International Journal of Engineering & Technology, 7 (2.28) (2018) 234-239

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Speech synthesis systems: disadvantages and limitations

Karolina Kuligowska[1]*, Paweł Kisielewicz [2], Aleksandra Włodarz[3]*****

[1]Department of Information Systems and Economic Analysis, Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland
[2]Profeosoft, Cracow, Poland
[3]Profeosoft, Cracow, Poland
**kkuligowska@wne.uw.edu.pl*
***profeosoft@gmail.com*
****aleksandra.rojek@profeosoft.pl*

Abstract

The present speech synthesis systems can be successfully used for a wide range of diverse purposes. However, there are serious and important limitations in using various synthesizers. Many of these problems can be identified and resolved. The aim of this paper is to present the current state of development of speech synthesis systems and to examine their drawbacks and limitations. The paper dis-cusses the current classification, construction and functioning of speech synthesis systems, which gives an insight into synthesizers implemented so far. The analysis of disadvantages and limitations of speech synthesis systems focuses on identification of weak points of these systems, namely: the impact of emotions and prosody, spontaneous speech in terms of naturalness and intelligibility, preprocessing and text analysis, problem of ambiguity, natural sounding, adaptation to the situation, variety of systems, sparsely spoken languages, speech synthesis for older people, and some other minor limitations. Solving these problems stimulates further development of speech synthesis domain.

Keywords: *Speech synthesis system; Speech synthesis limitations; Text-to-speech; TTS.*

1. Introduction

Speech synthesis involves the synthetic, artificial generation of human speech using computer technology. This area, defined as automatic process of converting written text into an acoustic speech signal [35] is also known as TTS (text-to-speech).

The historical approach to speech synthesis research indicates that the first systems, so-called têtes parlantes (fr. talking heads), appeared in the eighteenth century [19]. These pioneering efforts were focused on mechanical method for producing speech, however they constituted a very imperfect imitation of the human voice [12]. The evolution of speech synthesis lasted over the next centuries. Mechanical models were replaced by half-electric and electric models in the mid-twentieth century [15], [29], [28], [38]. At that time the two first approaches to simulation of the vocal tract resonance characteristic were formed, namely: articulatory synthesis and formant synthesis. With the development of information technology, in the 70s of the twentieth century, computer models of speech synthesis started to appear. Then emerged a third approach, significantly different from previous, known as concatenative synthesis [17].

The paper is organized as follows. Section 2 addresses the primary objective of our research, Section 3 presents a review of classification of existing speech synthesis systems, as well as describes their design and functioning. Section 4 analyzes drawbacks and limitations of speech synthesis systems. Finally, Section 5 presents our conclusions.

2. Research objectives

Modern technologies of the XXI century speech synthesis are based on complex methods and advanced algorithms. The most developed in 2010-2014, competing approaches to the synthesizing human speech are [6], [9], [14], [24], [43]: unit-selection synthesis, statistical parametric speech synthesis and hybrid methods of speech synthesis, that combine both previously mentioned approaches.

Despite such a variety of speech synthesis systems, they all face common limitations. Therefore in this paper we aim to present actual approaches to the synthesizing human speech and to examine thoroughly their drawbacks and limitations.

As a preliminary work, we examined over one hundred recent papers that contain the most important results of research in the design, construction and implementation of speech synthesis systems. Gathering this knowledge allowed us to diagnose weak spots of speech synthesis systems regardless of the chosen system design approach, and allowed to define issues to overcome in near future.

3. Classification, construction and functioning of speech synthesis systems

Text-to speech systems' functioning is based on the fact that the entered text is automatically converted into speech. The input constitutes written text in a digitalized form, and the output is a

synthetic speech. Two basic stages performed by the TTS system in order to synthesize speech are:

1. Analysis of the text (description of linguistic rules in the text)

2. Speech synthesis (production of speech sounds corresponding to the input text using the rules of linguistic description of the text)

In the first stage, linguistic rules determine not only how to pronounce individual words correctly, but also how to pronounce abbreviations, specialized terms, proper names, surnames etc. The process of linguistic analysis of the text is initiated by the NLP module (Natural Language Processing) [25]. Input sentences are decomposed into a list of individual words. Identified numbers, dates, abbreviations and acronyms are transformed into a full form (i.e. canonical). Then occurs a morphological analysis. To each word on the list all the possible names of parts of speech are assigned. All words are analyzed in their context. This allows to reduce the list of all possible parts of speech to a very restricted number of highly probable hypotheses, considering also the parts of speech of neighbouring words. In the last step, syntactic-prosodic parser determines the construction of the resulting text, which to the greatest extent refers to its expected prosodic realization.

In the second stage, the role of the synthesis algorithm is to simulate the action of the vocal tract system. Proper sounds of speech are generated and they represent the input text as a speech signal [22]. Therefore occurs automatic phonetization, i.e. automatic determination of the phonetic transcription of the input text. For this purpose is used LTS (letter-to-sound) conversion, which predicts pronunciation of words [39]. Then, through pro-sodic features, such as the pitch of the voice (varying between low and high), the length of sounds (varying between short and long), and loudness (varying between soft and loud), speech signal is implemented. Correct realization of melodic line is very difficult, because many different factors influence the prosody [36]:

- the meaning of sentence (neutral, imperative, question)
- feelings (happiness, sadness, anger etc.)
- speaker characteristics (gender, age, health etc.)

Finally, phonetic transcription and prosody obtained in the process of linguistic analysis are con-verted into acoustic wave of synthesized speech. For this purpose DSP (digital signal processing) module, known also as speech synthesizer, is used.

Knowing general operating principles of speech synthesizer, we will examine the structure of currently used and developed speech synthesis system. Their DSP modules will vary depending on the applied approach to speech synthesis. In general, currently used and developed speech synthesis systems can be distinguished into the following types: concatenative, statistical parametric and hybrid.

3.1. Concatenative speech synthesis

Concatenative speech synthesis concatenates (combines) individual units (phonemes, diphones, triphones, micro-segments, syllables) into a speech. It is divided into three main subtypes [14]:

1. Domain-specific synthesis

It is used in applications narrowed to a specific domain (e.g. speaking clock, speaking calculator, speaking weather forecast etc.). Such system is highly restricted by a vocabulary in its database and creates speech which is a combination of prerecorded words and phrases.

2. Diphone synthesis

It uses speech database which includes all the diphones of a given language – one recorded sample for each diphone (meaning the transition between two adjacent letters). Number of di-phones for each language can be radically different. For example Spanish has about 800 diphones and German has about 2500 diphones [36]. The target prosody of a sentence is a function modeled by using selected technique of digital signal processing, e.g. LPC (linear predictive coding), PSOLA (Pitch Synchronous Overlap and Add) or MBROLA.

3. Unit-selection synthesis (also known as corpus-based speech synthesis)

The biggest difference between unit-selection synthesis and di-phone synthesis is the length of speech segments. Unit-selection synthesis database keeps whole words and phrases. There-fore, it is many times larger than the diphone database. This causes that the system uses a large amount of memory while having a low central processing unit utilization.

The most effective, and in the same time the most popular type of concatenative synthesis, is unit-selection synthesis which in its segment database contains properly prepared corpora. It contains the recorded voice units of different lengths. Function called Cost Function is used to create an utterance. This function enumerates all the possible ways to generate a set of given expressions. The value of the cost function consists of target cost and join cost. Target cost measures how closely given unit is suitable for linguistic specification of the target sequence. Join cost checks the best way in which neighbouring units might be connected. Weighted of elements of the target cost, as well as cost optimization, have an effect on the synthesis quality [32].

3.2. Statistical parametric speech synthesis system

Statistical parametric synthesis is based on mathematical approach to generating speech signal. Statistical acoustic model is trained using context-dependent hidden Markov models. Modeled system is a Markov process with unknown parameters. The challenge here is to determine the values of hidden parameters based on observable parameters [8]. In this system, frequency spectrum (vocal tract), fundamental frequency (source voice) and prosody are a subject to statistical modeling.

One of the main advantages relating to the speech synthesis technique based on the hidden Markov models (HMM), in comparison with the unit-selection synthesis and concatenative synthesis, is that the change of speaking can be performed on the basis of the small size database and the quality of the synthesis is as good as in unit-selection and concatenative systems.

Although the speech generated by the concatenative system is virtually indistinguishable from the natural human speech, however, such system fails if the required segments are not in the basic database. This is due to the fact that even the largest corpora is not able to cover all the contextual variants of every speech segment. Such a strong dependence on data makes this approach very inflexible, since the characteristics of synthesized speech can be modified only by the construction of additional databases or using DSP algorithms which degrades the quality of synthesized speech. On the other hand, statistical parametric approach is not only able to generate speech signal which quality is comparable to the best unit-selection synthesizers, but also allows the synthesis of new segments and, almost unlimited - yet very effective, modification of the speech output characteristics. Therefore, speech synthesis technology based on the use of statistical models is gaining more and more recognition [10].

Comparison of the above-mentioned unit-selection and statistical parametric speech synthesis based on HMM is presented in Tab 1.

Table 1: Comparison of unit-selection and statistical parametric speech synthesis based on HMM

Unit-selection	HMM
Clustering (possible use of HMM)	Clustering (use of HMM)
Large run-time data	Small run-time data
Fixed voice	Various voices

Advantage: 1) High quality at waveform level Disadvantage: 1) Discontinuity 2) Hit-or-miss	Advantage: 1) Smooth 2) Stable Disadvantage: 1) Vocoded speech (buzzy)

Source: [6].

As we may observe on the Table 1, the unit-selection approach is based on clustering (with optional use of HMM), large run-time data and uses fixed voice, whereas statistical parametric speech synthesis based on HMM uses explicit HMM in clustering, small run-time data and allows various voices. The advantages and disadvantages of a chosen system flow from these differences. Therefore high quality at waveform level characterizes unit-selection approach, while smooth and stable voice defines statistical parametric speech synthesis based on HMM. On the other hand, unit-selection suffers from discontinuity of data and hit-or-miss method, while HMM exhibits vocoded (i.e. buzzy) speech.

3.3. Hybrid speech synthesis

Researches on the speech synthesis based on HMM indicates that it is possible to produce a speech signal based on a small amount of data used during the synthesis. Still, obtaining a high quality of natural voice is a challenge. On the other hand, unit-selection synthesis has proven that it is possible to recreate all the nuances and voice characteristics with sufficiently large database of resource materials. Unfortunately, the large size of database requires many hours of work on labelling and controlling results. Hence, the idea to combine HMM synthesis and unit-selection synthesis in one hybrid approach.

One example of a hybrid approach is the system resulting from a combination of synthesizer HTS-2007 (created at the University of Edinburgh by the Centre for Speech Technology Research in collaboration with Nagoya Institute Technology) with a commercial unit-selection synthesizer Cere Voice (created by Cereproc). It is an attempt to use strengths of both approaches in order to achieve more scalable tool able to mimic the voices of specific speakers. In this case, creators mimicked the voice of George W. Bush, the 43rd president of the United States. The advantage of this approach is that a large part of prosody can be produced within the unit-selection synthesis, while hybrid system can cope with the sparsity of data in many dimensions [2].

4. Drawbacks and limitations of speech synthesis systems

Numerous researches highlight a wide range of problems concerning speech synthesis systems. One of the fundamental limitations in the speech synthesis is generating correct prosody and pronunciation from text input. Written text does not contain any emotions, moreover the pronunciation of proper names and foreign words is sometimes very unusual. Speech synthesis is also difficult to generate in case of women's and children's voices. The female voice has a tone almost two times higher than the male voice, and in the case of children even up to three times higher. Estimation of formant frequency localization is more difficult with the higher fundamental frequency. There is also a number of problems associated with preprocessing of the text containing numbers, abbreviations and acronyms.

4.1. Emotions

The impact of emotions on interpersonal communication makes it necessary to take into account the emotional state as an integral part of human-computer interaction. If, for example, dialog systems could reliably determine that user is upset or angry, they could automatically switch to a potentially more appropriate mode of interaction. In research centres, it is estimated that expressive speech synthesis will play a key role in the believability and acceptance of future natural language interfaces. Adding emotions into synthesized speech requires that certain speech segments sound more joyful, soft or disrespectful, according to the communication situation. To achieve this goal, two problems should be solved: emotion must be identified on the basis of the input text, and appropriate signal changes must occur during the production of synthetic speech [7].

Currently, the most cost-effective commercial solutions, which convert text to speech, can produce neutral speech that in many cases is indistinguishable from human speech. It happens especially in certain types of applications, for example, in limited domain application scenarios such as synthesizing telephone numbers, speaking clocks or voice weather forecasting. However, there has been a lack of emotional affect in the state-of-art speech synthesis systems, and emotion simulation is not their (even optional) feature. This is mainly due to the fact that prosodic modules in these systems are not able to predict from the text prosody appropriate for emotional speech. Emotional speech has more prosodic variations then neutral speech [33]. Another reason for this is the complexity of human vocal expression: modern speech synthesis systems still face problems of generating understandable (for domain-independent systems), naturally sounding speech. One method of generating an emotional speech is based on unit-selection techniques that provide high quality speech synthesis. These techniques can deal with the emotional speech based only on the pre-included corpora, without any generalization concerning synthesizing specific emotions.

Emotions in synthesized speech constitute a complex domain in comparison to other research areas of language technology. The reason for this is the lack of standardized metrics to evaluate the emotional content of the speech sample. Automatic speech recognition can be evaluated easily by using objective results, such as error rate or phoneme words. There are no such parameters to classify the emotions expressed in synthesized speech. Moreover, people usually identify emotions subjectively, according to their temporary mood, opinion and cultural background. As a result, a sample of emotional speech may be perceived differently by every listener [26].

4.2. Prosody

Prosody is one of the key elements of the speech synthesizer that allows the implementation of complex psychological and phonetic effects [38]. They are necessary to express attitudes and emotions as a parallel channel in everyday communication. Prosody plays an important role in transferring a full communication experience between the speaker and the listener. Prosody determines how the sentence will be pronounced in terms of melody, phrasing, rhythm, accent and emotions. Very often it also may have a meaning, so it can help to distinguish the meaning of words even in the non-tonal languages. Prosody affects the naturalness and intelligibility of speech synthesis systems.

Quality of prosody is one of the main problems which face modern speech synthesis systems. Problems arise from prosodic bases (speech with a little presence of emotions) to the full range of nuances associated with the expression. While human reads specific text aloud, the listener collects contextual information. In real life prosody of the sentence is usually determined by the information presented few sentences earlier. However, current TTS systems are not able to make good use of such information. Therefore, synthetic speech lacks rhythm and other changes which man produces in a natural way [21].

Research on speech synthesis has brought significant improvements over the past decade that makes possible to generate natural speech from text. However, if the synthesized speech sounds

acoustically natural, it is often considered as not enough consistent with the human way of speaking. Therefore, now modeling the variability in the way of speaking (variations of prosodic parameters) is required to generate high quality expressive speech [28]. Despite growing attention to prosody modeling, one of the major drawbacks of actual prosody models is the monotony of the generated prosodic parameters. The prosody monotony is both related to poor dynamic as well as poor variability of the generated prosodic parameters. This causes the generation of stereotypical prosody, mainly due to the lack of linguistic knowledge extracted from the text [18].

Issue related to the prosody is the accent. Speech synthesis systems lack modeling of dialect variation. Although all synthesizers speak with a specific accent, there are still many of them that do not reflect what could be called as a "standard" accent of a specific language. The purpose is that synthesizers model an accent in any language. There are several practical reasons why systems should have this feature. One of them is the fact that people can change the accents and they often do that. Thus, speech synthesizers should also have this possibility in order to sound convincingly. The problem is that linguistics do not have much to offer in terms of the researches on the accent. Of course, there are large collections of recordings with accents of different languages and researches on pronunciation in a specific language. However, there is no such thing as theory of accent, which could form the basis for speech synthesis with many accents, and the ability to change them [34].

4.3. Spontaneous speech

The unit-selection synthesis simulates neutral speech quite well, both in terms of naturalness and intelligibility. For any application neutral ton is sufficient. However, there are some new applications in which speech synthesis plays an important role [9], [11], [22], e.g. in the dialog-oriented customer service applications, in navigation systems with traffic alerts or in educational systems. Such application should generate speech information about the attitude, the intention and the spontaneity observed in everyday conversations. In other words, we should simulate the way people speak, rather than the way in which they read.

Spontaneous speech exhibits many characteristics that are avoided in the current speech synthesis systems or they are poorly modelled:

- pronunciation variations (reduction, elision)
- lack of fluency (mispronunciations, hesitations, repetitions, repairs)
- voice quality and its amplitude variations (attitude, emotions)
- paralinguistic means of communication (laughter, sighs, breathing)

By using spontaneous speech that contains natural prosodic realizations of the above phenomena, it is possible to build speech synthesis which has natural conversational characteristics [3].

4.4. Preprocessing – text analysis

The first stage of speech synthesis systems is based on preprocessing, which is usually a very complex task, depending on the language. At the stage of preprocessing, the input text is converted into a sequence of words and symbols that will be processed by the rest of the system. This is called text normalization. Although the speech synthesis is an area where much attention is paid to the standardization of the text, dealing with the real text is a problem that also appears in other applications such as machine translation, speech recognition and detection of the topic of conversation. The most beneficial for speech synthesis system is the situation in which there is unambiguous relationship between spelling and pronunciation. But in real text, there are many unusual words: numbers, digit sequences, acronyms, abbreviations, dates. The main problem of text normalization module is converting non-standard words into regular words [27]. At this stage of preprocessing, system also identifies and makes decisions concerning punctuation, identifies and expands to a full from acronyms and numbers. The main problem in sentence segmentation, which is a part of preprocessing process, is the ambiguity of the period that is marking sentence boundaries or abbreviations. To determine the correct function of a period it is necessary to identify the acronyms and capital letters (proper names and beginning of the sentence). Difficulties arise from abbreviations which do not differ from normal sentence final words and from the fact, that proper names may appear on the beginning of a sentence [21].

4.5. Ambiguites

The problem of ambiguity exists in many different forms. The most basic is the ambiguity of the homographs, where the two words have different meaning but the same written form. Syntactic ambiguity also often occurs. All kinds of ambiguity cause additional problems in generating a good quality speech synthesis. Problems also arise with names, because people with the same name or surname can pronounce them differently, and the identification of such a name in the text is also a difficult task [30].

Ambiguity is often the result of a tension between opposing perceptions. It is this tension which can add to a user's curiosity and engagement. In a natural speech ambiguity is often used to create a specific effect, for example irony. This is about the utterance characterized by a intended contrast between obvious and intended meaning. One method used to generate irony is to use a contrasting emotion to the spoken content, for example sentence "What a wonderful day" said with an anger. Contrasting meaning and emotion in this way creates a complex picture of the speaker. It conveys more than straightforward utterance "What a horrible day". Current speech synthesis systems usually generate neutral speech. Researches on expressive speech synthesis that clearly communicate emotions, focus on evaluating a distinct set of emotional states, such as fear, anger, joy, surprise. This presents a problem for creating ambiguous utterance, because a very strong emotion in the voice will dominate the perception of the utterance. Therefore, more controlled approach is necessary which can offset other features in the utterance [4].

4.6. Naturalness

One of the most important tasks faced by researches on speech synthesis is to create a natural sounding synthetic speech system. Despite the fact that modern speech synthesis systems have reached a level of voice quality that no longer reminds robot-like voices, but rather a real human voice, various degradations still diminish the overall impression of the quality system [22]. Most synthesizers based on the PSOLA algorithm have artificial voices due to frequent concatenations of speech units. Synthesizers based on Hidden Markov Models can generate natural sounding voices, but also "noise" speech and the quality of choice of model units mainly depends on the size of the used corpus, on how well the units fit together and on how well this corpus fits to the text that is to be synthesized. All these impairments all sound differently, thus they degrade speech along different perceptual dimensions [13].

4.7. Adaptation of the system to the situation

Another issue is to adapt the system to the specific situation. In the usual communicative situation there is a possibility of reading aloud, reading silently or even having a conversation with other people. In the case of reading aloud, there are three important parts: the author of a text, the listener and the reader (i.e. the speech synthesis system). Most of the text was never written with the intention that it will be read aloud, and because of this, faithful

reading of a text can lead to situations where the reader is speaking something that the listener cannot understand, has no knowledge of or is not interested. When people take on the role of the reader, they often naturally stray from the literal text by adding some explanations, using paraphrases and synonyms to make the author's message understood. Very few systems that convert text into speech make any serious attempt at solving this issue [24], [35].

4.8. Disadvantages related to different types of systems

Concatenative, unit-selection speech synthesis systems rely heavily on the quality of the speech corpus used for building the systems. Creating speech corpora for this purpose is expensive and time consuming, so when the synthesized speech obtained is not as good as expected, it may be desirable to modify, correct or to update the corpus rather than record a new one. Usually corrections are limited to discarding mispronounced words or too noisy units [37]. Unit-selection speech synthesis simulates neutral speech quite well, both in terms of naturalness and intelligibility. However, when the speech corpus used for units selection does not provide good coverage, i.e. not every unit is seen in every possible context, there can be significant degradation in the quality of the synthesized speech [42].

The problems that arise in articulatory synthesis concern decisions how to find the right balance between very accurate model that closely follows human physiology and a more pragmatic representation that is easy to design and control [14].

4.9. Sparsely spoken languages

Speech synthesis systems for English and other well-researched languages use rich set of linguistic resources. Most often they have built-in modules such as: word-sense disambiguation, morphological analyzer, part-of-speech tagging and more. However, the minority languages are those which are not well-researched and often do not have enough linguistic resources. This involves many complications which appear when accumulating the text corpora in the digital format suitable for further processing. Linguistic components are not widely available in all languages of the world [23]. For this reason, building a good quality speech synthesizer for some languages is very difficult. What is more, the problem also arises when languages are purely spoken languages and they do not have standardized writing system. Also in this case, it is difficult to find appropriate corpora or linguistic resources. Such language could be language which is one of the official languages (e.g. in India) or dialect of a major language that is phonetically different from the standard language (e.g. in Egypt). Building speech synthesis system usually requires training which consist of a corpus with the proper transcriptions. However, in the case of languages that are not written in a unified manner, you can collect only corpus of prerecorded speech [31].

4.10. Speech synthesis for older people

More and more elderly people benefit from voice interfaces. They are also becoming more familiar with the computer technology, however they have problems with understanding the synthesized speech, particularly if they have hearing problems, and when they miss the contextual clues that compensate for weakened acoustic stimuli. Unfortunately, most of the research investigating potential reasons for these problems has not been carried out on unit-selection synthesis, but on formant synthesis. Formant synthesis lacks acoustic information in the signal and exhibits incorrect prosody. Since concatenative approaches preserve far more of the acoustic signal than formant synthesizers, lack of information should not be a problem anymore. Instead, there are problems with spectral mismatches between units with spectral distortion due to signal processing, and temporal distortion due to wrong durations [40].

4.11. Other limitations

It can often be seen that online speech synthesizers do not recognize special characters and symbols such as dot ".", question mark "?", or hash "#". Their databases usually contain only a few prerecorded voices that are used for synthesis. Modern software often leads to a different pronunciation of a particular text. What is more, there is a limit to the number of words for the input text that is going to be converted into speech [20].

5. Conclusion

Despite the fact that speech synthesis constitutes a dynamically developing technology, there are still some limitations in the currently developed speech synthesis systems. We have examined eleven weak points of speech synthesis systems implemented so far. Scope of the issues to improve in speech synthesis systems is very wide and includes: emotions, prosody, spontaneous speech, preprocessing and text analysis, ambiguities, naturalness, adaptation of the system utterances, disadvantages related to different types of systems, disadvantages associated with not commonly used languages, speech synthesis for older people, and finally some other limitations concerning special characters and symbols.

One of the most important features that needs to be improved concern natural sound of a synthetic speech system. Although the quality of speech generated by the concatenative systems is very good, however, such systems fail if the required segments of speech are not included in the primary database. This is due to the fact that even the largest corpora are not able to cover all variants of contextual segments of speech. Concatenative speech synthesis systems depend largely on the quality of the speech corpus used to construct these systems [17]. The creation of a comprehensive corpus for such a purpose is costly and time-consuming, because if the speech synthesis is not as good as expected, it is desirable to modify, improve or update the corpus, mainly by well-pronounced words or less noisy units [37].

On the other hand, synthesizers based on hidden Markov models can generate naturally sounding voices, but also a noisy speech [16], [41]. The performance and benefits of statistical HMM speech synthesis systems are impressive, but there are still some disadvantages associated with this approach [6]. Firstly, the parameters need to be automatically derived from the databases of natural speech. Secondly, parameters need to be lead to the high quality synthesis. Thirdly, the parameters must be possible to predict on the basis of the text. Hence the result of the application of statistical models of speech is understandable speech, but still it is not similar to natural human speech.

Identified by us weak spots of speech synthesis systems appear regardless of the chosen system design approach. We may observe that all speech synthesis systems face common limitations. Therefore hybrid methods of speech synthesis, that combine advantages and eliminate the disadvantages, are probably best suited here. Scope of the issues to improve in speech synthesis systems is very wide, however development of hybrid systems needs closer attention in the near future.

References

[1] Andersson S., Georgila K., Traum D., Aylett M., Clark R.A.J., "Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech for Unit Selection", *5th International Conference on Speech Prosody (Speech Prosody 2010)*, Chicago 2010, 1 – 2.

[2] Aylett M.P., Yamagishi J., "Combining Statistical Parameteric Speech Synthesis and Unit-Selection for Automatic Voice Cloning", *Proceedings of LangTech*, Rome 2008, 3.

[3] Aylett M. P., Potard B., Pidcock Ch.J., "Expressive speech synthesis: synthesising ambiguity", *8th ISCA Workshop on Speech Synthesis (SSW-8), ISCA*, Barcelona 2013, 217.

[4] Balyan A., Agrawal S.S., Dev A., "Speech Synthesis: A Review", *International Journal of Engineering Research and Technology IJERT*, Vol. 2, Issue 6, 2013, 57 – 75.

[5] Bellegarda J.R., "Toward Naturally Expressive Speech Synthesis: Data-Driven Emotion Detection Using Latent Affective Analysis", *7th ISCA Workshop on Speech Synthesis (SSW-7), ISCA*, Kyoto 2010, 200.

[6] Chandra E., Akila A., "An Overview of Speech Recognition and Speech Synthesis Algorithms", *International Journal of Computer Technology and Applications*, Vol.3, Issue 4, 2012, 1427.

[7] Chauhan A., Chauhan V., Singh G., Choudhary C., Arya P., "Design and Development of a Text-To-Speech Synthesizer System", *International Journal of Electronics and Communication Technology*, Vol. 2, Issue 3, 2011, 42 – 44.

[8] Demenko G., Wagner A. (eds.), *Speech and Language Technology*, vol. 14/15, Polskie Towarzystwo Fonetyczne, Poznań 2012, 32.

[9] Gruhn R.E., Minker W., Nakamura S., *Statistical Pronunciation Modeling for Non-Native Speech Processing, Signals and Communication Technology*, Springer-Verlag Berlin Heidelberg, 2011, 15 – 17.

[10] Hankins T.L., Silverman R.J., *Instruments and the imagination*, Princeton University Press, 1995, 186.

[11] Hinterleitner F., Norrenbrock Ch. R., Moller S., "Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech", *8th ISCA Workshop on Speech Synthesis (SSW-8)*, Barcelona 2013, 147.

[12] Indumathi A., Chandra E., "Survey on Speech Synthesis", *Signal Processing: An International Journal - SPIJ*, Vol. 6, Issue 5, 2012, 140 – 145.

[13] Kacprzak S., „Inteligentne metody rozpoznawania dźwięku", *master's thesis*, Wydział Fizyki Technicznej i Matematyki Stosowanej Politechniki Łódzkiej, Łódź 2010, 13.

[14] Kuczmarski T., "Overview of HMM-based Speech Synthesis Methods", [in:] Demenko G., Wagner A. (eds.), *Speech and Language Technology*, vol. 14/15, Polskie Towarzystwo Fonetyczne, Poznań 2012, 32 – 35.

[15] Nabożny A., „Przygotowanie korpusu do projektu korpusowego syntezatora mowy", *engineering thesis*, Wydział Inżynierii Mechanicznej i Robotyki Akademii Górniczo-Hutniczej, Kraków 2014, 14 – 18.

[16] Obin N., Lanchantin P., Avanzi M., Lacheret-Dujour A., Rodet X., "Toward improved HMM-based speech synthesis using high-level syntactical features", *Speech Prosody 2010 Conference Proceedings*, Chicago 2010, 2000.

[17] Ohala J.J., "Christian Gottlieb Kratzenstein: pioneer in speech synthesis", *17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong Kong 2011, 156 – 159.

[18] Padda S., Bhalla N., Kaur R., "A Step towards Making an Effective Text to speech Conversion System", *International Journal of Engineering Research and Applications*, vol. 2, issue 2, 2012, 1242 – 1244.

[19] Parssinen K., "Multilingual Text-to-Speech System for Mobile Devices: Development and Applications", *doctoral dissertation*, Tampere University of Technology, Tampere 2007, 18 – 42.

[20] Rabiner L.R., Schafer R.W, "Introduction to Digital Speech Processing", *Foundations and Trends in Signal Processing*, vol. 1, Issue 1–2, Now Publishers, 2007, 12 – 158.

[21] Raj A.A., Sarkar R., Pammi S.C., Yuvaraj S., Bansal M., Prahallad K., Black A.W., "Text Processing for Text-to-Speech Systems in Indian Languages", *6th ISCA Workshop on Speech Synthesis (SSW-6), ISCA*, Bonn 2007, 188.

[22] Raptis S., Chalamandaris A., Tsiakoulis P., Karabetsos S., "The ILSP Text-to-Speech System for the Blizzard Challenge 2012", *Proceedings of Blizzard Challenge 2012*, Portland, Oregon, USA 2012, 1 – 6.

[23] Rehm G., Uszkoreit H. (eds.), *The Polish Language in the Digital Age*, Springer 2012, 25.

[24] Saheer L., Potard B., "Understanding Factors in Emotion Perception", *8th ISCA Workshop on Speech Synthesis (SSW-8)*, Barcelona 2013, 59.

[25] San-Segundo R., Montero J.M., Giurgiu M., Muresan I., King S., "Multilingual Number Transcription for Text-to-Speech Conversion", *8th ISCA Workshop on Speech Synthesis (SSW-8)*, ISCA, Barcelona 2013, 65.

[26] Schroeder M., "Expressive Speech Synthesis: Past, Present, and Possible Futures", [in:] Tao J., Tan T. (ed.), *Affective Information Processing*, Springer, London 2009, 111 – 116.

[27] Schroeder M.R., "A Brief History of Synthetic Speech", *Speech Communication*, vol. 13, issue 1-2, Elsevier, 1993, 231 – 237.

[28] Schroeter J., "Text to-Speech (TTS) Synthesis", [in:] Dorf R. C. (ed.), *Circuits, Signals, and Speech and Image Processing*, CRC Press, 2006, 163.

[29] Sitaram S., Anumanchipalli G. K., Chiu J., Parlikar A., Black A. W., "Text to Speech in New Languages without a Standardized Orthography, Multilingual Number Transcription for Text-to-Speech Conversion", *8th ISCA Workshop on Speech Synthesis (SSW-8)*, Barcelona 2013, 95.

[30] Szklanny K., "System korpusowej syntezy mowy dla języka polskiego", [in:] *XI International PhD Workshop OWD 2009*, Conference Archives PTETiS, Vol. 26, Wisła 2009, 235 – 240.

[31] Tang H., Zhou X., Odisio M., Hasegawa-Johnson M., Huang T. S., "Two-Stage Prosody Prediction for Emotional Text-to-Speech Synthesis", *9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, Brisbane 2008, 1 – 2.

[32] Tatham M., Morton K., *Developments in Speech Synthesis*, Wiley, 2005, 143 – 144.

[33] Taylor P., *Text-to-Speech Synthesis*, Cambridge University Press, 2009, 1 – 50.

[34] Thakur B.K., Chettri B., Shah K.B., "Current Trends, Frameworks and Techniques Used in Speech Synthesis - A Survey", *International Journal of Soft Computing and Engineering IJSCE*, Volume 2, Issue 2, 2012, 444 – 445.

[35] Violante L., Rodriguez Zivic P., Gravano A., "Improving speech synthesis quality by reducing pitch peaks in the source recordings", *Proceedings of NAACL-HLT 2013*, Association for Computational Linguistics, Atlanta, Georgia 2013, 502.

[36] Wagner A., "A comprehensive model of intonation for application in speech synthesis", *doctoral dissertation*, Wydział Neofilologii Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań 2008, 129 – 137.

[37] Wang D., King S., "Letter-to-sound Pronunciation Prediction Using Conditional Random Fields", *IEEE Signal Processing Letters*, vol. 18, No. 2, 2011, 39.

[38] Wolters M., Campbell P., DePlacido C., Liddell A., Owens D., "Making Speech Synthesis More Accessible to Older People", *6th ISCA Workshop on Speech Synthesis (SSW-6), ISCA*, Bonn 2007, 1 – 2.

[39] Yang C.Y., Chen C.P., "A Hidden Markov Model-Based Approach for Emotional Speech Synthesis", *7th ISCA Workshop on Speech Synthesis (SSW-7), ISCA*, Kyoto 2010, 126 – 129.

[40] Yang Wang W., Georgila K., "Automatic Detection of Unnatural Word-Level Segments in Unit-Selection Speech Synthesis", [in:] Nahamoo D., Picheny M., (eds.), *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, ASRU 2011, 289.

[41] Zen H., Senior A., Schuster M., "Statistical parametric speech synthesis using deep neural networks", *Proceedings of 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver 2013, 7962.