ISCA Archive
http://www.isca-speech.org/archive

SPECOM'2004: 9th Conference
Speech and Computer
St. Petersburg, Russia
September 20-22, 2004

Development of multi-voice and multi-language Text-to-Speech (TTS) and Speech-to-Text (STT) conversion system (languages: Belorussian, Polish, Russian)

Ruediger Hoffmann (1); Edward Shpilewsky (2); Boris Lobanov (3); Andrey Ronzhin (4)

(1) Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, Mommsenstr. 13, D - 01062 Dresden, Germany,
ruediger.hoffmann@ias.et.tu-dresden.de
(2) Institute of Computer Sciences, University Bialystok (ICS UB), Poland,
edwshp@hotmail.com
(3) Speech Synthesis and Recognition Laboratory, United Institute of Information Problems of the National Academy of Sciences of Belarus (UIIP NASB), Surganov Str. 6, Minsk, 220012, Belarus,
lobanov@newman.bas-net.by
(4) St. Petersburg Institute for Informatics and Automation, 39, 14th line, St. Petersburg, Russia,
ronzhin@iias.spb.su

Abstract

This proposal was submitted to INTAS Thematic Call on Information Technology 2004. The participants from 4 countries (Belarus, Poland, Russia and Germany) give efforts to the activity. The overall coordination, monitoring and control of the project will be implemented by Project Coordinator: Prof. Dr. Ruediger Hoffman *(Team 1, IAS TUD, Germany)* and Decision Board: Prof. Dr. Ruediger Hoffmann *(Team 1, IAS TUD, Germany);* Prof. Dr. Edward Shpilewsky *(Team 2, ICS UB, Poland);* Prof. Dr. Boris Lobanov *(Team 3, UIIP NASB, Belarus);* Dr. Andrey Ronzhin *(Team 4,* SPIIRAS, *Russia).* Duration of the research is two years.

1. Research Objectives

This research aims to fill the gap in introducing and promoting computerized speech technology for Slavonic languages, in particular for Belarusian, Polish and Russian. The main research objectives include:

- Creation of linguistic and acoustical resources for Text-to-Speech (TTS) synthesis (languages: Belarusian, Polish and Russian) and for Speech-to-Text (STT) recognition (Russian).
- Creation of multi-lingual and multi-voice TTS conversion system for Belarusian, Polish and Russian.
- Creation of data driven voice «cloning» technology for high quality personalization of speech provided by TTS conversion system.
- Creation of STT conversion system for continuous Russian speech recognition based on the morphemes level of speech representation.
- Creation of the complex PC-based software model and a pilot application based on TTS synthesis for Belarusian, Polish, Russian and STT continuous Russian speech recognition.

2. Background & Justification

Objective 1. Creation of Slavonic linguistic and acoustical resources for TTS and ATT conversion. To this date, certain experience in creating linguistic and acoustic resources has been gained for Russian and Polish. As for Belarusian, these resources are practically nonexistent. A decisive factor in achieving high quality of speech synthesis and recognition is the completeness of the resources and databases used. The research objective is to further develop the available linguistic resources (both vocabulary and grammar) and multi-voice acoustical databases for Polish and Russian, and to create new resources and databases for Belarusian in collaboration with research institutions from the INTAS member states and NIS involved in the project.

Objective 2. Creation of TTS conversion system for Slavonic languages. Slavonic language and speech systems, in particular, those of Belarusian, Polish and Russian, have very much in common. This is true of their phonetic, lexical, morphological and syntactic structure. This fact enables the researchers to set as objective the creation of an integrated algorithm of multi-language TTS conversion and the construction of a new TTS system common for all these languages. One may expect that such a system will be also applicable to other Slavonic languages, such as Czech, Slovak, Serbo-Croatian, Slovenian, Bulgarian and Macedonian. It should be noted that most of these languages are spoken in the countries that become or will soon become EC members. At present, only a few TTS systems of Polish and Russian speech generation are available. However, the quality of the synthesized speech is still far from natural, and the number of synthetic voices is very restricted. Belarusian TTS systems are not present at all.

The project objective can be achieved due to the use of original algorithms of multi-language and multi-voice

TTS synthesis, developed in research institutions from the INTAS member states and NIS involved in the project [1-5]. The Synthesis of phonemic characteristics of speech is based on the Allophones Natural Waves (ANW) method of speech signal concatenation. The basic principle of synthesizing the prosodic features of speech is the division of an utterance into accentual groups and the formation on their basis of entire tonal, rhythmical and dynamic contours of a syntagm and utterance as a whole [1,2]. By using Data Driven (DD) approach, the TTS system will resort to a vast multilingual set of phones and multuphones of ANW and prosodic features databases for the synthesis of speech sounds and intonation. To synthesize prosodic features, the system will be also resort to deep morphological and syntactic analysis of sentences [3]. The two modules are expected to achieve a high quality of synthesized speech.

Objective 3. Creation of data driven voice «cloning» technology. The quality of TTS synthesis largely depends on how close the model of human voice and pronunciation can be made. The voice "cloning" technology ensures a high quality of speech imitation for a specific individual by mean of TTS synthesis. There has been so far only one mention of an example of personal voice cloning by TTS (see www.naturalvoices.att.com), which however was only confirmed by a popular science publication (see: *"Voice Cloning - Software Recreates Voices of Living & Dead"*, By Lisa Guernsey, New York Times, 8 -1-1).

UIIP NAS Belarus has developed the basic rules of an original technology for cloning personal voice and pronunciation particularities [6, 7]. To successfully solve the problem of cloning of personal voice and pronunciation particularities, the following strategy will be used:

- The maximum use of a complex of acoustic characteristics carrying information about the individual voice and pronunciation properties of the speaker being imitated by utilizing of data-driven approach.
- The minimally possible distortions of the wave form elements of concatenation at all stages of their 'extraction', as well as the maximally possible accuracy of prosodic modifications in the process of speech synthesis.

A significantly larger amount of data concerning the individual voice and speech characteristics is generated by the *acoustic, phonetic and prosodic* processors. A reasonable approach to the construction of the acoustic processor consists in using fragments of natural speech waves as the minimal "genetic material" for 'cloning' a person's voice. Significantly, the sound wave contains all personal peculiarities of voice production as they manifest themselves in a given concrete allophone. Successful 'cloning' of individual phonemic peculiarities depends, for the main part, on the successful imitation of the peculiarities of phoneme-allophone transformation inherent to a given speaker. The individual features of prosodic organization manifest themselves both in the structural and the functional aspects. The structural aspect refers to the shape of the tonal, amplitude and timing configurations coextensive with the separate accentual groups. The functional aspect is concerned with the distributional characteristics of prosodic units, reflecting the frequency of their occurrence in the speech of the given speaker.

The cloning procedure is based on two types of texts-corpus and corresponding them audio-data from a speaker: a) for data-driven 'cloning' of individual voice and phonemic peculiarities, b) for data-driven 'cloning' of individual features of prosodic organization of the speech. The text's information is an input of TTS synthesizer, while the audio-data is an input of speech signal parameterization block. The output of TTS synthesizer conveys the information in the form of speech signal parameters labeled according to the allophone sequences. The next block uses the DTW algorithm to fulfill a labels transferring from synthetic speech to natural speech spoken by a certain speaker. The last block carries out the procedure of data-driven personal features collection [8].

Objective 4. Creation of STT conversion system for continuous Russian speech recognition. All the languages of Slavonic group have the common problem, which significantly decreases the quality of speech recognition with large vocabulary. It is the complicated mechanism of word formation. In contrast to English the Russian language have much more variety on word-form level and so the size of recognized vocabulary is sharply increased as well as quality and speed of the processing are decreased. Moreover the using syntactic constraints leads to that the errors of declensional endings cause the recognition error of the whole pronounced phrase. To decide these problems the morphemes level of speech representation is introduced [9].

Since during the process of word formation the same morphemes are often used so it will be useful to insert the additional level of speech representation – morphemic. Owing to word-form separation into morphemes the vocabulary size of recognized lexical units is significantly decreased. On the basis of existent dictionaries and word-formation rules of Russian language the databases of various types of morphemes (root, affix, inflexion) and also the methods for automatic text processing, which provide the morphologic word parsing and formation of word hypotheses from morphemes sequence will be developed. The developed databases of morphemes further will be used for collecting the statistics of morphemes co-ordination by text corpuses. At that during recognition the degree of co-ordination between root morphemes will have main significance [10]. As a result of such processing the speed of recognition and robustness to syntactical deviations in the pronounced phrase will be improved. The compiled speech databases for training phoneme models and using the hidden

Markov models will provide the speaker independence for natives of Russian languages.

Objective 5. *Creation of the complex PC-based software model and a pilot applications based on TTS and STT conversion.* The common PC based model proposed in the present project provides a natural integration of functions of multi-lingual and multi-voice TTS and STT conversion, based on large-vocabularies of spoken words and has no equals, at least for Slavonic languages. Moreover, the proposed model envisages a possibility to achieve data-driven «cloning» of personal voice and pronunciation particularities of the speaker by TTS synthesis. The final aim is a development of the complex PC-based software model, its testing, evaluation and creation of a pilot application – oral speech interpreter from Russian into Belarussian and Polish based on TTS and STT conversion.

3. Research activities

The research and development work done within the scope of the project includes the creation of reusable Belorussian, Polish and Russian linguistic resources and multi-voice acoustical databases, methodologies and algorithms of multi-lingual TTS synthesis, voice «cloning» technology and STT recognition as well as the construction of their pilot application prototypes. The proposed research down into the following individual tasks and sub-tasks:

Task 1. Development of linguistic and acoustical resources for Text-to-Speech (TTS) synthesis (languages: Belarussian, Polish and Russian) and for Speech-to-Text (STT) recognition (Russian).

Sub-Task 1.1. Development of *Belarussian and Russian* linguistic and acoustical resources for TTS

Sub-Task 1.2. Development of *Polish* linguistic and acoustical resources for TTS

Sub-Task 1.3. Development of *Russian* linguistic and acoustical resources for STT.

Objectives: Provide the lexical level information: words stress position, grammatical category of words and acoustical databases for *Belarussian, Polish and Russian* TTS conversion.

Inputs: Existing knowledge and experience of team members and task leaders.

Outputs: Large-size *Belarussian, Polish and Russian* pronunciation vocabularies that include data on word-forms stress position and grammatical category. Digital sound recording data that include the phonemic transcription and prosodic labeling for two male and two female voices and for two types of reading text, namely: one providing phonemic and another - prosodic peculiarities. Results of evaluation and testing.

Task Methodology: Creation of linguistic resources based on an advanced methodology multilingual natural language processing developed at the *IAS TUD, Germany* that provide morphological and syntactic information on sentence level and lexical level information on word stress position and grammatical category of words. Development, testing and evaluation of a multi-voice acoustical database based on the well-known methodology of the DARPA TIMIT acoustic-phonetic speech corpus creation, and partly, on an original methodology, developed at the *UIIP NASB, Belarus* for data-driven voice 'cloning' of personal and phonetic peculiarities of the speech.

Task 2. Development of multi-lingual and multi-voice TTS-synthesizer for Slavonic languages.

Sub-Task 2.1. Development of language-specific rules and algorithms of TTS synthesis for *Belarussian and Russian*

Sub-Task 2.2. Development of language-specific rules and algorithms of TTS synthesis for *Polish.*

Objectives: Provide the set of rules and algorithms for the creation of *Belarussian, Polish and Russian* TTS synthesis system.

Inputs: Linguistic resources and multi-voice acoustical database for *Belarusian, Polish and Russian*. Existing knowledge and experience of team members and task leaders.

Outputs: TTS synthesis rules and algorithms for Belarussian and Russian that provide:

- dividing a text into phone periods, phrases and syntagms, word accentuation, letter-to-phoneme transformation;
- Prosodic marking of syntagms, dividing syntagms into accentual groups, prosodic patterns creation for various intonation types;
- Positional and combinatory allophone creation, phoneme-to-allophone transformation;
- Waveform allophones creation, waveform concatenation and prosodic modification.

Results of evaluation and testing.

Methodology: Development, testing and evaluation of TTS-system based on an original methodology developed at the *UIIP NASB, Belarus* for the creation of the language-specific rules of textual, prosodic, phonetic and acoustical processors.

Task 3. Development of data driven voice «cloning» system for high quality personalization of speech pronunciation provided by TTS synthesis.

Objectives: Provide the set of rules and algorithms of language independent data-driven voice «cloning» technology for high quality personalization of speech provided by TTS synthesis.

Inputs: Linguistic resources and multi-voice acoustical databases for Slavonic languages. The set of rules and algorithms of multi-lingual and multi-voice text-to-speech synthesis.

Outputs: A set of rules and algorithms of data-driven voice «cloning» technology that provides personalization of acoustic, phonemic and prosodic features of speech generated by TTS synthesizer.

Results of evaluation and testing.

Methodology: Development, testing and evaluation of voice «cloning» technology based on an original methodology developed at the UIIP NASB (Belarus) for the creation of:

- a specialized TTS synthesis model supplied with an output conveying the information in the form of speech signal parameters labeled according to allophone sequences;
- Speech signal parameterization based on pith-synchronous spectrum analysis.
- Allophonic labels transfer from synthetic to natural speech spoken by a certain speaker;
- Several procedures for data-driven personal acoustic, phonemic and prosodic features collection.

Task 4. Development of STT conversion system for continuous Russian speech recognition based on the morphemes level of speech representation.

Objectives: Provide the set of rules and algorithms for high-performance speaker independent STT recognition.

Inputs: Linguistic resources and multi-voice acoustical databases for Russian language.

Outputs: computer model of voice interface, which provides speaker independent input of Russian speech

Methodology: Development, testing and evaluation of recognition system based on an original methodology developed at the SPIIRAS (Russia) for the creation of:

- the optimal structure of phonetic and morphemic units and the corresponding toolkit for their processing and database creation;
- the software module for accumulating the statistics of the morphemes to morphemes association by text corpus;
- creation of acoustic models with the help of the multi-language and multi-voice TTS synthesizer;
- multi-source speech signal analysis;
- compiled speech databases for training phoneme models and using the hidden markov models will provide the speaker independence for natives of Russian language;

Task 5. Development of the complex PC-based software model and a pilot application based on TTS synthesis for Belarussian, Polish, Russian and STT continuous Russian speech recognition.

Objectives: Development of the pilot applications that Integrate a full complex of scientific results of the project.

Inputs: The set of rules and algorithms of multi-lingual and multi-voice TTS synthesis.

The set of rules and algorithms of data-driven voice «cloning» technology.

The set of rules and algorithms of STT recognition.

Outputs:

- TTS synthesis prototype system for Belarussian;
- TTS synthesis prototype system for Polish;
- TTS synthesis prototype system for Russian;
- Voice «cloning» prototype system;
- STT recognition prototype system;
- Complex PC-based software models provided Belarusian, Polish, Russian TTS synthesis, voice «cloning» and Russian STT recognition.
- Oral speech interpreter from Russian into Belarussian and Polish (on an example of the ticket booking office) based on TTS and STT conversion

Methodology: Development of a complex PC-based software model based on C++ Programming Language running under Windows XP-Professional in accordance with the methodologies employed to achieve tasks 1-4. Testing and evaluation results.

4. Expected results

The research and development work done within the scope of the project includes the creation of reusable Belorussian, Polish and Russian linguistic resources and multi-voice acoustical databases, methodologies and algorithms of multi-lingual TTS synthesis, voice «cloning» technology and STT recognition as well as the construction of their pilot application prototypes. The work packages of the project includes:

1) Large-size Belarusian, Polish and Russian pronunciation vocabularies
2) Belarusian, Polish and Russian multi-voice acoustical databases
3) The set of rules and algorithms of Belarusian, Polish and Russian TTS synthesis
4) The set of voice «cloning» technology rules and algorithms
5) The set of rules and algorithms of STT recognition
6) The set of pilot applications:
   - TTS synthesis prototype system for Belarusian;
   - TTS synthesis prototype system for Polish;
   - TTS synthesis prototype system for Russian;
   - TTS voice «cloning» prototype system;
   - Speaker independent STT recognition prototype system for Russian;
   - Complex PC-based software prototype system provided Belarusian, Polish, Russian TTS synthesis and Russian STT speech recognition.
   - Oral speech interpreter from Russian into Belarussian and Polish prototype system based on TTS and STT convertions (on the example of a ticket booking office).

5. Potential for applications and dissemination of the results

In scientific field: The expected results will be of use in new directions of research in applied linguistics, in particular, in the study of phonetics and prosody of Slavonic languages, in theoretical issues of multilingual speech communication systems. The results will be used in the preparation of a book and articles to be published on the materials of the project as well as in lecture courses in the theory and applications of speech technologies at the Universities of Bialystok (Poland), Minsk (Belarus), Saint-Petersburg (Russia) and Moscow (Russia).

In technical field: The TTS and STT conversion systems developed during the project as well as its separate components and methods could be successfully applied for the Slavonic speech synthesis and recognition with large vocabulary for creation of industrially produced technical devices (cars, domestics robots, etc.) with speech interface for organization of more effective and robust control. Our results and models also can be embedded into the modern telecommunication applications, new generation of mobile devices and new perspective intelligent services and applications (smart room), where speech becomes the most effective mean for the inquiry and obtaining the information.

In economic and social fields: The results obtained will contribute to the development in Belarus, Poland and Russia of new areas of business activities and services, which are connected with the creation of: (1) on line telephone information services; (2) Internet portal navigation by voice, using the facilities of the speech synthesis and speech recognition module; (3) computer-based system for the blind and some other socially oriented systems, etc.

Dissemination will be done mainly through the following channels:

- Paper submissions to conferences and journals;
- Special interest e-mail distribution list for project news;
- Project homepage presenting the project and current results;
- Contacts to project teams and organizations that work on potentially relevant topics;
- Technology transfer management;
- Workshop to communicate and discuss results of the project;
- Involving customers and research partners;
- Reports will be created and published.

Potential technology user organizations will be associated when the project starts ensuring requirements transfer and evaluation. Besides the public exploitation activities like publications there will also be actions taken to protect intellectual property rights (IPR): Project management will provide an IPR manager contact managing all IPR-inquiries in order to harmonize activities. The project results have a high potential for application, especially voice «cloning» technology and result. All partners have contacts to potential technology or end-users, who show high interest in the new possibilities provided by this approach.

6. References

[1] Lobanov B. *The Phonemophon Text-to-Speech System*. Proc. of the XI International Congress of Phonetic Sciences, Tallin, 1987, pp. 61-64.

[2] Lobanov B., Karnevskaya H. *MW Speech Synthesis from Text*. Proc. of the XII International Congress of Phonetic Sciences. Aix-en-Provense, France, 1991, pp. 406-409.

[3] Boguslavsky I., Lobanov B. and Karnevskaya H. *Generation of Intonation and Accentuation of Synthetic Speech on the Base of Morpho-Syntactic Knowledge*, Proceedings of the International Workshop “Integration of Language and Speech”, Moscow, 1996, pp. 11-28.

[4] Lobanov B., Jokisch O. et al. *A Bilingual German/Russian Text-to-Speech System.* Proceedings of the 3rd International Workshop “Speech and Computer” – SPECOM’98, St.-Petersburg, 1998, pp327-330.

[5] Lobanov B., Shpilewski E. et al. *Polish TTS in Multi-Voice Slavonic Languages Speech Synthesis System* . Proc. of International Conference SPECOM’2004, St. Petersburg, 2004.

[6] Lobanov B. and Karnevskaya H. *TTS-Synthesizer as a Computer Means for Personal Voice (On the example of Russian).* In the Book: Phonetics and its Applications. Stuttgart: Steiner. 2002, pp. 445-452.

[7] Lobanov B. and Tsirulnik L. *Phonetic-Acoustical Problems of Personal Voice Cloning by TTS*. Proc. of International Conference SPECOM’2004, St. Petersburg, 2004.

[8] Wolff M., Eichner M. Hoffmann R. *Improved Data-Driven Generation of Pronunciation Dictionaries Using an Adapted Word List.* Proc. of the 7th European Conference on Speech Communication and Technology – EUROSPEECH 2001, Scandinavia, 2001, pp. 1433-1436.

[9] A.L. Ronzhin, A.A. Karpov. *Implementation of morhpemic analysis to Russian speech recognition*. Proc. of International Conference SPECOM’2004, St. Petersburg, 2004.

[10] Kosarev Yu, Ronzhin A., Karpov A., Lee I. "Continuous Speech Recognition without Use of High-Level Information", 15th International Congress of Phonetic Sciences, Barcelona, August 2003, pp 1373-1376.