Prosody visualization seems to be more complex. However, speech analysis has been used for teaching L2 (second language) intonational patterns since 1970s e.g. [6], [7]. 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

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 only of the authors, 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 .

product from Kay Elemetrics called Visi-Pitch which has been available for a number of years for DOS-based personal computers. With Visi-Pitch, students are able to see both the model speaker's and their own intonational curve simultaneously.

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

- Technical aspects:
  - Weak speech signals;
  - 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 evaluate results;
  - Examples and exercises consist of word and sentence-level intonation;
  - Lack of integration of prosodic features as tone, duration, loudness;
  - Lack of voice quality analysis - even if the 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.

In order to develop an effective audio and visual training that improves learners' perception and production of intonation and rhythm we need to better understand the relationship between perception and production at the level of segmentals and especially suprasegmentals. There are generally three types of prosodic phenomena which are not understood well enough to develop effective prosody training tools: 1) those which divide the speech into *chunks* or *units* , 2) those which lend *prominence*, and 3) paralinguistic/nonlinguistic phenomena [8].

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 from learners to be able to monitor and evaluate themselves critically. Apart from visual display, no further feedback is provided and there is a lack of objective assessment. Another question concerns the long-term effects of any of the brief training sessions.

B. State of the art

Innovative PC-based pronunciation teaching systems like Pronunciation Power, American Sounds, Phonics Tutor and Eyespeak offer multimodal feedback which includes (verify [9]) analysis and recognition of the input speech by the learner who can record his/her own utterance which is then displayed as spectrum enabling acoustic and visual comparison between the learner’s and the reference voice’s utterances. It is common to include animations showing movements of articulators: jaw, tongue and lips while producing given sounds. Some systems (e.g., Fonix iSpeak 3.0, ProNunciation) make use of synthesized speech or TTS solutions [10]. During the last decade, speech recognition technology was implemented into innovative interactive systems like Istra and Pronto [11] and in the European research project Interactive Spoken Language Education ISLE [2], [12].

Various speech databases which contain speech from non-native speakers have been created and some researchers have investigated the possibilities to improve the performance of speech recognizers on non-native utterances and tried various probabilistic models to produce pronunciation scores from the phonetic alignments generated by HMM-based acoustic models. In [13] the task of predicting the degree of ‘nativeness’ of the learner utterances was addressed. To achieve the best results not only the segmental features of the speech signal were used, but also prosody. For example, rate of speech appears to be a very effective predictor of linguistic competence; two other important determinants of reading fluency are the rate at which the speakers articulate the sounds and the number and length of pauses.

Unfortunately, current speech recognition systems are poor at handling information contained in the speaker’s prosody. Therefore, there is still room for the improvement of the ability of speech recognition systems to recognize accented or mispronounced speech and to provide meaningful evaluation of the pronunciation quality.

In the FLUENCY project [14] speech recognizer was used to detect foreign speakers’ pronunciation errors for L2 training. The research also involved prosody and correlation between pronunciation and prosody errors was investigated. However, neither the placement of the intonation errors, nor suggestions on how to improve intonation were provided, leaving the comparison to the users. Well-known software, Tell Me More of Auralog, improved the detection and the feedback for pronunciation practice by pointing out erroneous phonemes and showing a 3D animation to visualize the ‘standard’ articulation. However, its technology for suprasegmentals (concerning only intonation aspects) is very limited.

III. Euronounce Project

A. Towards optimal technology for L2 language learning

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. In accordance with EU’s policy of promoting less widely spoken languages the project focuses on Slavonic-German language pairs: Polish-German (PL/DE), Slovak-German (SK/DE), Russian-German (RU/DE) and Czech-German (CZ/DE). 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 software AzAR (German acronym for *Automat for Accent Reduction*) aimed at teaching Russians German pronunciation was created [5]. Following the baseline developed in these projects the Euronounce project aims at creating software for pairs L1 DE – L2 RU, CZ, SK, PL and L1 RU, CZ, SK, PL – L2 DE beside segmental adding also suprasegmental exercises.

B. Speech databases, speaker selection and text corpus

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 ([15], [16]). 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 [17]. 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 [5]. For that reason in the development of the Euronounce software three speech databases are created for each language pair:

- Reference database - target language speech by target language native speakers.
- Non-native speech database - target language speech by non-native speakers.
- Source-language accent database - source language speech by source language native speakers.

The reference database consists of the whole 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, and the latter 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 pronunciation and prosody follow the rules of the standard form of the language being taught. The reference speakers have been recorded in sound-studio conditions of high standard with practically no perceivable reverberation or background noise.

The second database consists of recordings of non-native speech produced by 18 speakers per language pair. 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 [19], i.e. levels A1-A2, B1-B2, C1-C2. The database was balanced with respect to sex and proficiency of the speakers (i.e., 6 speakers per proficiency level were recorded). The proficiency ratings were based partly on the information provided by the students in a questionnaire including among other things a self-judgment of their language skills. This information was verified by an expert who listened to a sample recording of student's speech produced during a short spontaneous speech production test.

For the purpose of recording the non-native speech database a special text corpus was created. Generally, six tests including different types of texts were proposed. The *accent* and *dialectological* tests serve the purpose of investigating L1 interferences and consist of sentences created by experienced teachers of phonetics. The dialectological test includes also words/sentences for which alternative pronunciations exist. The *Phondat test* consists of phonetically rich and balanced sentences designed for the purpose of ASR training (recognition of non-standard pronunciation), analysis and assessment of pronunciation errors. Part of the text corpus contains also texts for the *fluent reading test* which aims at the analysis of the realization of segmental, but also suprasegmental and discourse features of non-native speech.

The main objective of *spontaneous speech test* is to investigate different aspects of spontaneous speech produced by non-natives.

The last and probably most innovative part of the corpus is *prosodic test*. The purpose of the test is to investigate the realization of prosodic/intonational features by advanced L2 learners and L1 interferences in the domain of prosody. The text material for the prosodic test was created according to the same criteria as the exercises for prosody training described in the next section. Together with spontaneous speech test the prosodic test is addressed to advanced students only.

Part of the text material (accent, dialectological, phondat and prosody test) has been read by both non-native and native speakers. The resulting speech material serves as a reference for the assessment of non-natives' pronunciation and prosody.

The recordings have been conducted using the WiGE rec software, in a studio with low noise and reverberation, and with a two-channel input i.e., a close-talk and condenser microphone. Basic quality requirements are: sampling frequency 44,1 kHz, minimal resolution 16 bit, minimal SNR of 35 dB.

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. Basic requirements for the recordings are the same as in case of the non-native speech database.

C. Annotation of speech datatbases

The whole speech material has been segmented and phonetically transcribed using force alignment. The subset of the B speech database (including *accent, dialectological* and *phondat* sentences) has been manually verified. Verification includes: adjustment of segment boundaries, marking of noise, disfluency and pauses, checking the transcription and annotation of automatically inserted primary and secondary stress markers, marking deviations from the canonical pronunciation (insertions, deletions and substitutions of phonemes, accents and phrase boundaries). Fig. 1 below shows an example of annotated utterance from the German-Polish non-native speech corpus (i.e. learners of Polish with L1 German). From top to bottom, the waveform, the spectrogram and annotation panel are depicted. In this example the speaker made four substitutions: he pronounced the initial /z/ as a voiceless /s/, substituted /o/ for /u/, and reduced twice the word-final /e/ vowel (marked as /e-@/). The speaker found it probably difficult to realize the cluster /z/ +/r/ and inserted the vowel /y/ in between the two consonants.

D. Suprasegmental structure

In accordance with the current emphasis on communicative and sociocultural competence, more attention should be paid to discourse-level communication and to cross-cultural differences in pitch. As natural discourse exhibits anything but "default" intonational patterns, L2 learners must be made aware on of how stress, emphasis, contrast, and illocutionary speech are expressed in the L2. In order to meet these goals, the version of AzAR developed within the

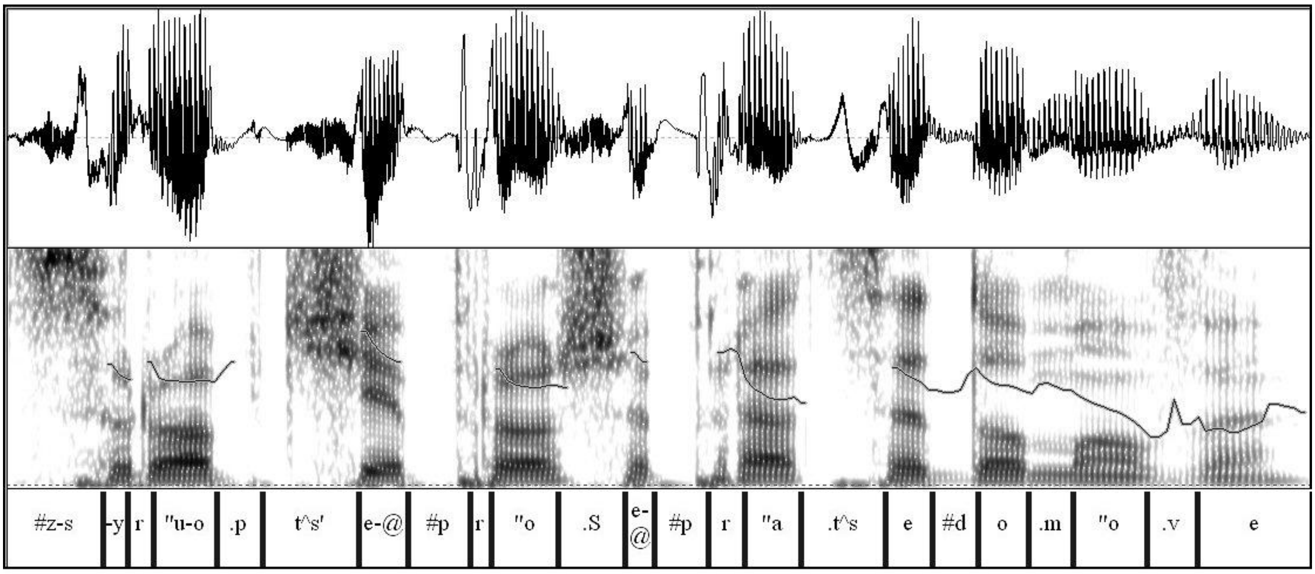


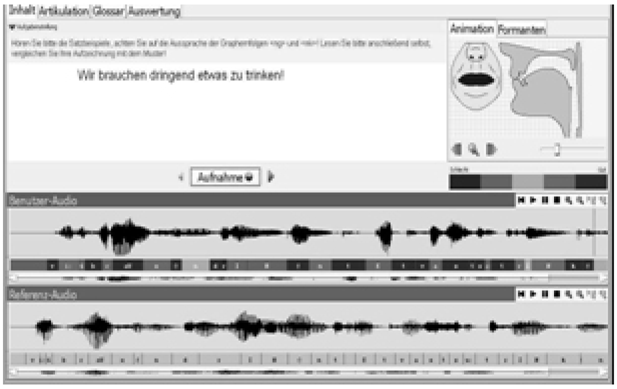
Fig 1. Example of annotated utterance from the German-Polish speech corpus (male speaker). The text was: Zróbcie proszę prace domowe (Eng. Please do you homework).

framework of the Euronounce project contains module for prosody training. The exercises are devised in order to test and practice prosody in smaller and larger syntactic units. In isolated words suprasegmental identification is devoted mainly to the perception and production of regular and irregular lexical stress and feet structure as well as types of nuclear accents, duration, intensity, identification of mono-, di-, tri-, four-syllable words, prosodical word, enclitics, proclitics linking. At the level of simple and complex sentences exercises consist in production and recognition of different types of sentences, i.e. declaratives, commands, wh-questions, etc. on the basis of their suprasegmental features. Also building the awareness of relationship between focus and meaning needs close attention. Identification and production of emphatic stress, relating focus with meaning and performing communicative functions with focus should be practiced e.g. showing emotions, disagreement, calling attention to new information. Special awareness is also given to contrastive pitch patterns conveying various meanings: fall (finality, authority), rise (unfinished, insinuating, tentative), level (unfinished, unresponsive), fall-rise (reservation, contrast, calling), rise-fall (insistence, surprise, irony).

IV. Feedback System

A. Segmental structure

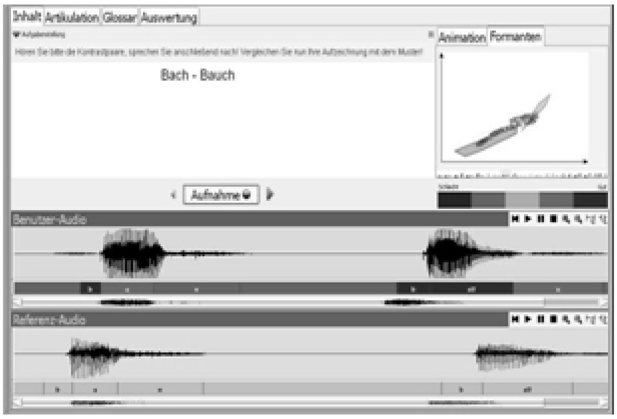
Lack of proper (or any) feedback is often named as the most serious flaw in educational software ([11], [16]). A good software should not only assess if the correctness of pronunciation but also instruct on how to improve it, show where exactly the error has been made, e.g. which phone has been produced erroneously and offer feedback that is easy to interpret. To answer these needs the AzAR 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 software uses HMM-based speech recognition and speech signal analysis on the learner's input which makes a visual and aural comparison of user's own performance with that of the reference voice possible. Most importantly, the system also performs an automatic error detection on the phonemic level. 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 model voice as well as see the speech signal for a particular utterance, record and listen to his/her own utterance and see the speech signal for his/her own utterance and finally get feedback on his/her own pronunciation. Additional visual mode includes animated visualization of the vocal tract (lips area and articulators movements) and a formants graph for particular phones. A typical AzAR template for an exemplary phrase is showed in Fig. 2 and 3. From top to bottom the panel containing the text to be produced is shown together with the formants/articulation graph, the spectrograms of user's and reference speaker's utterances: below them the transcription and segmentation panel can be seen. Pronunciation quality is visualized with colors (here in a greyscale). The segments that appeared problematic in the second 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.



Hören Sie bitte die Satzbeispiele, achten Sie auf die Aussprache der Graphemfolgen »ng« und »nk«! Lesen Sie bitte anschließend selbst, vergleichen Sie Ihre Aufzeichnung mit dem Muster!

Wir brauchen dringend etwas zu trinken!

Fig 2. The editing window and stylization of an intonation contour with the Pitch Line program.



Hören Sie bitte die Kontrastpaare, sprechen Sie anschließend nach! Vergleichen Sie nun Ihre Aufzeichnung mit dem Muster!

Bach - Bauch

Fig 3. AzAR template for pronunciation assessment of a minimal pair Bach – Bauch (DE).

Positive results of this kind of audio-visual feedback have been reported especially in the context of prosody teaching ([15], [20]). For pronunciation training [21] also traditional instruction is being recommended since visuals can be too difficult for the user to interpret and listening drill is not enough when one keeps in mind that L2 learner tends to associate foreign sounds with more familiar L1 sounds. Therefore, beside audio-visual feedback, AzAR software includes also text tutorial on articulatory and basic acoustic phonetics with glossary, phonemes description and classification, anatomic information, etc.

B. *Suprasegmental structure*

In order to provide an effective feedback in prosody training software should visualize the "relevant" intonation pattern of a given utterance as realized by L2 student and native speaker. Apart from that it should draw attention to acoustic features involved in the realization of intonation [22]. 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. Pitch Line program [23] designed for approximation and parameterization of intonation contours answers these needs and could probably be successfully implemented in the AzAR environment.

The method behind the Pitch Line stylization 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). At the moment, identification of pitch targets' position is carried out manually. The parts of F0 contour corresponding to the events are approximated with functions described as follows:

$$0<x<1 \quad y=x^{\gamma}$$

$$1<x<2 \quad y=2-(2-x)^{\gamma}$$

The stretches of contour between subsequent events are called connections and are approximated with 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. [24]) 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.

At the output the program [1] provides a file containing the values of the stylized F0 curve (which can be used for pitch resynthesis in Praat) 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.

Fig. 4 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 (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.

The usefulness of the approach adopted in Pitch Line was tested on a speech corpus including recordings of two speakers (male and female) reading a novel passage, altogether 1000 phrases [24]. The stylization accuracy was evaluated objectively by measuring the NMSE value between original and stylized F0 contours and subjectively in a perception study. The average NMSE value for the two speakers is 0.003, which indicates that the proposed method provides an accurate approximation of F0 contours. As regards the perception test results the general impression of the listeners was that the phrases resynthesized with the

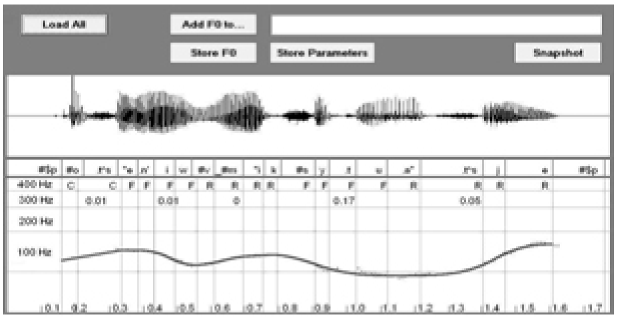


Fig 4. AzAR template for pronunciation assessment of a minimal pair Bach – Bauch (DE).

stylized F0 contours sounded very natural. Informal tests were also carried out on a subset of German utterances from the Euronounce speech database and the first results are promising.

However, for the assessment and training of realization of specific intonation patterns a higher-level representation of intonation could be more useful. Prosody transcription systems such as ToBI [25], INTSINT [26] or the one used in the Polish module of BOSS unit selection TTS system [27] use discrete categories to describe not only perceptually, but also linguistically distinct utterances. On the contrary, phonetic descriptions such as Tilt or Pitch Line use continuous parameters which enable encoding of perceptual differences between intonation contours, but not necessarily the linguistic ones. Since the purpose of prosody training is to help students acquire foreign language prosody, a linguistic representation should be used to analyze and assess the realization of prosodic features, because errors in this domain are most of all categorical. Of course such representation can be mapped onto/derived from a lower-level description in terms of acoustic-phonetic parameters [28]. At the moment frameworks of prosody description and analysis (the higher-level categorical ones and the lower-level phonetic–acoustic ones) are being tested with respect to their application in the prosody training and assessment module of the AzAR system.

As indicated in the discussion above in the latter both linguistic and paralinguistic aspects should be considered.

V. Conclusions

This paper presented specifications and assumptions for the Euronounce project which aims at creating an intelligent system with multimodal feedback functions for acquiring pronunciation and prosody of German, Polish, Slovak, Czech and Russian as L2. The article outlines basic theoretical foundations as well as the core technology established in the preceding projects based on a German-Russian language pair. This baseline coupled with new cross-lingual databases are to help improve the visualization and quality assessment methods and to allow including prosodic factor in the final software.

References

[1] M. Liu, Z. Moore, L. Graham, and S. Lee (2002). “A Look at the Research on Computer-Based Technology Use in Second Language Learning: A Review of the Literature from 1990–2000”. *Journal of Research on Technology in Education*, 34(3):250-273, 2002.

[2] E. Atwell, P. Howarth and C. Souter (2003). "The ISLE Corpus: Italian and German Spoken Learners' English", *ICAME Journal*, 17:5-18, 2003.

[3] B. Mak, M. Siu, M. Ng, Y. Tam, Y. Chan, K. Chan, K. Leung, S. Ho, J. Wongand, J. Lo (2003). “PLASER: Pronunciation Learning via Automatic Speech Recognition”. *Proc. HLT-NAACL Workshop on Building Educational Applications using Natural Language Processing 2003*.

[4] H. Franco, V. Abrash , K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossierand, F. Cesari (2000). “The SRI EduSpeak(TM) System: Recognition and Pronunciation Scoring for Language Learning”. *Proc. InSTIL 2000* Dundee, Scotland, pp. 123-128.

[5] O. Jokisch, U. Koloska, D. Hirschfeld and R. Hoffmann, (2005). “Pronunciation learning and foreign accent reduction by an audiovisual feedback system”. *Proc. 1 st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII)* , Beijing, 2005, pp. 419-425.

[6] E. Abberton & A. J. Fourcin (975). “Visual feedback and the acquisition of intonation”, In E. H. Lenneberg & E. Lenneberg (Eds.), *Foundations of Language Development* (New York: Academic Press, 1975), pp. 157-165.

[7] K. de Bot & K. Mailfert (1982). “The teaching of intonation: Fundamental research and classroom applications”, *TESOL Quarterly, 16* , 1982, pp. 71-77.

[8] A. Cruttenden (1997). *Intonation.* Cambridge: Cambridge University Press, 1997.

[9] Learning Village. Educational Software Review, Retrieved on 15th July 2008 from http://www.learningvillage.com/html/guide.html

[10] J. Burston (2000). The CALICO Software Review. Computer Assisted Language Instruction Consortium homepage. Retrieved on 15th July 2008 from http://calico.org/CALICO Review/

[11] J. Dalby & D. Kewly-Port (1999). “Explicit Pronunciation Training Using Automatic Speech Technology”, *CALICO Journal, 16* (3), 1999, pp. 425-445.

[12] Interactive Spoken Language Education (ISLE), project homepage of Hamburg University. Retrieved on 15th July 2008 from http://nats-www.informatik.uni-hamburg.de/ ~_\isle/

[13] C. Teixeira, H. Franco, E. Shriberg, K.Precoda, K. Sönmez (2000). “Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners”, *Proc. 6th ICSLP* , Beijing, 2000, pp. 187-190

[14] M. Eskenazi , S. Hansma (1998). “The Fluency pronunciation trainer”, *Proc. Speech Technology in Language Learning* , Marholmen, 1998. Retrieved on 15th July 2008 from http://www.cs.cmu. edu/~max/mainpage_files/Esk-Hans-98.pdf

[15] M. Eskenazi (1999). “Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype”, *Language Learning & Technology, 2* (2), 1999, pp. 62-76.

[16] O. Engwall, P. Wik, J. Beskow & G. Granström (2004). |”Design strategies for a virtual language tutor”, *Proc. 8th ICSLP,* Jeju Island, 2004, pp. 1693-1696.

[17] J. C. Wells (2000). “Overcoming phonetic interference”, *English Phonetics, Journal of the English Phonetic Society of Japan, 3,* 2000, 9-21.

[18] Common European Framework of Reference for Languages. Retrieved on 21st August 2008 from http://www.coe.int/t/dg4/linguistic/Illustrations_EN.asp

[19] D.M. Chun (1998). “Signal analysis software for teaching discourse intonation”, *Language Learning & Technology* , *2* (1), 1998, 61-77.

[20] A. Neri, C. Cucchiarinim and H. Strick (2002). “Feedback in Computer Assisted Pronunciation Training: When technology meets pedagogy”, *Pr oc. 10th Int. CALL Conference on "CALL professionals and the future of CALL research",* Antwerp, 2002, pp. 179-188.

[21] J. t'Hart, R. Collier & A. Cohen, *A Perceptual Study of Intonation.* Cambridge: Cambridge University Press, 1990.

[22] G. Demenko & A. Wagner (2007). “Prosody annotation for unit selection text-to-speech synthesis”, *Archives of acoustics, 32* (1), 2007, pp. 25-40.

[23] P. Taylor (2000). “Analysis and synthesis of intonation using the tilt model”. *J. Acoust. Soc. Am 107* (3), 2000, pp. 1697-1714.

[24] A. Wagner (2006). “A comprehensive model of intonation for application in speech synthesis”, *Proc. 8 th International PhD Workshop OWD* , Wisła, Poland, 2006, pp. 91-96.

[25] Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P.J., Pierrehumbert J. & Hirschberg J. (1992). "ToBI: A standard for labeling English prosody" . *Proceedings of ICSLP '92* , pp. 867-870.

[26] Hirst, D.J., Di Cristo, A. & Espesser, R. (2000). "Levels of representation and levels of analysis for intonation" . In M. Horne [ed] *Prosody : Theory and Experiment* . Kluwer: Dordrecht.

[27] Demenko G. & Wagner A. (2007). "Prosody annotation for unit selection text-to-speech synthesis" . *Archives of acoustics* , 32(1):.25-40.

[28] Wagner A. (2008). "A utomatic Labeling of Prosody". *Proc. ISCA Tutorial and Research Workshop on Experimental Linguistics* , 25-27 August 2008, Athens, Greece.