Domain-Specific Evaluation of Croatian Speech Synthesis in CALL

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

Abstract: - Formant speech synthesis method mimics the time-varying formant frequencies of human speech and does not use prerecorded speech samples. In this paper, related work is discussed and an experiment is conducted using formant synthesis-based text-to-speech tool CroSS (Croatian Speech Synthesizer), in order to assess and evaluate the quality of synthesized Croatian speech, according to five criteria, then by domain suitability, affective attitudes and appropriateness of implementation in broader public use and in Computer-assisted Language Learning (CALL). The aim of integrating speech synthesis technology in Computer-assisted Language Learning resulted from the need to provide an interactive learning and teaching environment. This paper also addressed weaknesses and problems of Croatian language (e.g. input preprocessing of word classes) in the process of text-to-speech synthesis. The results are discussed and suggestions for further research mentioned.

Key-Words: - formant speech synthesis, Croatian domain-specific evaluation, Computer-assisted Language Learning

1 Introduction

Formant synthesis method synthesizes speech by attempting to imitate the time-varying formant frequencies of human speech. Resonances are produced in the vocal tract while a human speaks [1]. 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. It still could have 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.

Speech synthesizers can be used for various education purposes and in interactive educational applications (e.g. in tutorial systems) for impaired persons, or in the range of applications used in Computer-assisted Language Learning (CALL), e.g. in spelling and pronunciation teaching, transcribing activities, listening with comprehension and answering questions, reading aloud, etc.

Computer-assisted Language Learning (CALL) implements various computer applications in language teaching and learning [2] and embraces a wide range of Information and Communication Technologies and approaches.

Speech synthesis technology in Computer-assisted Language Learning has come out from the need to provide ideally interactive environment. According to [3] it has unique potential benefits, such as generation and editing of speech models, and various uses, e.g. talking dictionaries offering pronunciation of mostly headwords or in some cases whole phrases, talking texts, text dictation, pronunciation training and dialogue partner.

Speech synthesis can be integrated into learning environments which provide controlled interactive speaking practice outside the classroom [4]. Namely, speech synthesis systems may assume three different roles within Computer-Assisted Language Learning: reading machine, pronunciation model and conversational partner [5].

After the related work dealing with speech synthesis evaluation in CALL, the experiment is presented: test set description, evaluation criteria, tool and methods performed. Results are analyzed, followed by conclusion.

2 Related work

Formant analysis, used in this experiment, does not use human samples of speech at runtime. Instead, it uses synthesized speech by using acoustic modelling, 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. The research and the results on Croatian speech synthesis are presented in the paper by [6].

In the paper presented by [1] speech synthesis by diphone concatenation method is presented and evaluation performed on the criteria of quality, intelligibility, naturalness of sound and error frequency.

Speech recognition and speech synthesis are point of interest in language learning software, whose evaluation would be useful for scientists, industry, teachers, students and everyday users. [5] report on progress made in benchmarking of adequacy of speech synthesis in CALL in order to evaluate suitability and benefits of text-to-speech application.

[7] has described use of formant parametric synthesizer in laboratory assignments, i.e. learning activities in undergraduate courses in speech communication technology.

Evaluation of speech synthesizers is topic of interest of various speech software as presented by [8] predicting the following domains of implementation:

- entertainment as major business area including applications in sport, music, art, etc.,
- education and training, especially in foreign language learning,
- customization of voice synthesizer by speaking with proper style, emotions, accent and programming for the specific purpose (e.g. in telephone answering machines),
- improvement in expressiveness and voice humanity when replacing everyday human voice (e.g. in sending messages, information services, games, customer-care, etc.),
- use of syllable as basic unit of speech synthesis,
- evaluation of current speech synthesizers,
- interaction of engineering work and phonetic science with cognitive research and neuropsychological studies.

3 Experiment

The experiment was conducted at the Faculty of Humanities and Social Sciences, among students enrolled in Computer-assisted Language Learning course. The evaluation of Croatian speech synthesis was performed using the criteria of correctness of synthesized speech, usability in CALL and everyday life and their affective attitudes.

The evaluation is performed on 20 formant-synthesized test sentences in four different domains:

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

Evaluation was done by 12 philological graduate students (language and literature studies, history and linguistics) focusing on names, numbers, dates, general and special terminology in each of the twenty sentences.

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	

Table 1. Test sentences statistics.

In this research, the benchmarking criteria included viability and potential benefits of text-to-speech application in CALL and in everyday life, adequacy of use, potential implementation in Computer-assisted Language Learning programs and affective attitude. In this case, the experiment was divided into several segments:

- input preprocessing in form of text normalization (expansion of numerals, dates, abbreviations, etc. into text),
- dividing sentences into logical units by punctuation or spaces,
- synthesizing speech,
- conducting the survey,
- evaluation of results.

3.1 Evaluation criteria

Although various types of criteria are used, some appear more frequently. [5] used appropriateness, acceptability, accuracy and comprehensibility. In [9] the criteria of naturalness and intelligibility are pointed out as the most important ones in speech synthesis evaluation process.

[3] distinguished between two levels of readiness to use text-to-speech technology: acceptability or state of being prepared to use technology in various CALL applications representing "additional value" and adequacy of use comparing it with other media.

He also uses the following criteria in evaluation process: adequacy, acceptability and quality of the speech (comprehensibility, intelligibility, choice of pronunciation, precision of phonemes, appropriateness of prosody, naturalness of phonemes, naturalness of prosody, expressiveness, and appropriateness of register).

3.2 Tool

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 is 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 [10]. 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.

3.3 Methods

Preprocessing of textual input and preparing text for speech synthesis had to be performed manually, as the input is rarely structured, clean or unambiguous enough for this to happen directly [11].

Preprocessing tasks included the normalization of:

- abbreviations (km > *kilometar*, Eng. "kilometer"),
- acronyms (Zg > *Zagreb*, Eng. "Zagreb"),
- cardinal numbers (8:53 > *8 sati i 53 minute*, Eng. "7 minutes to nine"),
- dates (2013. > *dvijetisućetrinaeste*, Eng. "2013"),
- decimal numbers (1,5 > *1 i pol*, Eng. "one and a half"),
- nominal numbers (tb. 103 > *telefonski broj 1 0 3*, Eng. "telephone number 1 0 3"),
- ordinal numbers (3. > *treći*, Eng. "3rd") and
- special symbols (10.4€: *10 eura i 4 centa*, Eng. "10 euros and 4 cents").

All word classes were separated by spaces and transformed into full-textual form [12]. This kind of preprocessing is highly language and context dependent, due to the fact that word classes are pronounced differently in different situations.

All sentences were saved in UTF-8 format in order to avoid interoperability problems with CroSS and guarantee correct handling.

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

Human evaluators that were sitting cca. about half a meter in front of loudspeakers were asked to fill out a questionnaire for every single sentence after careful listening (three times) of generated synthesized speech. Every sentence was rated using the following criteria:

- appropriateness of the speech for the specific sentence including names, numbers, dates, general and special terminology in each of the twenty sentences,
- comprehensibility of the whole sentence,
- intelligibility or words,
- correctness of pronunciation of words,
- naturalness of synthesized speech.

For each criterion the Likert scale from -3 to 3 was used.

The following set of criteria related to:

- domain suitability (selecting one domain),
- adequacy for public use,
- affective attitude.

The criteria of adequacy for public and affective attitude use also used Likert scale from -3 to 3. Loudspeaker's output was measured with a sound meter to be cca. 90 dB. The goal was to obtain the evaluator's view of the quality and usability of the synthesized speech.

3.4 Results and discussion

In order to assess the quality and adequacy of the formant speech in different domains, the Mean opinion score (MOS) was used to evaluate CroSS tool. Figure 1 presents average results in four domains: hotel reservation, insurance, automobile industry and weather forecast. The best average result is obtained for weather forecast domain, followed by hotel reservation. The worst result is obtained for automobile industry domain. When comparing specific terminology, the best results are 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 are 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, while names always have the lowest scores.

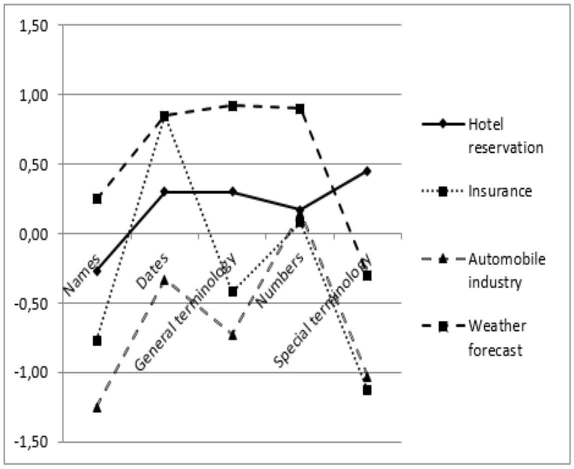

Fig. 1. Average scores in four domains per domain and specific terminology.

Figure 2 presents results of five criteria representing the quality of synthesized speech achieved for the domain of weather forecast, having the highest grades. Among five criteria of appropriateness, comprehensibility, intelligibility, correctness of pronunciation and naturalness of speech the best average scores are obtained for appropriateness, followed by the comprehensibility of the sentence.

Medium results are achieved for intelligibility of words, followed by correctness of word pronunciation. The worst results are obtained for naturalness of synthesized speech.

Comparing specific terminology the best score is obtained for dates, followed by numbers and general terminology. The worst results are scored for names and specific terminology.

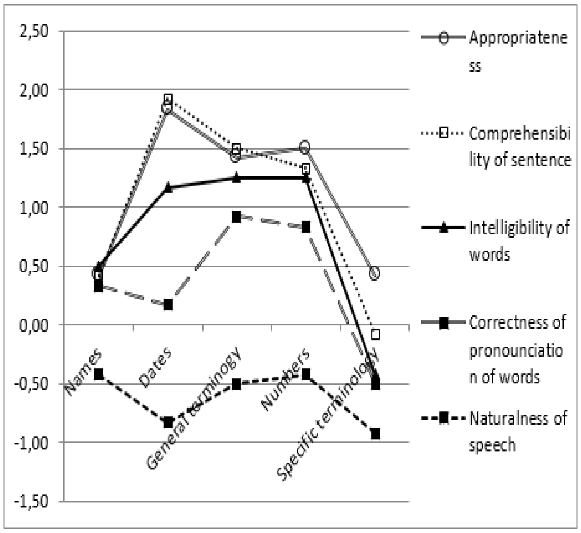

Fig. 2. Quality scores for five criteria in weather forecast domain.

The evaluation of domain suitability criteria shows that the domain of weather forecast was chosen as the most suitable by 83.33% of evaluators. Hotel reservation and automobile industry are equally presented by 8.33% of evaluators, while insurance domain was not selected (Fig. 3).

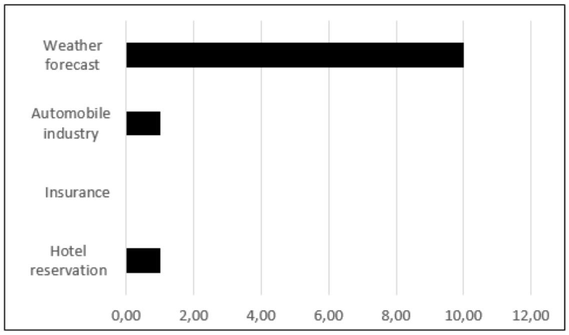

Fig. 3. Domain suitability of formant speech synthesis.

Figure 4 presents average values per domain and criteria. The best results are scored for weather forecast domain, described in Figure 2.

The second best results are given to hotel reservation domain for the criteria of comprehensibility followed by appropriateness and intelligibility of words. Negative results for all four domains are obtained for the criterion of naturalness of speech. Comprehensibility, intelligibility and correctness of word pronunciation are negatively score for automobile industry and insurance.

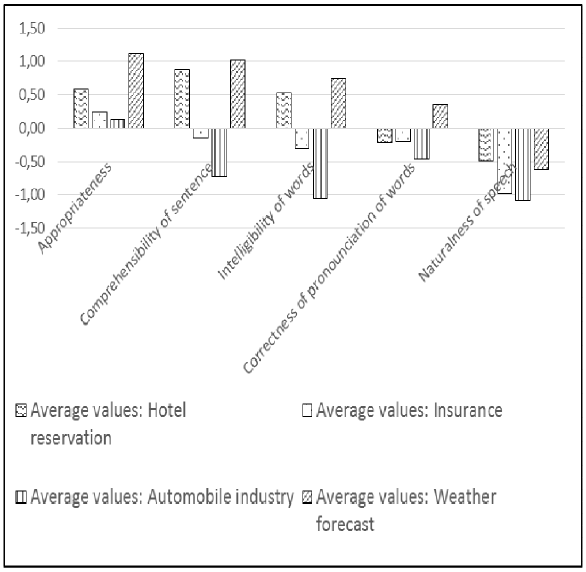

Fig. 4. Average scores in four domains per domain and criteria.

Figure 5 presents results of adequacy for broader public use of formant-synthesized speech and affective attitude of CALL students towards speech synthesis. Grades obtained for adequacy for broader public use range from mostly -1 to 3.

The most frequent grade is -1, followed by double less frequent 1 and by triple less frequent 0, 2 and 3. Grades for affective attitude range from -3 to 1.

The grade -1 is mostly represented followed by double less represented 1 and then followed by scores of -3 and -2. Average grade for affective attitude is -1.2 and average grade for adequacy for broader public use is 0.17.

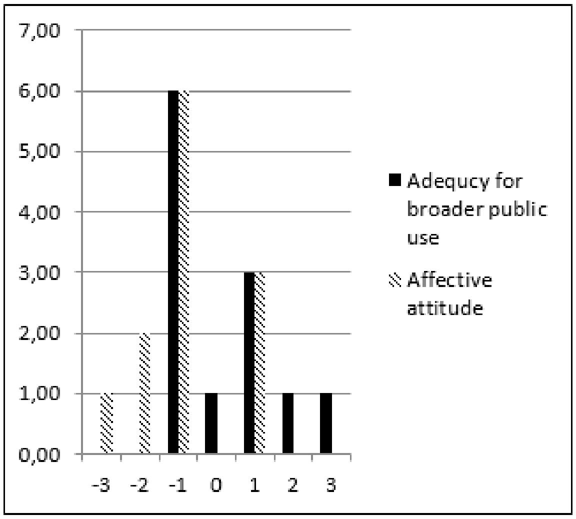

Fig. 5. Adequacy and affective attitude for formant speech synthesis.

Human evaluators were also asked whether they have had any experiences with speech synthesis before. 66.7% have not had former experience with speech synthesis, while 33.3% have already used it in dictionaries and online translation tools.

4 Conclusion

The paper presents evaluation results of formant synthesis-based text-to-speech tool for Croatian language. The experiment was conducted by Computer-assisted Language Learning students in four domains of hotel reservation, insurance, automobile industry and weather forecast. Evaluation was performed using five criteria to evaluate the quality and three criteria to evaluate adequacy and affective attitudes.

The best scores are obtained in the domain of weather forecast, which is perceived as objective, informative and the most suitable for formant speech synthesis. This domain is followed by ten times less scored domains of hotel reservation and automobile industry.

Among five criteria relating to quality the best scores are given to appropriateness and sentence comprehensibility, followed by intelligibility of words and correctness of pronunciation. Naturalness of speech has obtained negative results in all four domains.

The use specific terminology has shown the best results for dates, numbers and general terminology, where the human voice does not play the major role. Names and specific terminology are scored negatively since they require specific pronunciation and human-sounding prosody.

Average grade for affective attitude is -1.2 and average grade for adequacy for broader public use is 0.17. In all four domains the results are not evaluated as extreme (grades -3 or -2), but generally range from -1.5 to 2.

Although, the formant analysis is not perceived with high values, it still can have its implementation due to language independency and possibility to be integrated into various embedded systems, e.g. Computer-assisted Language Learning software used for spelling and pronunciation teaching, transcribing activities, listening with comprehension and answering questions or reading aloud.

The following research could possibly investigate the possibilities of CroSS tool implementation for weather forecast industry or bilingual language learning software.

References:

[1] M. Pobar, S. Martinšić-Ipšić, I. Ipšić. Text-to-speech Synthesis: A Prototype System for Croatian Language, *Engineering Review*, Vol. 28, No. 2, 2008, pp. 31-44.

[2] M. Levy. *CALL: Context and Conceptualisation*, Oxford University Press, 1997.

[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] F. Ehsani, E. Knodt. Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm, *Language Learning & Technology*. Vol. 2, No. 1, 1998, pp. 54-73.

[5] Z. Handley, M.-J. Hamel. Establishing a Methodology for Benchmarking Speech Synthesis for Computer-Assisted Language Learning (CALL), *Language Learning & Technology*, Vol. 9, No. 3, 2005, pp. 99-120.

[6] B. Damir, N. Lazić. Aspects of a Theory and the Present State of Speech Synthesis, *29th International Convention MIPRO: Computers in Technical Systems*, 2006, pp. 187-190.

[7] J. Beskow. A Tool for Teaching and Development of Parametric Speech Synthesis, *Fonetik - Swedish Phonetics Conference*, 1998, pp. 162-165.

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

[9] A. Chauhan, V. Chauhan, G. Singh, C. Choudhary, P. Arya. Design and Development of a Text-To-Speech Synthesizer System, *International Journal on Electronics & Communication Technology*, Vol. 2, No. 3, 2011, pp. 42-45.

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

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

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

[13] A. W. Black. Speech Synthesis for Educational Technology, *SLaTE Workshop on Speech and Language Technology in Education*, 2007.

[14] F. Hinterleitner, S. Möller, C. Norrenbrock, U. Heute. Perceptual Quality Dimensions of Text-to-Speech Systems, *InterSpeech: International Speech Communication Association*, 2011, pp. 2177-2180.

[15] M. Malcangi, P. Grew. Toward Language-independent Text-to-speech Synthesis, *WSEAS: Transactions on Information Science and Applications*, Vol. 7, No. 3, 2010, pp. 411-421.

[16] R. Sproat, J. Olive. *Text-to-Speech Synthesis*, in Digital Signal Processing Fundamentals, Ed. V. Madisetti, CRC Press, 1999.