*The Third International Conference on e-Learning (eLearning-2012), 27-28 September 2012, Belgrade, Serbia*

# SPEECH TECHNOLOGIES FOR PRONUNCIATION AND PROSODY TRAINING OF MACEDONIAN LANGUAGE

IVAN KRALJEVSKI
TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany, ivan.kraljevski@tu-dresden.de

OLIVER JOKISCH
TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany, oliver.jokisch@tu-dresden.de

SUZANA LOSKOVSKA
Faculty of Computer Science and Engineering, UKIM, Skopje, R. Macedonia, suzana.loshkovska@finki.ukim.mk

RÜDIGER HOFFMANN
TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany, ruediger.hoffmann@tu-dresden.de

***Abstract:*** *In this paper, we introduce a concept of Computer assisted pronunciation training (CAPT) e-learning platform on Macedonian language where the issues for speech technologies (Automatic Speech recognition and synthesis) on Macedonian language are also described. Review of the speech technologies used in CAPT is given and detailed description of the proposed framework with the development steps and a number of technological challenges is presented. The presented CAPT concept will take advantages of speech technologies to provide audio-visual feedback to the language learners and provides language training on both subsegmental (pronunciation) and suprasegmental level (prosody). The specifications for the required speech databases (native speech database, non-native speech database and source-language accent database) and the speech corpora development are described. Also, general guidelines are given for preparation of the speech material and its lexical structure for prosody tests and exercises, based on the contrastive analysis research studies for Macedonian as L2.*

***Keywords:*** *Language learning, Computer-assisted pronunciation training, Speech technologies*

## 1. INTRODUCTION

It is well known, that learning foreign languages is performed at best in small interactive groups of learners where the language teacher is able to receive appropriate feedback and provide immediate correction. The trained language teacher interaction with one learner exhibits even more advantages regarding the traditional classroom instruction. However, this one-to-one approach is costly, not ever possible and mostly not feasible for larger language learner population [1]. Also, in larger groups, because of the restricted time frame, it is not always possible to provide immediate individual feedback for each language learner. This especially applies to oral skills like proper phoneme pronunciation, prosody intonation and lexical stress.

In the European countries space, the increased mobility indicates that proficiency of one or more foreign languages beside the mother language is required and encouraged. The real situation is far from desired and the problem is worldwide present, effectively limiting the intercultural communication, mobility and integration. To improve the current situation, new natural and stimulating language learning methods are required [2]. Beside this, in many countries there is an increase in interest for particular language learning, which causes the situation where the demand is bigger than the available language learning possibilities. Similar problem apply also in the area of e-inclusion where efficient language learning is one of the keys to social inclusion. This concerns disabled people as well, like deaf and hard to hearing population where the oral skills can be drastically improved by using computer-assisted language learning (CALL) systems in combination with Speech technologies. Speech technologies are already intensively used for communicative needs for disabled persons, as for example speech synthesis in screen readers and reading machines for visually impaired people and as a speech aid for non-vocal persons. Speech-enabled technology, speech analysis, automatic speech recognition (ASR) and speech synthesis implemented CALL offers new perspectives for language learning. These systems allow the possibility to address individual learners' problems, additional learning time and materials, specific feedback on individual errors etc. Speech-enabled CALL systems can also help learner to simulate realistic interaction to improve their language skills [3].

Automatic speech recognition (ASR) technology is of particular interest for developing new methodology for improving literacy, reading, oral proficiency, speaking fluency, and vocabulary. Thanks to these technologies the focus has changed from teaching writing skills, grammar and vocabulary to teaching oral skills and also

pronunciation and prosody. Acquiring L2 (second language) pronunciation and prosody is of great importance for social integration as mentioned in [4], strong foreign accent may cause integration problems which makes it particularly important in the times of global migration and the policy of integration.
The growing interest in teaching and learning foreign language pronunciation and prosody has been reflected in the development of Computer-assisted pronunciation training (CAPT) systems [5]. Common system approaches include the phonetic quality assessment, highlight problematic sections in the speech signal and usually rely on automatic speech recognition (ASR) regarding the target language L2.

To develop CAPT applications with high level of performance and usability, it is necessary to have already available and trained ASR system. But for development of new ASR system acquisition of large quantity of carefully prepared and transcribed acoustic data is required for the particular language. This requirement is especially emphasized in the case of under-resourced languages, where little or no electronic speech and language resources exist. This is the case with the languages used by smaller speaker population, but also for large speaker population languages with sparse language resources. The acquisition and transcription of speech data is very time consuming and costly process and presents one of the major limitation factors for speech applications development. Thus, availability of sophisticated CAPT systems with ASR and speech synthesis components is limited on the most wide spread worlds languages.

Beside of the established interest for learning world languages, also in accordance with EU's policy there is increased need for promoting and learning less widely spoken languages and this is case also with the Macedonian language. Given the fact that the country has the EU candidate status, it could be foreseen that there will be increased interest for learning Macedonian language by foreigners in the near future.

In Republic of Macedonia in the communities where larger population of one of the ethnic minorities exists, the usage of the minority's mother language as second official language in primary, secondary and tertiary education as well as in the public administration is guaranteed by the constitution. Therefore, it is of great importance to improve oral skills (particularly pronunciation and prosody) in using Macedonian language within these communities for better communication and social integration.

In this paper, a concept for development of CAPT e-learning platform on Macedonian language is presented, and the issues for speech technologies (ASR and synthesis) development on Macedonian language are introduced. Firstly, a review of the speech technologies used in CAPT is given. In the following section, detailed description of the proposed framework with the development steps and a number of technological challenges is presented. Next section describes the required speech databases and the speech corpora creation. Last section provides the guidelines for preparation of the speech material and its lexical structure which will be used in the practical exercises. Finally, some conclusions are drawn and challenges and opportunities for the future are considered.

## 2. SPEECH TECHNOLOGIES USED IN CAPT

Computer assisted pronunciation training (CAPT) [6] encompasses a range of tools and techniques which include the use of spectrograms, pitch contours and statistical information on acoustic features of oral production. Speech analysis techniques are used in pronunciation and prosody training, primarily for displaying the speech waveform and its spectrogram, together with the extracted pitch contour. This allows comparing the speech sequences uttered by the learner with the reference voice. Such combination of audio and visual feedback improves the perception of the target language.

Speech synthesis is not widely used in CAPT for audio feedback production because the synthetic speech sounds artificially and unnatural. Therefore, recordings of real native speakers are preferred in most current learning systems. However, Text-to-Speech (TTS) is used as a step for visual speech synthesis to convert free text input into a phoneme string, together with the corresponding phone durations and syllable boundaries [7]. In [8] the authors report an effort on training the incorrect phone models by making use of synthesized speech data using special formant speech synthesizer to filter the correctly pronounced phones into incorrect phones by modifying the formant frequencies.

Automatic Speech Recognition (ASR) systems have been already widely used in numerous CALL applications. The simplest approach to take advantage over ASR for CAPT would be to use commercially available ASR dictation packages as a tool for training language skills. In [9] the authors investigated the ASR for CALL by evaluating the performance of a standard dictation package "Dragon Naturally Speaking Preferred" to identify pronunciation errors in the L2 speech of Cantonese and Spanish learners of English, and they express doubt about the usefulness for L2 learning of specific dictation package. These conclusions do not apply to ASR in general, but the ASR dictation package tested in this study was never intended to be used for L2 learning.

Use of automatic speech recognition provides the possibility of dialog between the learners and the computer in conversational manner, but only restricted in a limited speech domain. It is clear from various research studies that speech recognition system which is created and trained on native speakers of one language performs worse at recognizing foreign-accented speech. One of the solutions is creation of non-native speech databases for developing speech recognizer specific to learners' mother tongue. In [10] the researchers conclude that speech recognition provides an optimal solution to pronunciation learning. Other studies like [11] stated that the

effectiveness of CALL systems could be improved by careful design of the language learning activities and inclusion of some form of corrective audio-visual feedback.

The main task of an ASR component in CAPT system is to analyze learner's input speech and using various probabilistic models to produce pronunciation scores from the phonetic alignments generated by HMM based acoustic models. However, this should not only assess if the pronunciation is correct or incorrect but also instruct on how to improve it, to show the placement of the intonation errors, suggests how to improve intonation [4] and offers feedback that is easy to interpret [12]. Thus, it is of great importance to develop appropriate audio-visual learning environment and provide useful and robust feedback on learner errors [13].

## 3. CAPT OF MACEDONIAN LANGUAGE

In this section, description of framework for CAPT system of Macedonian language is presented. In current CAPT systems the main focus is set on the global quality of the user's phones compared to a previously defined average acoustic model which is provided by existing HMM (Hidden Markov Models) based automatic speech recognizer trained on native speakers. The requirements for a good and intelligent CAPT system include:

- precise identification the error location and type;
- monitoring of the learner's performance, to identify specific problems and to adapt the exercises;
- provide feedback that is relevant for the error type;
- individualized feedback that indicates on what features more practice its needed;
- natural interaction with the system to practice all aspects of language learning, from articulation training to conversations.

Usually, the visual feedback uses waveforms and pitch curves to indicate prosody differences between the user and the model, and highlights the part of the speech utterance with most deviant pronunciation. The following different views could be implemented to illustrate differences in input and the reference pronunciation:

- animation of clearly visible lips and other facial features;
- midsagital animation for tongue movements presentation;
- palatal display to show regions of contact;
- speech waveforms;
- spectrograms;
- formant graphs.

The proposed framework includes HMM-based ASR and speech signal analysis on the learner's input from where after feedback production he can visually and aurally compare his own performance with that of the reference voice. Also, the proposed system includes automatic error detection on the phonemic level.
All uttered phones are marked using colour scale from red for mispronounced phones to green for those pronounced correctly. Colouring provides additional quality in the feedback system instead of showing two comparable displays, one representing the learner's utterance and one representing the referent utterance. This avoids the situation where the learner will try to produce a speech utterance that closely corresponds to that of the referent voice. Two sequences with the same content may both be very well pronounced and still have waveforms or spectrograms that are very different from each other. But, adding coloration to the wrongly pronounced phonemes gives the opportunity for the learners to focus on particular error type and location. The learner can listen to and play back the reference voice as well as see the speech signal for a particular utterance, record and listen to his own utterance and see the speech signal for his own utterance and finally get feedback on his own pronunciation.

The proposed architecture is presented on Figure 1. The structure is modular and it is separated from the content, as well as the universal linguistic tools from language specific components. This allows flexibility for further development and adaptation on new user groups, new language or new set of exercises. The Exercise manager consists of exercise examples created to practice production and perception at the phonemic and prosodic level in isolated words, simple phrases, complex phrases and continuous speech.

The ASR system could recognize spoken utterances from the learner even if there are deviations in pronunciation and intonation, and it would be able to locate at phoneme level the pronunciation errors made by the speaker. The role of the ASR module is to recognize and label the utterance using forced alignment recognition with known transcription, which is restricted to the current exercise example.

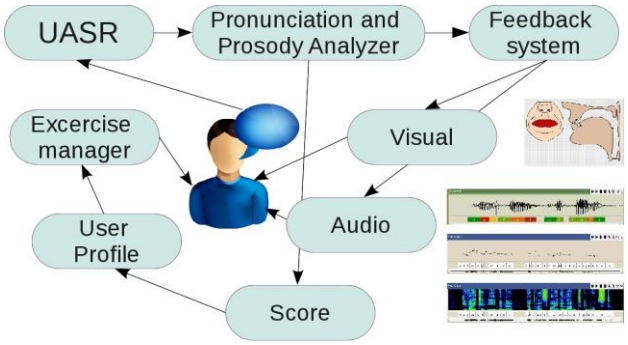


**Figure 1.** Diagram of the Macedonian-CAPT system

For the proposed framework, the UASR (Unified Automatic Speech Recognition and Synthesis) [14] system will be used for speech recognition. It is a speech dialogue system where the speech recognition process and the speech synthesis components use common databases at each processing level. During the recognition process, prosodic information is separated from the data flow and later it can be used for intonation curve presentation and as well as for synthesis, improving the naturalness of the synthesized speech.

The structure of the UASR system is based on arc-e-mission HMMs with one single Gaussian density per arc and an arbitrary topology. This structure is built iteratively during the training process by state splitting from an initial HMM model. The acoustic model structure will consists of 33 monophonic (Macedonian phoneme inventory count) HMMs plus one pause and one garbage model.

To perform the feature extraction, each frame of the recorded speech (L2) is processed using Hamming window and filtered through Mel DFT filter bank, where the log of the energy and the delta values were calculated at the each channel output. The obtained feature vectors were standardized to a mean of 0 and a standard deviation of 1 and Principal Component Analysis (PCA) is used for feature vectors transformation to reduce the dimension to 24 most significant components.

Then, the pronunciation and prosody analyzer will use the outcome of the recognition and transcription and performs assessment where the erroneously recognized phonemes and non-matched intonation curves will be pinpointed and provided as a feedback to the learner. The comparison between the spoken and referent speech utterance will be performed using the proposed UASR setup by forced alignment procedure. The outcome of the phoneme recognition will be the confidence score for each phoneme (subsegmental level) – GOP (Goodness of Pronunciation).

The problem that must be solved it the appropriate threshold calculation to decide about the quality of the pronunciation. To achieve that, the distribution pairs of the GOP scores of each phoneme will be estimated in two occasions: correctly pronounced and incorrectly pronounced phonemes. Then the value of the decision threshold (phoneme pronunciation is good or not) will be calculated by the equal error rate (EER) of both distributions.

As for the prosody (suprasegmental level), to calculate the distance between two intonation patterns several metrics can be used like: mean-absolute-frequency-deviation, the pitch target points and the use of temporary shifts in f0 patterns, but none of them performs better than the classic RMSE (root-mean-square error) or the correlation coefficient of Person R2. First, the spoken intonation curve must be aligned with the referent one using DTW (Dynamic Time Warping) and then RMSE can be calculated.

The quantitative outcomes of the pronunciation and prosody analyzer are written in the current learner's profile database and passed to the Feedback module. The Feedback module provides audio-visual information that can be presented via animated sagital and facial features, recorded or synthesised speech waveforms and audio output, spectrograms and pitch contours, in the same time providing enriched multimedia environment and experience for the pronunciation and prosody learner.

## 4. SPEECH DATABASES

Usually CAPT systems rely on recordings of native vs. non-native speakers to evaluate the signals recorded by the learners. Therefore, a database developed specifically for the verification of phonemes recordings must have both types of learners. The most important development step will be creation of native speech databases on Macedonian language. These databases will be used for acoustic model training for the proposed UASR system.

Beside the L2 (Macedonian) speech corpus, also, complex non-native speech databases are also required including spontaneous speech, continuous speech as well as simple and complex sentences designed to investigate specifically selected phenomena. Most errors result from L1 and L2 interference and consist primarily in transferring allophonic and phonotactic rules from the learner's mother tongue to the target language and replacing L2 phonemes with their most similar L1 counterparts [15]. 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 [4].

Important part of the speech corpus is the speech material for the prosodic test which purpose is to investigate the realization of prosodic/intonational features and L1 interferences on L2 in the domain of prosody. Therefore, it is desirable to perform a cross-reference mapping of linguistic features for each language, to predict the possible difficulties the learners are likely to have.

For the use in the proposed system, the following speech databases should be created for pair L1-L2:

- reference database – L2 read speech by L2 native speakers;
- non-native speech database – L2 speech by non-native speakers, reflecting typical pronunciation and prosody mistakes in L2;
- source-language accent database – L1 speech by source language native speakers for ASR training and comparative study of interferences.

The text material for the prosodic test will be created after detailed contrastive analysis based on the theoretical findings but also using modern empirical methods taking the advantage of the recorded speech databases. Special considerations must be taken, as the L2 learner's competence to perceive and produce difficult and new phonetic contrasts depends on the mother tongue. For this purpose, the analysis will be carried out in two stages:

1. summarizing theoretical research findings after gathering and speech data processing of the recorded databases;
2. detailed analysis of the corpora by selecting a language structure at a certain level, its description, classification of typical errors and their explanation with special emphasis on cross-linguistic influences between L1 and L2. The data analysis will be performed on phonetic-phonological, morphosyntactic, semantic and discourse level.

Contrastive analysis research studies for Macedonian and different language exist or are currently conducted. One of those studies concerns the L1 - Macedonian and L2 - English [16].

Prosody tests will be recorded both by non-native and native speakers. The resulting speech material will serve as a reference for the assessment of non-native prosody. The recordings will be conducted in a quiet room or studio with low noise and reverberation, using 2-channel input, i.e. close-talk and table/condenser microphone. Basic quality requirements are: sampling frequency 44.1 kHz, minimal resolution 16 bit, minimal SNR of 35 dB.

The whole speech material will be segmented and phonetically transcribed using force alignment by trained speech recognizer. Since there are no existing acoustical models on Macedonian language, UASR system trained on other language can be used for acoustic model bootstrapping and adaptation. Both knowledge-based and data-driven approaches for source and target language phoneme mapping can be used for initial transcription and labelling of small amount of recorded speech data. Initial experiments with German acoustic models confirm the usability for cross-language modelling as a first step toward fully trained acoustic model on Macedonian language.

Manual verification will be performed on parts of non-native speech database and where necessary following intervention will be applied: segment boundaries adjustment, marking of noises, disfluencies and pauses, check of the transcription and automatically inserted primary and secondary stress markers. Also, deviations from the canonical pronunciation (insertions, deletions and substitutions) will be marked as well.

## 5. STRUCTURE OF SPEECH EXAMPLES

Exercises will be 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. For the case of segmental features (phonemes, diphones...), the test structure will consist of material for production, perception and discrimination of L2 sounds in minimal pairs, in contrast with L1 sounds and in larger syntactic units to practice assimilations within and between words and phrases as well as missing/inserted syllables, words and phrases.

For the case of suprasegmental features, the exercise examples will be created to test and practice prosody in smaller and larger syntactic units. In isolated words suprasegmental identification will be 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.

At the level of simple and complex sentences exercises will consist in production and recognition of different types of sentences, i.e. declaratives, commands, wh-questions, yes/no questions, compounds, requests on the basis of their suprasegmental features.

Also identification and production of emphatic stress, relating focus with meaning and performing communicative functions with focus will be practiced and tested e.g. showing emotions, disagreement, correcting wrong information, calling attention to new information.

## 6. CONCLUSION

In this paper, a framework for creation and development of CAPT-Macedonian e-learning system is presented. To produce solid and useful pronunciation and prosody training environment, many technological challenges must be solved.

First, recording and processing of various speech databases must be performed, because this is one of the most important requirements to develop the important underlying speech technologies like ASR and speech synthesis for the target language (L2-Macedonian). Existing ASR and speech synthesis modules are vital components of the audio-visual feedback system of CAPT learning platform. Also, contrastive analysis research studies for Macedonian as L2 and the mother tongue of the target learners group will be used for prosody tests and exercises creation by the linguistics experts.

The proposed framework is in its early stage of development and further work will be focused on its complete realization and usability assessment by foreign learners of Macedonian language.

## LITERATURE


[1] H. Strik, A. Neri, C. Cucchiarini, „Speech technology for language tutoring“, (2008) Proceedings of LangTech-2008, Rome, February 28-29, 2008, pp. 73-76.

[2] Björn Granström, „Speech technology for language training and e-inclusion“, INTERSPEECH 2005: 449-452

[3] Hincks R., “Computer Support for Learners of Spoken English“. Doctoral Thesis in Speech and Music Communication. KTH, Stockholm, 2005

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

[5] Cylwik, N.; Demenko, G.; Jokisch, O.; Jäckel, R.; Rusko, M.; Hoffmann, R.; Ronzhin, A.; Hirschfeld, D.; Koloska, U.; Hanisch, L: „The use of CALL in acquiring foreign language pronunciation and prosody - General specifications for EURONOUNCE project“. In Proc. Workshop on Speech Analysis, Synthesis and Recognition (SASR), September 2008. Piechowice, Poland.

[6] Levis, J. (2007). „Computer technology in teaching and researching pronunciation“. Annual Review of Applied Linguistics, 27, 1-19.

[7] Ka-Ho Wong, Wai-Kit Lo, Helen Meng: Allophonic variations in visual speech synthesis for corrective feedback in CAPT. ICASSP 2011: 5708-5711

[8] Helen Meng, Wai-Kit Lo, Alissa M. Harrison, Pauline Lee, Ka-Ho Wong, Wai-Kim Leung and Fanbo Meng, "Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience," in the Proceedings of the 2nd Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA2010), Biopolis, SINGAPORE, 14-17 December 2010.

[9] Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). „Does popular speech recognition software work with ESL speech?“ TESOL Quarterly, 34, 592-603

[10] Neri A., Cucchiarini, C. and Strik H., 2001. „Effective feedback on L2 pronunciation in ASR-based CALL“. In Proc. of the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference, San Antonio, Texas

[11] Wachowicz K. A. and Scott B., 1999. „Software That Listens: It’s Not a Question of Whether, It’s a Question of How“, CALICO Journal 1999, Volume 16, Number 3

[12] Zinovjeva, N. 2005. „Use of speech technology in learning to speak a foreign language“. Retrieved on 16th July 2008 from http://www.speech.kth.se/~rolf/NGSLT/gslt_papers_2005/Natalia2005.pdf.

[13] Mixdorff, H., Külls, D., Hussein, H., Gong, S., Hu, G., Wei, S., “Towards a Computer-aided Pronunciation Training System for German Learners of Mandarin”, Proc. of SLaTE Workshop, Wroxall Abbey Estate, Warwickshire, England, September 2009.

[14] R. Hoffmann, M. Eichner, and M. Wolff: “Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system“. In: A. Esposito et al. (eds.), Verbal and Nonverbal Communication Behaviours. Berlin etc.: Springer 2007, Lecture Notes in Artificial Intelligence vol. 4775, pp. 200--218.

[15] Wells, J.C. 2000. Overcoming phonetic interference. English Phonetics, Journal of the English Phonetic Society of Japan, 3, pp. 9–21.

[16] "Analysis of Macedonian English Interlanguage at proficiency levels A1, A2 and B1 of the European Framework of Reference for Languages (CEFR) with special emphasis on cross-linguistic influence" - Project supported by the Ministry of Education and Science No. 13-3965/1