Abstract. This paper presents the development of a Croatian text-to-speech (TTS) system capable of synthesizing speech from arbitrary text. For speech generation Pitch Synchronous Overlap and Add (PSOLA) and Hidden Markov Models (HMM) based methods were used. The generated Croatian speech was evaluated subjectively by the mean opinion score (MOS) of the 40 evaluators and objectively by the automatic speech recognition (ASR) system. In the paper we propose an evaluation approach which combines objective and subjective evaluation results.

Keywords. speech synthesis, Hidden Markov Models, diphone speech synthesis, PSOLA, MOS evaluation, ASR.

1. Introduction

Text-to-speech (TTS) systems or speech synthesis systems transform the input text into speech. The use of TTS systems is widely spread over different IT areas and applications: in telecommunication services access, in information enquiring over the telephone, in multimedia applications, in audio books and toys, in computer assisted language learning, in eyes and hands free applications. TTS can help blind and visually impaired persons accessing the written electronic content. But TTS systems are still far from natural speech: the generated speech can be oddly intonated, distorted at concatenation boundaries or the clarity and intelligibility can be poor.

There are several methods of speech signal generation from a given phonetic string and prosody. Each of them has its specific characteristics regarding intelligibility, naturalness, flexibility, system requirements and time required for implementation. Which method will be used depend mainly on the system requirements (like available memory) and expected voice quality. Most speech synthesis systems use concatenative synthesis methods [1], since this method produces the most natural speech.

Concatenative speech synthesis concatenates stored speech units into the word sequences according to the pronunciation dictionary [1]. Special signal processing techniques such as SOLA (Synchronous Overlap and Add), PSOLA (Pitch Synchronous Overlap and Add) or MBROLA (Multi Band Resynthesis Overlap Add) are used to smooth the unit (segment) transitions and to model the intonation [5]. Methods which are able to produce more natural speech are generalizations of the concatenative synthesis which is based on dynamic selection of different speech units from a large speech corpus. Those methods are also known as corpus synthesis [1]. The hidden Markov model (HMM) TTS systems using context dependent phone models are representatives of statistically corpus based speech synthesis methods [2]. In this work we used PSOLA and HMM speech synthesis methods.

The main issue in using voice interfaces in different applications is the quality of TTS generated speech. Speech quality is a perceptual value, so no exact measures for generated speech evaluation are proposed yet [10]. The TTS systems are therefore evaluated through series of tests that measure the listener’s subjective views using mean opinion scale (MOS) and listener’s preferences. The MOS tests are designed according to the ITU recommendation [9].

In this work we used the subjective and objective evaluation methods. For the subjective evaluation we gathered the opinion of 40 evaluators and for the objective evaluation we used an automatic speech recognition system. Finally we propose a combination of subjective and objective evaluation of synthesized speech which enables the verification of subjective evaluation.
The paper is organized as follows: in section two we present the common structure of a TTS system, the third and the fourth parts describe the phonetic analysis and speech generation methods of a Croatian speech synthesis system. The fifth part is dedicated to subjective, objective and combined evaluation results. We conclude the article with a discussion and some suggestions for future work.

2. Overview of a Text-to-speech System

The common architecture of a text-to-speech system is modular and consists of text analysis, phonetic analysis, prosodic analysis and speech synthesis modules [1].

Text analysis module processes the input text, transcribing numbers, symbols, acronyms and other elements that cannot be unambiguously pronounced into words. This is called text normalization. For example, “2. sv. rat trajao je” is transcribed into “drugi svjetski rat trajao je”. In this process an extensive set of rules is used.

Phonetic analysis module transforms the normalized text into the corresponding stream of phones. For Croatian this is a rather simple process because there are specific rules that can be applied to grapheme-to-phoneme transformation. Foreign words and some exceptions are handled by a phonetic dictionary which comprises words and their phonetic transcription.

The goal of a prosodic analysis module is creating synthesis parameters that will give the most natural output for a given sentence. These parameters include volume, pitch, pauses, speed and rhythm. Pitch is represented with fundamental frequency of glottal pulses F0 in Hertz. Besides the fundamental frequency, phoneme duration and speech volume are estimated.

Speech synthesis module is responsible for speech generation. In concatenative speech synthesis a speech segment is synthesized by simply reproducing segments of previously recorded natural speech [1]. A desired utterance is synthesized by reproducing a sequence of such segments that correspond to the given phonetic string. If segments are extracted from different contexts, distortion can occur at concatenation boundaries. These segments can also have completely different prosodic characteristics that cannot occur in natural speech. Overlap-and-add (OLA) and pitch synchronous overlap-and-add (PSOLA) algorithms are used for prosody modification and joining segments in concatenative speech synthesis [5]. If there are well-fitting segments in the database which are used to synthesize the desired utterance, synthetic speech can sound almost as good as natural speech.

In HMM speech synthesis the modified Viterbi algorithm is used to find the most probable path through HMM states that can generate the speech signal feature vectors [2]. During the phone duration period the feature vectors consisting of mel-cepstrum parameters and excitation parameters are generated. The speech signal can then be synthesized from such generated feature vectors using the source-filter model.

3. Phonetic Analysis

3.1. Speech Corpus

Croatian speech synthesis was developed using 2.5 hours or 2332 utterances of selected male speaker’s speech from the VEPRAD corpus [3]. The phonetic dictionary contains 6222 different accented words and their phonetic transcriptions. The standard set of 30 Croatian phonemes according to the Speech Assessment Methods Phonetic Alphabet (SAMPA) was extended with accented vowels (a:, e:, i:, o: and u:), vibrant phoneme /r/ as a vowel and silence. For the Croatian HMM and PSOLA TTS the same set of 37 Croatian phonemes was used.

The automatic segmentation of the speech signals is performed using forced alignment of the spoken utterances and their corresponding word level transcriptions [7]. The results of automatic segmentation are exact time intervals for each phone. Automatically segmented speech was used for HMM TTS training and for diphone database construction in PSOLA TTS.

3.2. Diphone Database

Acoustic units used in concatenative speech synthesis are phonemes, sub-phonetic units, diphones, triphones, syllables, words and phrases. The number of concatenation points and possible places where distortion can occur is reduced when longer units such as words or phrases are used as in corpus based synthesis methods.

The diphone O-B in the word “oblak” is a unit that spans from the middle of the phone o to the middle of the phone b. In the set of 37
Croatian phonemes it is possible to differentiate $37 \times 37 = 1369$ diphones which correspond to all possible phoneme pairs. The selected speaker’s speech contained only 1057 diphones each represented with one instance at least. The automatically segmented speech segments were divided into diphones. To obtain the most constant and neutral diphone prosody we preferred segments extracted from the middle part of the words to those extracted from word beginnings or endings.

3.3. Croatian Grapheme-to-Phoneme Conversion

The essential goal of the phonetic analysis phase in speech synthesis is transforming graphemes into phonemes. The normalized input text is transformed into the corresponding stream of phones according to orthographic-to-phonetic rules. Croatian orthographic rules are based on the phonological-morphological principle which enables phonetic transcription automation. Standard orthographic-to-phonetic rule according to which there is one phonetic symbol for one grapheme was extended with additional rules; for example [3]:
- sequences $ds$ and $ts$ were phonetically transcribed as [c]
- suffix $naest$ was phonetically transcribed as [n a j s t]
- words with četiri were phonetically transcribed as [C e t r i] etc.

4. Croatian TTS Systems

4.1. HMM TTS

The hidden Markov model based trainable speech synthesis uses the Croatian speech for context-independent and context-dependent acoustic models training and uses HMM as a generative model for speech production [2, 3]. The model training was performed using HTS (HMM Based Speech Synthesis System) [8] which is an extension of the HTK Toolkit [9].

The HMMs were simultaneously embedded-trained with the feature vectors consisting of spectrum, pitch and their dynamic features in a unified framework of multi-space probability distribution HMMs and multi-dimensional Gaussian distribution. Since the observation sequence of fundamental frequency is composed of one-dimensional continuous function for the voiced speech segments and a constant for the unvoiced speech segments, multi-space probability distribution is used [2]. The HMM state output feature vector consists of spectrum and excitation part using a multi-space probability distribution [4]. State duration densities for the speech feature vectors generation are estimated according to the probabilities obtained in the last iteration of embedded re-estimation.

4.2. PSOLA TTS

In concatenation based TTS system for Croatian the implemented method is based on PSOLA (Pitch Synchronous Overlap and Add) algorithm [5]. The input to the synthesizer is a string of phones and optionally between - word pause duration in samples [6]. The output signal is synthesized by overlapping and adding the diphones according to the input phone sequence. First the signals are overlapped and aligned with the pitch marks. Then the overlapping parts of the signals are multiplied by a Hanning window and added together, and the non-overlapping parts are copied directly into the output signal.

Basic functionality of the phonetic analysis module is provided by a Matlab function that takes some normalized text as an input and outputs the corresponding phone string.

5. Evaluation of Generated Speech

The synthesized speech was evaluated subjectively by the mean opinion score (MOS) of 40 evaluators and objectively by the speech recognition system (ASR). The combined approach compared objective and subjective
results. Fig. 1 presents evaluation of synthesized speech using: the subjective (a), the objective (b) and the combined (c) test methods. For the purpose of the test, 30 s (76 words in 9 sentences) of synthesized Croatian speech concerning weather forecast, temperature, sea and wind conditions, visibility at the sea and the river water level was generated.

5.1. Subjective evaluation results

The HMM generated speech was played to 21 evaluators and PSOLA generated speech was played to 19 evaluators who were asked to fill in the questionnaire. The questionnaire consists of three parts: in the first part, the evaluator’s opinion about the overall quality of generated speech, intelligibility, naturalness, pronunciation and usability in the telephony applications is inquired. For each attribute except for the usability, five possible answers were offered. And three possible answers were offered for the usability. Later, to each qualitative answer a score from 1 to 5 was assigned. In the second part of the questionnaire, the intelligibility was estimated through 4 questions. In the third part of the questionnaire, evaluators were given the opportunity to write down their remarks and opinion.

Mean Opinion Score

<table>
<thead>
<tr>
<th>Quality</th>
<th>HMM</th>
<th>PSOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligibility</td>
<td>3.29</td>
<td>3.9</td>
</tr>
<tr>
<td>Naturalness</td>
<td>3.05</td>
<td>3.92</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3.14</td>
<td>3.92</td>
</tr>
<tr>
<td>Usability</td>
<td>3.19</td>
<td>3.92</td>
</tr>
<tr>
<td>OVERALL</td>
<td>3.02</td>
<td>3.22</td>
</tr>
</tbody>
</table>

Figure 2. The MOS results for the HMM and PSOLA synthesized speech.

The female and male evaluators were mainly students of electrical engineering, informatics or linguistics, some professionals with experience in speech technologies, linguists, and one person with vision impairment who uses a screen reader on daily basis. The results were produced according to the mean opinion score (MOS) scale. Fig. 2 presents the MOS results for HMM and PSOLA TTS. The highest score was achieved in intelligibility for PSOLA speech (3.9) and the lowest in the quality of HMM generated speech (2.43). Overall results 3.02 for HMM and 3.22 for PSOLA speech are showing that subjective perception of generated speech for both systems is average.

Intelligibility test

<table>
<thead>
<tr>
<th>Question</th>
<th>HMM</th>
<th>PSOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather condition</td>
<td>83.2%</td>
<td>86.9%</td>
</tr>
<tr>
<td>Temperature</td>
<td>76.9%</td>
<td>92.86%</td>
</tr>
<tr>
<td>Visibility</td>
<td>92.86%</td>
<td>92.86%</td>
</tr>
<tr>
<td>River water level</td>
<td>54.21%</td>
<td>86.89%</td>
</tr>
<tr>
<td>Overall</td>
<td>83.10%</td>
<td>86.89%</td>
</tr>
</tbody>
</table>

Figure 3. Intelligibility test results.

In the second part of the questionnaire, the intelligibility of synthesized speech is inquired through 4 questions about the weather. The results are presented in Fig. 3. Intelligibility test shows better percentage of correct answers in the HMM system (88.69%) over the percentage of correct answers in the PSOLA system (82.89%). But still there is no significant difference between intelligibility of HMM and PSOLA generated speech.

5.2. ASR for objective evaluation

The ASR for objective evaluation of generated speech was trained on radio speech using the VEPRAD corpus [3].

In speech recognition, the speech signal feature vectors consist of 12 mel-cepstrum coefficients and their derivatives and acceleration.

The automatic speech recognition (ASR) system was trained on the radio data: from the weather forecast domain and from the news domain. The radio ASR system was trained on 8030 (82%) utterances spoken by 8 male and 8 female speakers and tested on 1712 (18%) utterances spoken by 3 male and 6 female speakers. In all experiments backoff bigram language model was used. Estimated perplexity of the radio bigram language model is 17.16 (10230 different words).

The development of the ASR system followed HTK procedures [7]. The monophone models with continuous Gaussian output probability functions described by diagonal covariance matrices were trained for the set of 30 standard Croatian phonemes and 4 additional models for silence, breathing sound,
mispronounced words, hesitations and noise. Each monophone model consists of 5 states, where the first and the last states have no output functions. The initial Baum-Welch algorithm training on monophone HMMs resulted in a monophone recognizer, which was used for the automatic segmentation of the speech signals.

The automatic segmentation of the speech signal to the phone level is performed using the forced alignment of the spoken utterance and their corresponding transcription on the word level. The automatically segmented speech for selected speaker was also used for the speech synthesis developed almost without any manual intervention.

Furthermore, the monophone models were trained on 10 passes of the Baum-Welch algorithm and such monophone models were used to initialize cross-words context-dependent triphone hidden Markov models. The triphone models consist of 5 states with the mixture of continuous density output, described by diagonal covariance matrices. The state tying was performed due to the lack of the acoustic material using 83 Croatian phonetic rules. The same phonetic rules were used for state tying in the context-dependent modelling of the speech synthesis acoustic model.

The number of output Gaussian probability density functions per state used in the triphone recognizer was increased to 20.

| Table 1. ASR correctness and accuracy of synthesized speech. |
|-----------------|-----|---------|---------|
| HMM speech      | No. words | Corr. % | Accurac. % |
| approach (1)    | 92            | 75.00   | 51.09    |
| approach (2)    | 76            | 84.21   | 68.42    |
| PSOLA speech    | No. words    | Corr. % | Accurac. % |
| approach (1)    | 92            | 78.00   | 73.00    |
| approach (2)    | 76            | 86.25   | 85.00    |

5.3. Objective Evaluation Results

The developed speech recognition system for the weather domain was used for the objective evaluation of the synthesized speech. Correctness and accuracy for HMM and PSOLA generated speech achieved are presented in Table 1. In the first approach (1) all incorrectly recognized silences and inspirations are counted as errors. In the second approach (2) all, correctly or incorrectly, recognized silences and inspirations were ignored.

Accuracy $WA$ is computed according to:

$$WA = 100 \% \left(1 - \frac{W_S + W_D + W_I}{N}\right),$$

where $W_S$, $W_D$ and $W_I$ are substituted, deleted and inserted words, while $N$ is the total number of words. $W_S$, $W_D$ and $W_I$ are computed using the Levenshtein distance between the transcribed and recognized sentences. Correctness $WC$ is computed according to:

$$WC = 100 \% \left(1 - \frac{W_S + W_D}{N}\right).$$

Relatively high recognition correctness and accuracy for HMM and PSOLA speech is achieved partly due to the possibility of repeated listening to the speech. The objective evaluation results showed improvement of accuracy in ASR of PSOLA over HMM speech of 22% for approach 1 (strictly error counting) and of 17% for approach 2 (smoother error counting). The improvement ground lies in less incorrectly inserted words which points to the good smoothing at transition boundaries in PSOLA.

5.4. Combined evaluation results

For the purpose of the combined evaluation, 8 evaluators were additionally asked to write down the transcription of the synthesized speech. Two of them wrote down the transcription of both HMM and PSOLA synthesized speech. The correctness and accuracy of their transcription was calculated in the HTK environment which makes it comparable with the correctness and accuracy achieved in the objective test. In the objective test the same synthesized speech was recognized by the speech recognition system. During the evaluation, incorrectly recognized silences and inspirations were ignored. Fig. 4 presents the comparison of results achieved by speech recognition system and by evaluators for
HMM and PSOLA speech. The results are compared using the second approach.

Evaluators got better results than the ASR. The main reason is a great ability of humans to adapt fast to the speech of lower quality, unusual accent etc. The adaptation ability of humans is directly associated with the exposure to the hundreds and thousands of hours of different speech which is simply not available for ASR system development.

5.5. The evaluators’ opinion

In their remarks, evaluators noticed that the intelligibility of the synthesized speech is lower due to the fast speaking rate, “sharp” pronunciation and due to the too short pauses between sentences. Two of them noted that certain amount of effort is needed for complete understanding. The remark about higher attention was expected since majority of evaluators had no previous contact with artificial speech of any kind.

6. Conclusion

In this paper we presented the development of a Croatian text-to-speech system capable of synthesizing speech from arbitrary text. For speech generation, PSOLA and HMM based methods were used. The Croatian speech synthesis was developed using 2.5 hours of selected male speaker’s speech from the VEPRAD corpus. A set of 37 Croatian phonemes was used in both systems. 1057 diphones were automatically segmented.

Croatian speech generated by both systems was evaluated subjectively by the mean opinion score (MOS) of the 40 evaluators and objectively by the automatic speech recognition (ASR) system. The combined approach compared objective and subjective results.

The combination of objective and subjective evaluation results put different perspectives on the quality of the synthesized speech. Although the overall score is only 3.02 or 3.22, the human evaluators performed better on the comparison test. Subjective evaluation enabled better understanding of user opinions and needs and therefore should not be omitted in the synthesis evaluation, while the objective evaluation enabled different perspective on the generated speech.

The combined evaluation showed that generated speech quality could be improved, especially in the speech technology applications. For further improvement an extended phoneme set, pronunciation rules and advanced methods of corpus synthesis will be considered.

7. References