Slavonic TTS and SST Conversion for “Let's Fly” Dialogue System

Ruediger Hoffmann (1), Oliver Jokisch(1), Boris Lobanov (2),Liliya Tsirulnik (2), Edward Shpilewsky (3), Bozhena Piurkowska(3), Andrey Ronzhin (4), Alexey Karpov (4)

(1) Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, Germany
(2) United Institute of Informatics Problems, National Academy of Sciences, Belarus
(3) Institute of Computer Sciences, University of Bialystok, Poland
(4) St. Petersburg Institute for Informatics and Automation, Russia

ruediger.hoffmann@ias.et.tu-dresden.de, lobanov@newman.bas-net.by, edwshp@hotmail.com, ronzhin@iias.spb.su

Abstract

The present paper sketches some outlines of the large vocabulary text to speech (TTS) and speech to text (STT) systems for Russian, Belarussian and Polish languages. The mentioned languages have similar phonetics and grammar structure thus it is possible to create baseline speech processing models, which will be easy-to-apply for most Slavonic languages. The developed systems were integrated in multilingual dialogue system “Let’s Fly” providing information on flight schedule.

1. Introduction

Information exchange and cross-lingual communication is of fundamental importance for multilingual Europe. The barriers to communication can be lowered through the use of appropriate human-computer interfaces technologies to support multilingual speech engines [1]. At the present time the problem of natural human-computer interaction is actively investigated in the framework of on-going projects.

Slavonic languages and speech systems and, in particular, those of Belarusian, Polish and Russian, have very much in common, which is true of their phonetic, lexical, morphological and syntactic structures. This fact enables the researchers to set as an objective the creation of an integrated algorithm of multi-language TTS conversion system common for all these languages. At present, only a few TTS systems for Slavonic speech generation are available. However, the quality of synthesized speech is still far from natural, and the number of synthetic voices is but restricted. TTS systems for some Slavonic languages do not exist at all. This work is an attempt to fill the gap in the creation of Slavonic TTS and its application.

The inflective nature of Slavonic languages leads to rich morphology and complex structure of word formation which complicates text and speech parsing. A word-based model, commonly used for English, is not suitable here. To improve the performance of speech to text synthesis (STT) the morpho-phonological tree structure is proposed for decoding the input speech signal. Fusion of the identical morphs in different words significantly reduces the search space for a large vocabulary. Recently similar approach, based on lexical-tree decoder, has been successfully applied for other inflective and agglutinative languages [2]. The key feature of the developed model is the preliminary calculation of a probabilities pool of all phonemes by a multiprocessor system and the following speech decoding by a morpho-phonological tree.

The peculiarities of the Slavonic languages and those of the developed multilingual TTS synthesizer are presented in Section 2. Section 3 describes the essence of the multithread STT using the morpho-phonological tree decoder and root-based language model and the application of Slavonic TTS and STT for the “Let’s Fly” dialogue system.

2. Slavonic Speech Synthesis Baseline System

The TTS conversion system under discussion has a common structure for all Slavonic languages concerned but it uses different linguistic and acoustical resources for each language. The objective is the development of a high-quality multi-lingual and multi-voice TTS-system on a common platform [3]. The speech signal synthesis is based on Allophone and Multi-Allophone Natural Waves (ANW and MANW) concatenation. The speech prosody synthesis is based on an Accentual Units Portrait (AUP) model of stylizing the entire tonal, rhythmical and dynamic contours of a phrase and an utterance as a whole [4]. These two modules operating jointly are expected to produce a high quality of synthesized speech. The quality of TTS synthesis largely depends on how close to human voice and pronunciation the model can be made. The voice "cloning" technology ensures a high quality of speech imitation for a specific individual by means of TTS synthesis [5].

2.1. Peculiarities of Belarussian, Polish and Russian phonetic systems

Phonetic systems of the Slavonic languages group have much in common among themselves, however, each of them also possesses specific features, sometimes significant.

The phonetic systems of the Belarussian, Polish and Russian languages, in particular, are rather close, especially Russian and Belarussian. There are 42 phonemes in Belarussian and Russian (6 vowels and 36 consonants), but the inventories of the phonemes are different. The phonetic system of Polish is more varied. There are 51 phonemes in it: 8 vowels and 43 consonants.

The distinctive features of the phonetic systems of Belarussian and Russian consist in the fact, that some of the consonants found in Russian are missing in Belarussian, namely: soft alveolar /*Т'/, /Д'/, /Ш'/, /Ч'/, /Р'*/, hard velar /*Г*/ and soft velar /*Г'*/ (here and later the phonetic symbols are denoted by the national alphabets). On the other hand, there are a number of specific consonants in Belarussian, which are missing in Russian: bilabial liquid /*Ў*/, alveolar voiceless affricates - hard /*Ч*/ and soft /*Ц'*/, alveolar voiced affricates - hard /*Дж*/ and soft /*Дз'*/, velar voiced - hard /*Гх*/ and soft /*Гх'*/.

Comparing the phonetic systems of Polish and Russian, we shall also note some features of similarity and difference. In the Polish language there are all the phonemes, characteristic of Russian, however, the pronunciation of soft fricative /*Ш'*/ and affricative /*Ч'*/ differs from the Polish soft /*Ś*/ and /*Ć*/, the place of articulation of which is intermediate between the soft Russian /*С'*/, /*Ш'*/ and /*Ц'*/, /*Ч'*/, accordingly. Besides, in the Polish language there are a number of specific phonemes, which are absent in Russian: liquid bilabial /*Ł*/, affricates soft - /*С'*/, /*Ć*/ and hard - /*Cz*/, voiced affricates soft - /*Dź*/**,** and hard - /*Dż*/, /*Dz*/, nasalized vowels - /*Ą*/ and /*Ę*/.

2.2. Allophonic Representation of Phonemes Stream

As it is known, phonemes in the speech flow are realized in the form of allophones, or otherwise, in the form of positional and combinatory variants of phonemes. Experience of creating Russian-speaking TTS has shown, that synthesized speech of sufficiently good quality can be reached under conditions of generating the so-called mini-set of positional and combinatory allophones.

For the mini-set of allophones for Russian speech synthesis 3 types of positional allophones of vowels are created: stressed - (0), pre-stressed -(1) and post-stressed - (2). To present the left context the following combinatory allophones of vowels are created: after a phrase pause - (0), after hard consonants labial - (1), alveolar - (2), velar - (3), and after soft consonants - (4). In all there are 5 types of left contexts. For the right context there are the following combinatory allophones of vowels: before a phrase pause - (0), before hard consonants labial - (1), alveolar and velar - (2), and before soft consonants - (3). In total for 6 vowels we avail of Nv = 3*5*4*6 = 360 allophones. Allophones of consonants are created only with regard to the right context: before a pause - (0), before unvoiced - (1) and voiced - (2) consonants, before unstressed- (3) and stressed - (4) vowels.

In total for 36 consonants we obtain Nc = 5*36 = 180 allophones. Overall, in a mini-set for Russian we have 360+180 = 540 allophones and the same number for Belarussian. Similar calculations for Polish give the number of 480+215 = 695 allophones.

Estimations of allophone number calculated theoretically are, a-priori, considerably higher since, firstly, many positional and combinatory situations do not occur in speech altogether and, secondly, because for many allophones the acoustic distinctions are so insignificant, that they can be neglected. As is proved by experience, the number of allophones used in a mini-set appears to be about 1.5 times smaller.

At the beginning stage the mini-sets of Allophone Natural Waves (ANW) for TTS synthesis of each language are created manually. At the next step the mini-sets of ANW are utilized for an automatic creation of maxi-sets of Multi-Allophone Natural Waves (MANW) - sequences of two and more ANWs. MANWs are associated with the most frequent diallophones and syllables. Automatic creation of MANW DB is realized by data driven voice “cloning” technology. The total number of MANWs obtained in this way and utilized for concatenative synthesis is around 6 - 9 thousand, depending on the language.

2.3. General Description of the Slavonic Multi-language TTS-synthesizer

The general structure of the multi-lingual and multi-voice TTS-synthesizer is shown in Fig.1. The incoming orthographic text undergoes a number of successive operations carried out with the help of specialized processors.

The textual processor is devised to transform the incoming orthographic text into a prosodically marked one. The processor performs the following tasks:
• dividing an orthographic text into utterances;
• transforming numbers, abbreviations, etc;
• dividing an utterance into phrases;
• placing word’s accents (weak and strong);
• dividing phrases into accentual units (AU);
• marking the intonation type of the phrases.

The prosodically marked text is then sent to the phonemic processor, which performs the following tasks:
• phonemic transcription of the text;
• transforming the phonemic text into allophonic;
• combining the allophones into allosyllables.

The prosodic processor performs the following tasks:
• splitting AU into the elements of accentual units (EAU): pre-nuclear, nuclear and post-nuclear parts;
• generating the fundamental frequency (F_0) contour as well as the amplitude (A) and phoneme duration (T) assigning values according to the accent unit portrait for each accent unit.

The acoustical processor uses the information coming from the phonemic and prosodic processors to provide:
• the prosodic parameters modification of ANWs and MANWs;
• concatenation of ANWs and MANWs to the appropriate sequence.

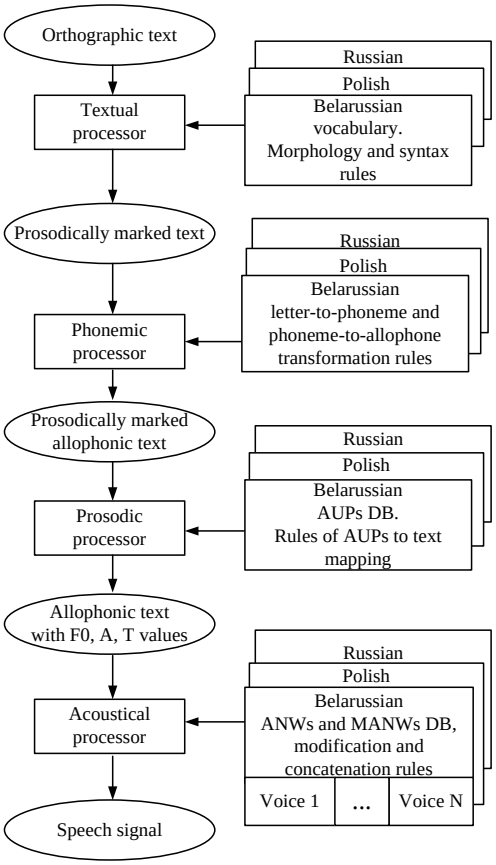

Fig. 1. General structure of the Slavonic TTS-synthesizer

Finally, by concatenating ANWs and MANWs and their modifications in accordance with the current values of F_0, A, T it generates the speech signal.

3. Using TTS and STT Conversion in "Let's Fly" Inquiry System

The multi-language TTS synthesizer described above was applied in the "Let's Fly" dialogue system providing flight information. The architecture of the developed inquiry system "Let's Fly" is presented in Fig. 2. The dialogue model for this task was initially obtained in Wizard of Oz (WoZ) mode and the accumulated speech material was used for building the main databases for speech recognition.

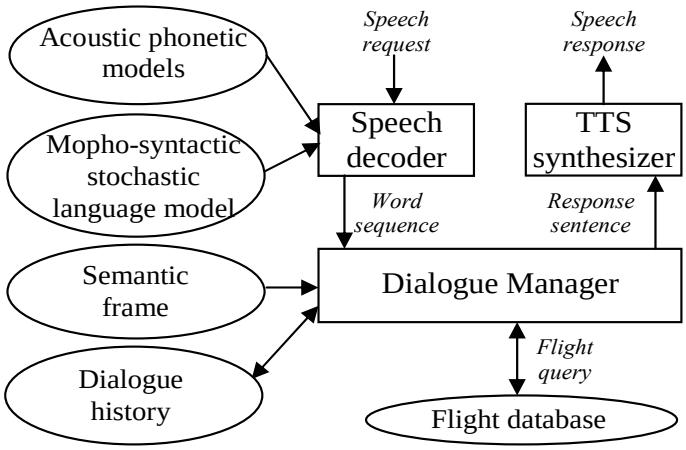

Fig. 2. General architecture of "Let's Fly" inquiry system

In WoZ mode 83 telephone calls and dialogues were recorded via telephone channel and processed manually. In each dialogue the user's and the operator's phrases were labeled accordingly. Separate acoustic, orthographic, and transcription files were created for every recorded phrase. Besides, breath and filled pauses, the clicking of the lips or the mouth and other speech artifacts are marked for training the corresponding acoustic model. A word recognition vocabulary was constructed based on flight databases and analysis of real dialogues. Now it contains over five hundred words including cities, countries, dates, times, ticket class, many common words, speech artifacts and noise models, which arise in real dialogues in the situation of air ticket booking.

The speech decoder based on morpho-phonological tree uses a time-synchronous Viterbi search with token passing and effective beam pruning techniques applied to re-entrant copies of lexical tree. During decoding a token keeps the indexes of passing morphs and their acoustic probabilities.

The proposed dialogue manager is able to process phrases formulated in free-forms. At that a user possesses complete freedom of discourse and can choose the most natural suitable conversation style. The values of concept attributes are extracted from the input sentences using the frame-based approach to dialogue analysis. The risk of dialogue frustration is minimized due to explicit and implicit confirmation strategy and analysis of dialogue history. The phrases generated by the dialogue manager can include two types of information: the flight information or a clarification request. The latter is used for recovering the lost information and filling the frame request to the flight information database. The operator's phrases accumulated during the WoZ mode together with the flight information as well as different types of general inquiry information were used in the system response and generated by TTS syntesizer.

4. Conclusion

Common and specific features of the phonetic, lexical, morphological and syntactic structures of three Slavonic languages are investigated. A multilingual speech synthesizer was created based on MANW and AUP methods. To improve the performance of speech recognition for Slavonic languages a morpho-phonological tree decoder with a preliminary calculation of the probabilities pool of all phonemes on parallel computers was proposed. The developed models of speech processing were integrated in an airline information inquiry system, which was initially trained in WoZ mode and is now tested in an automatic mode. Naturalness of communication is provided by using a mixed initiative dialogue strategy.

Acknowledgement

This paper was supported by the European Commission under grant INTAS No 04-77-7404.

References

1. Human Language Technologies for Europe. *TC-STAR Public Report #17*, April 2006, 64.

2. *Kurimo, M., Creutz, M. et al*. Unsupervised segmentation of words into morphemes - Morpho Challenge 2005, Application to Automatic Speech Recognition. Interspeech, Pittsburgh, 2006, 537-540.

3. *Hoffmann, R., Jokisch, O. et al*. A multilingual TTS system with less than 1 mbyte footprint forembedded applications. ICASSP, Hong Kong, 2003, v. 1, 532-535.

4. *Lobanov, B., Tsirulnik, L. et al*. Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis. Speech Prosody, Dresden, 2006, 553-556.

5. *Lobanov, B., Karnevskaya, H*. TTS-Synthesizer as a Computer Means for Personal Voice “Cloning”. In: Braun, A., Masthoff, H.R. (eds), *Phonetics and its Applications*. Stuttgart: Franz Steiner Verlag, 2002, 445-452.