Automatic word-level evaluation and error analysis of formant speech synthesis for Croatian

SANJA SELJAN[1], IVAN DUNĐER[2]
[1]Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb
CROATIA
[2]PhD student at the Department of Information and Communication Sciences
sanja.seljan@ffzg.hr, ivandunder@gmail.com

Abstract: - In the paper the evaluation of formant speech synthesis for Croatian is conducted in four domains: hotel reservation, insurance, automobile industry and weather forecast. Human evaluation is performed in order to evaluate quality of speech according to criteria of sentence comprehensibility, word intelligibility and correctness of word pronunciation. Automatic evaluation is conducted using word error rate metric across four domains and across specific lexical units (names, numbers, dates, general and special terminology). Correlation between human evaluation and word error rate metric is given by Pearson's correlation. The results are discussed and suggestions for further research mentioned.

Key-Words: - formant speech synthesis, word error rate, human evaluation, correlation, Croatian

1 Introduction

Automatic evaluation of speech synthesis is point of interest of various applications, often compared with human evaluation. Speech technologies, including speech recognition and speech synthesis, together with machine translation, belong to one of 10 emerging technologies, which would enable significant changes in communication [1].

Automatic evaluation of speech synthesizers, which could be used as stand-alone or as embedded solutions in various applications (e.g. in tutorials, in educational learning environment, in language learning software, in assistive technologies for impaired persons, in dialogue systems, on mobile devices, etc.) has been analyzed in numerous researches, mostly in single systems across different phases or among systems based on different technologies.

Formant synthesis method generates speech by attempting to imitate the time-varying formant frequencies of human speech. Resonances are produced in the vocal tract while a human speaks. These resonances, known as formants produce peaks in the energy spectrum of the speech wave.

The formant speech synthesis does not implement various speech components, such as natural sound, human voice, appropriate emphasize (accent) of words, chunking words into meaningful phrases, longer or shorter pronunciation of some words in certain sentence positions, breaks because of punctuation, intonation, etc. Still, it could have the practical implementation because of its suitability for voice quality and smooth transitions between segments, language independence and possibility to be integrated into various embedded systems. Such speech synthesis systems are especially valuable for less spoken languages with scarce languages resources.

Human evaluation is mostly subjective, time-consuming and costly, but considered to be a gold standard for evaluation. Automatic metric, aiming to approach as much as possible to human evaluation tends to be consistent, fast, low-cost and objective. The correlation between two metrics is usually computed using the Pearson's correlation coefficient.

The evaluation in this paper is performed on four domains of hotel reservation, insurance, automobile industry and weather forecast using word error rate (WER) automatic evaluation metric, which is then correlated with human evaluation. The findings and directions for future work are summarized in the conclusion.

2 Related work

Application and usage of speech synthesizers are widespread. Reference [2] indicated hot topics in speech synthesis: entertainment as major business area including applications in sport, music, art, etc.,

followed by education and training, especially in foreign language learning, customization of voice synthesizer by speaking with proper style, and programming for the specific purpose (e.g. in telephone answering machines), improvements in expressiveness and voice humanity when replacing everyday human voice (e.g. in sending messages, information services, games, customer-care), etc.

Speech synthesizers are often investigated in the range of Computer-assisted Language Learning (CALL) applications, as in spelling, transcribing activities, listening with comprehension and answering questions, talking dictionaries, etc. According to [3], speech synthesis systems may assume three different roles within Computer-Assisted Language Learning: reading machine, pronunciation model and conversational partner.

Reference [4] points out the need to build a mixed-language, text-to-speech synthesizer which could be used in embedded systems, but also to switch the language or speaker.

Reference [5] presented corpus-based concatenation text-to-speech system which can be used in applications of mobile telephony and portable computing using reduced lexicon and speech corpora, evaluated by on a five-point Mean Opinion Score or MOS scale.

Reference [6] describes audio processing methods in interaction with multimedia information when using palm devices. Reference [7] presented an implementation of natural language interface module in intelligent employment system, where text-to-speech synthesis provides employment information, namely in real-life environment.

Speech synthesis is often integrated with machine translation technologies. The only example for Croatian is TONGUES project presented by [8] targeted only for Croatian language.

The system integrated speech recognition and speech synthesis for English and Croatian, and translation system in both directions, with interface allowing active communication. In the paper presented by [9] the basic version of the system DIPLOMAT was developed for several languages (Croatian, Korean, Spanish and Haitian Creole) for conversation in the very restricted domain about non-military issues such as medical supplies, refugees, etc.

Evaluation of text-to-speech system has been point of interest of various researches, e.g. [10] analyzed an application of Bayes' theorem to evaluate synthetic speech in CALL, in the situation where a text-to-speech synthesizer would significantly reduce the time to prepare audio or video materials. The aim was to investigate whether synthetic sounds have the same effectiveness as natural sounds in enhancing listening skills of students.

Reference [11] presented evaluation of text-to-speech system by measuring similarity to human voice (naturalness) and intelligibility, by using Degradation Mean Opinion Score (DMOS) method measure on the scale 1-5.

In the paper presented by [12], the impact of tutor voice quality was evaluated, by comparing differences in student learning outcomes (measured by learning gains), system usability (measured by a survey) and dialogue efficiency using two-tailed t-tests of WER metric.

It was shown that students with higher word error rate metric use the system with synthesized voice more regularly, than the system with pre-recorded voice.

In order to evaluate text-to-speech synthesis in CALL, i.e. in the dictation exercise designed for students of Portuguese as a second language, 40 sentences were extracted from the student book [13]. In order to perform evaluation, 20 sentences were synthesized and compared with 20 taken from pre-recorded audio documents, with similar lexical content and difficulty level. In that research, the average word error rate (WER) was 34.1% and without considering orthographic accents 30.2%. The manual error categorization showed that the same types of errors were made in both sets. In the paper presented by [14] the experiment was made with crowdsourcing the assessment of the intelligibility of synthetic speech, by measuring the extent to which listeners can reproduce utterances produced by a given speech synthesis system, under different environmental and subjective factors, different language knowledge. WER scores are lower than in the laboratory situation.

In the research [15], in order to assess the quality and adequacy of the formant Croatian speech in different domains, the Mean opinion score (MOS) was used to evaluate the ***CroSS*** (**Cro**atian **S**peech **S**ynthesizer) tool. The best evaluation results were obtained for weather forecast domain, followed by hotel reservation. The worst result was obtained for automobile industry. When comparing specific terminology, the best results were achieved when synthesizing dates and numbers, and general terminology in weather forecast and hotel reservation domains. In insurance and automobile industry domains, general terminology is not well scored. The worst results were achieved for names in all four domains and special terminology in three domains, except in hotel reservation, having the best score for special terminology. The reason for this is

probably in human perception, not giving too much of attention in pronunciations of numbers and dates, whereas names always have the lowest scores. Among five criteria of appropriateness, comprehensibility, intelligibility, correctness of pronunciation and naturalness of speech, the best average scores were obtained for appropriateness, followed by comprehensibility of the sentence. Medium results were achieved for intelligibility of words, followed by correctness of word pronunciation. The worst results were obtained for naturalness of synthesized speech. The evaluation of domain suitability criteria showed that the domain of weather forecast was chosen as the most suitable by more than 80% of the evaluators, followed by hotel reservation and automobile industry being equally presented by less than 10% of evaluators, whereas insurance domain was not selected.

3 Tool CroSS - Croatian Speech Synthesizer

Formant synthesis, used in this experiment, does not use human samples of speech, but uses synthesized speech by using acoustic modeling, including parameters such as volume, pitch, pauses, speed and rhythm.

Although it produces robotic sounding utterances, it can still have its application, especially for not widely spoken languages, such as Croatian.

In the experiment the tool for formant speech synthesis is used, named CroSS - Croatian Speech Synthesizer. CroSS is a text-to-speech synthesizer based on formant synthesis.

It is capable of producing Croatian speech from corresponding text input and aims to enable better communication and accessibility for people with voice disorders, language impairments, reading disabilities and for Computer-assisted language learning.

CroSS is a Microsoft Windows desktop application written in C++ and synthesizes clear speech that can be used at high speeds. But it is not as natural as larger synthesizers which are based on human speech recordings.

CroSS was created in 2013 for the research purpose, using Microsoft Visual Studio 2012 and requires Visual C++ Redistributable for Visual Studio 2012 Update 1 and Microsoft .NET Framework 4 or higher to be run. It operates on Microsoft Windows 8 (x64) and Microsoft Windows 7 (x64). CroSS is based on eSpeak speech engine, which is a compact open source formant synthesizer and allows Croatian language to be provided in a small size [16].

The synthesized speech is clear and can be used at high speeds, but it is not as natural as larger synthesizers which are based on human speech recordings.

In order to produce appropriate prosody, such as pause at comma sign or a rising intonation in interrogative sentence, CroSS considers punctuation characters in a sentence. It incorporates technologies that can be useful in the process of learning and teaching languages and therefore can be applied in CALL environments.

The prosodic characteristics of synthesized speech can be investigated and analyzed in order to train and improve pronunciation or practice phonetic transcription.

4 Research

Research methodology and test set, evaluation criteria, results and discussion are given in the following subsections.

4.1 Methodology and test set

This work represents continuation of previously published article, presented in [15], where the focus was on domain-specific human evaluation of formant speech synthesis for Croatian language, relating to quality of synthesized speech, adequacy for public use and affective attitudes.

In this experiment, evaluation was performed on 20 formant-synthesized test sentences in four different domains:

- hotel reservation,
- insurance,
- automobile industry,
- weather forecast.

It was conducted at the Faculty of Humanities and Social Sciences, among 12 students enrolled in CALL course in April 2013. In each of the domains, the following lexical units were represented:

- names,
- numbers,
- dates,
- general terminology,
- and special terminology.

Preprocessing of textual input and preparing text for speech synthesis had to be performed manually, as the input was rarely structured, clean or unambiguous enough for this to happen directly [17]. Preprocessing tasks included the normalization of

- abbreviations,
- acronyms,
- cardinal numbers,
- dates,
- decimal numbers,
- nominal numbers,
- ordinal numbers,
- and special symbols [18].

CroSS was then used to import prepared test sentences and generate speech output audible on loudspeakers at the rate of 175 words per minute.

Human evaluators (12 students) that were sitting ca. half a meter in front of loudspeakers were asked to write down what they heard. Loudspeaker's output was measured with a sound meter to be cca 90 dB. Each sentence was repeated three times.

Table 1 presents test set description across four domains, each sentence containing 15 words.

Table 1 Test sentences statistics.

Domains	Hotel reservation	
	Insurance	
	Automobile industry	
	Weather forecast	
Sentences per domain	5	
Total sentences	20	
Words per sentence	15	
Total characters	Hotel reservation	484
	Insurance	521
	Automobile industry	559
	Weather forecast	502
Average characters	516,5	

4.2 Evaluation criteria

Detailed evaluation criteria are presented in [15]. In this research, the human evaluation of Croatian speech synthesis was performed using the following criteria, for each using Likert scale from -3 to 3:

- comprehensibility of the whole sentence,
- intelligibility or words,
- and correctness of pronunciation of words.

The human evaluation of mentioned criteria is correlated with WER (word error metric) automatic metric, widely used in speech technologies and machine translation systems. The Word Error Rate (WER) is based on the Levenshtein distance [19], which performs at character level and is based on misrecognized items on the word level.

WER is based on the minimum number of insertions, deletions and substitutions that have to be performed to convert the generated text (hypothesis) into the reference text. Every word in the hypothesized sentences is compared with reference sentence and every word which does not match (inserted, deleted or substituted) is counted as an error and divided by total number of words in the reference sentence.

The WER of the hypothesis *hyp* with respect to the reference ref is calculated, as in (1):

$$\mathrm{WER}=\frac{1}{N_{\mathrm{ref}}^{*}}\sum_{k=1}^{K}\min_{r} d_L(ref_{k,r}, hyp_k) \quad (1)$$

where $dL(ref_{k,r},hyp_k)$ is the Levenshtein distance between the reference sentence $ref_{k,r}$ and the hypothesis sentence hyp_k. In other words, the sum of lexical units which differ from lexical units in reference sentence (all substituted (S), deleted (D) and inserted (I) words) is divided by the total number of words in the reference sentence (N), as presented in (2).

$$WER = \frac{S + D + I}{N} \quad (2)$$

Although, the main disadvantage of WER is the fact that it does now take permutations of words into consideration, this does not affect the results, since evaluators wrote down exactly what they perceived. In this research Hjerson tool was used for word error rate calculation [20].

4.3 Results and discussion

Figure 1 presents average human evaluation per domain, using Likert scale (from -3 to 3) and indicating the highest scores for weather forecast (0.71), followed by hotel reservation domain (0.39). The worst result is obtained for automobile industry domain (-0.75).

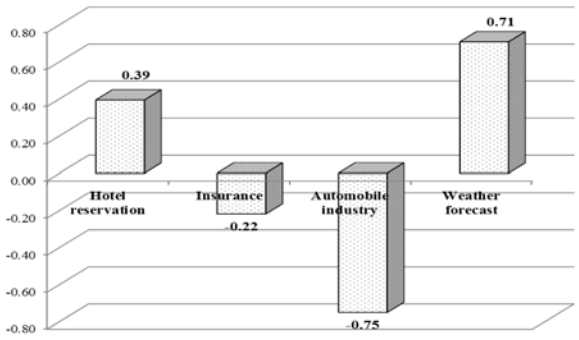

Fig. 1 Average human evaluation per domain.

Figure 2 presents scores of human evaluation, according to criteria of sentence comprehensibility, word intelligibility and correctness of word pronunciation, relevant for human perception. Using Likert scale from -3 to 3, the highest score is obtained for sentence comprehensibility in all domains, except in automobile industry. The highest scores for comprehensibility is obtained for weather forecast (1.02), followed by hotel reservation (0.87).

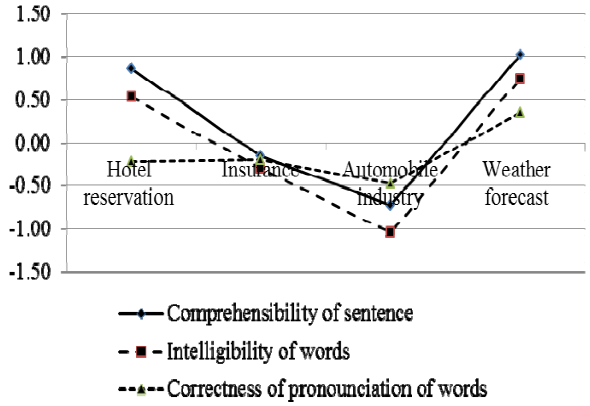

Fig. 2 Human evaluation across domains according to three criteria.

Table 2 shows the average human evaluation scores per domain, according to indicated criteria. The criterion of comprehensibility received the highest score. The other two criteria are in average negative. Intelligibility of words is positive for weather forecast and hotel reservation, whereas correctness of pronunciation of words is positive only for weather forecast. When comparing the three criteria across domains, the highest score is obtained for weather forecast (0.71), followed by hotel reservation (0.39).

Table 2 Average human evaluation scores per domain.
(Legend: H = Hotel reservation, I = Insurance, A = Automobile industry, W = Weather forecast)

	H	I	A	W	Average
Comprehensibility for whole sentence	0.87	-0.15	-0.73	1.02	**0.25**
Intelligibility of words	0.53	-0.30	-1.05	0.75	-0.02
Correctness of pronunciation of words	-0.22	-0.20	-0.47	0.35	-0.14
Average	0.39	-0.22	-0.75	**0.71**	

The following lexical units were analyzed in each of the twenty sentences in four domains: names, numbers, dates, general and special terminology.
Figure 3 presents average grades obtained for the specific lexical units. The highest grades are obtained for numbers (0.48) and dates (0.45) across four domains, and the lowest for specific terminology (-0.47) and names (-0.38).

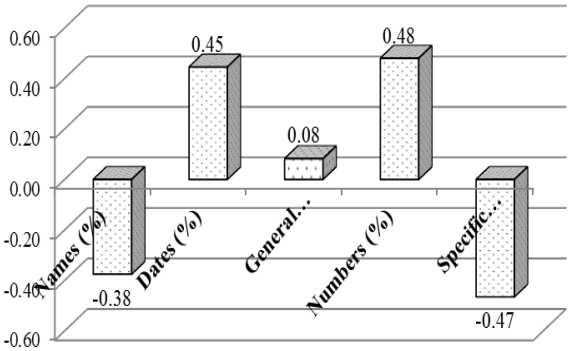

Fig. 3 Average human evaluation per analyzed lexical units.

Figure 4 presents average WER and human evaluation results per domain where higher word-error rate indicates lower quality and vice versa. The domain of hotel reservation having lowest WER was best scored, followed by weather forecast domain.

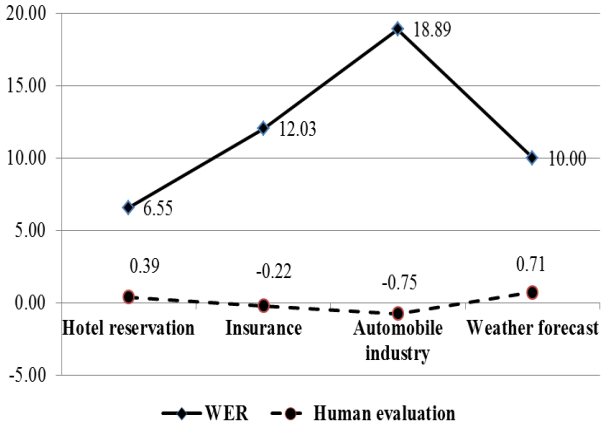

Fig. 4 Average WER (lower is better) and average evaluation per domain.

Figure 5 presents average WER and human evaluation scores per analyzed lexical units, indicating that the best result is the one having lowest WER scores, which are obtained for numbers, dates and general terminology.

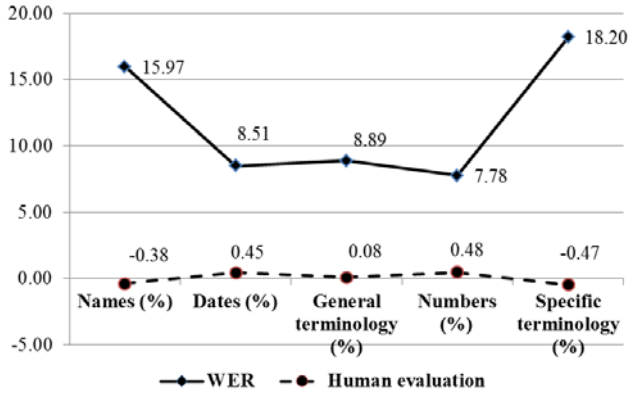

Fig. 5 Average WER and human evaluation per analyzed lexical unit.

Pearson's correlation, as shown in Table 3 indicates the best score for the domains of weather forecast and insurance, where there is the highest level of agreement between human evaluation and WER automatic metric, in evaluating them as the best and worst domain of synthesized speech.

Table 3 Pearson's correlation: Human evaluation - WER per domain.

(Legend: H = Hotel reservation, I = Insurance, A = Automobile industry, W = Weather forecast)

Correlation human evaluation - WER per domain	H	I	A	W
	-0.25	0.75	-0.65	-0.78

Pearson's correlation, as shown in Table 4, between human evaluation and WER range from -0.66 to -0.86 for all types of lexical units, except numbers.

The best correlation is obtained for general terminology (-0.86), followed by dates (-0.73) and specific terminology (-0.71).

Table 4 Pearson's correlation: Human evaluation - WER per lexical unit.

(Legend: N = names, D = dates, G = general terminology, No = numbers, S = Specific terminology)

Correlation human evaluation - WER per lexical unit	N	D	G	No	S
	-0.66	-0.73	-0.86	0.19	-0.71

5 Conclusion

The paper presents correlation of word-error rate and human evaluation for Croatian format speech synthesize. Human evaluation is performed according to criteria of sentence comprehensibility, word intelligibility and correctness of word pronunciation, with the highest score obtained for sentence comprehension, especially in the domain of weather forecast and hotel reservation, due to the fact that weather forecast domain is perceived as less subjective and more informative text and therefore more suitable for speech synthesis. Automobile industry received the worst scores by human evaluation.

WER automatic metric has shown the best results for the two domains, but in different order, scoring the best hotel reservation domain, followed by weather forecast.

The Pearson's correlation per domain gave the best results in evaluation of weather forecast domain as the most suitable (-0.78), followed by insurance domain (-0.75) and automobile industry (-0.65).

When analyzing specific lexical units, the worst result given by human evaluation is obtained for the specific terminology, followed by names which are always out of context. The best results are obtained for numbers and dates.

Pearson's correlation per lexical unit gave the best results for general terminology, dates and specific terminology.

The average score for Pearson's correlation across four domains is -0.61 and across all lexical units -0.55.

Human evaluation and WER automatic metric generally agree in selection of the two best domains (weather forecast and hotel reservation), and in selecting types of lexical units which are context-independent.

The main default of this research is relatively small test set, which could possibly be enlarged in the following research, or compared with diphone-based concatenative synthesis for Croatian.

References:

[1] S. Yamamoto, "10 Emerging Technologies That Will Change Your World", *Engineering Management Review - IEEE*, vol. 32 no. 2, 2004, pp. 32-51.

[2] G. Bailly, N. Cambell and B. Mobius, "ISCA Special Session: Hot Topics in Speech Synthesis", in *Proceedings of European Conference on Speech Communication and Technology*, 2003, pp. 37-40.

[3] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?", *Speech Communication*, vol. 51, no. 10, 2009, pp. 906-919.

[4] M. Malcangi and P. Grew, "Toward Language-independent Text-to-speech Synthesis," in

WSEAS Transactions on Information Science and Applications, vol. 7, 2010, pp. 411-421.

[5] J. Žganec Gros, "Text-to-speech synthesis for embedded speech communicators," in *Proceedings of 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2006, pp. 189-193.

[6] M. Malcangi, "Audio Interaction with Multimedia Information," in *Proceedings of 8th WSEAS International Conference on Computational Intelligence, Man-machine Systems and Cybernetics*, 2009, pp. 196-199.

[7] T. Šef, "An Employment Agent with a Speech Module", *WSEAS Transactions on Computers*, vol. 2, 2003, pp. 680-685.

[8] A. Black, R. Brown, R. Frederking, R. Singh, J. Moody and E: Steinbrecher, "TONGUES: Rapid Development of a Speech-to-Speech Translation System", in *Proceedings of 2nd International Conference on Human Language Technology Research – HLT* in San Diego, 2002, pp. 183-186.

[9] A. Black, R. Brown, R. Frederking, K. Lenzo, J. Moody, A. Rudnicky, R. Singh and E. Steinbrecher, "Rapid Development of Speech-to-Speech Translation Systems", in *Proceedings of International Conference on Spoken Language Processing* in Denver, States of America 2002.

[10] M. Kang, H. Kashiwagi, Y. Zou, K. Ohtsuki and M. Kaburagi, "An Application of Bayes' Theorem to Evaluate Synthetic Speech for Computer-Assisted Language Learning", in *WSEAS Int. conf. on Education and Educational Technology EDU '10*; Selected topics, 2010, pp. 253-257.

[11] Z. Orhan and Z. Gormez, "The Framework of the Turkish Syllable-Based Concatenative Text-to-Speech System with Exceptional Case Handling", *WSEAS Transactions on Computers*, vol. 7, 2008, pp. 1525-1534.

[12] K. Forbes-Riley, D. Litman and S. Silliman, "Comparing Synthesized versus Pre-Recorded Tutor Speech in an Intelligent Tutoring Spoken Dialogue System," in *Proceeding of 19th International FLAIRS (Florida Artificial Intelligence Research Society) Conference*, 2006, pp. 509-514.

[13] T. Pellegrini, Â. Costa and I. Trancoso, "Less errors with TTS? A dictation experiment with foreign language learners," in *Proceedings of INTERSPEECH 2012*, 2012.

[14] M. K. Wolters, K. B. Isaac and S. Renals, "Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk", in *Proceedings of 7th Speech Synthesis Workshop (SSW7)*, 2010, pp. 136-141.

[15] I. Dunđer, S. Seljan and M. Arambašić, "Domain-specific Evaluation of Croatian Speech Synthesis in CALL", in *Recent Advances in Information Science 2013 proceedings of the 7th European Computing Conference* in Dubrovnik, 2013, pp. 142-147.

[16] J. Duddington, "eSpeak text to speech". 2006, http://espeak.sourceforge.net/ (accessed in October 2012).

[17] U. D. Reichel and H. Pfitzinger, "Text Preprocessing for Speech Synthesis", *TC-STAR Workshop on Speech-to-Speech Translation*, 2006.

[18] D. Sasirekha and E. Chandra, "Text to Speech: A simple Tutorial", *International Journal of Soft Computing and Engineering*, vol. 2, no. 1, 2012, pp. 275-278.

[19] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", *Soviet Physics Doklady*, vol. 10, no. 8, 1966, pp. 707-710.

[20] M. Popović, "Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output", *The Prague Bulletin of Mathematical Linguistics*, vol. 96 no. 1, 2011, pp. 59-67.