RPKOM-GEN

A System for Testing Speech Recognition in Adverse Acoustic Conditions Using Speech Synthesis

Marián Trnka, Milan Rusko, Sakhia Darjaa, Róbert Sabo, Juraj Pálfy, Štefan Beňuš, Marian Ritomský, Martin Dravecký

Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

{marian.trnka, milan.rusko, utrrsach, robert.sabo, juraj.palfy, stefan.benus, marian.ritomsky, martin.dravecky}@savba.sk

***Abstract*—Training and testing of current state-of-the-art speech recognition systems require huge speech databases whose creation is time-consuming and expensive. This paper presents a novel approach for testing speech recognition in adverse acoustic conditions that uses speech synthesis, which facilitates optimizing and adjusting speech recognition to various environmental conditions. RPKOM-GEN is a complex system of multiple synthesizers that generates synthetic speech and testing signals with well defined characteristics. It might be used to produce public announcements, sets of utterances for spoken dialogue systems or other speech excerpts. The acoustic parameters of synthetic voices, such as speech rate, pitch, intensity, and others, can be pre-defined from a broad range of options. By using this novel technique, the system can also vary vocal effort imitating thus the Lombard effect and so-called long-distance speech. It is also possible to model the characteristics of the transmission channel since the system includes noise generators and digital effects such as the setting of environmental noise or reverberation levels. The paper presents the system architecture, describes graphical user interface and a rich array of usage possibilities, and discusses the results of pilot experiments testing the effect of added noise on speech recognition accuracy.**

Keywords-speech recognition; adverse conditions; noise; speech synthesis.

I. INTRODUCTION

One of the most demanding tasks in research and development of automatic speech recognition applications employed in situations with the transmission channel difficulties are the preparation, elicitation, annotation, and processing of speech databases. In an ideal case, high-quality databases should include speech from multiple and varied speakers recorded in real communicative situations under various physical (i.e., acoustic) conditions of the environment and the channel. The recording should be statistically representative in the sense that they should cover many combinations of factors under which a speech recognition system might be deployed. The variability of the factors, and consequently their combinations, is, however, enormous and creating a database that would cover all of them is, in effect, impossible.

One of the possible ways of approaching this problem is to substitute the recordings of real speech with synthesized acoustic signals, which allows imitating of various factors on the transmission channel with the help of synthesized signals and by introducing various effects through digitally processing the signal. This approach enables the creation of huge number of speech signals that can be used not only to verify understandability of a particular speech synthesis system in noise conditions but also to test the efficiency of a speech recognition system under various levels and types of noise present in the environment compounded with various characteristics of the transmission channel. Besides testing synthetic voices, the approach that exploits Text To Speech (TTS) synthesis also provides an option to set the characteristics of a particular synthetic voice, and imitate thus the changes speakers make in adverse acoustic conditions.

The paper contributes to the Activity 3.3 "Automatic speech recognition in adverse environments", which is included in the EU-funded project "Technology research for the management of business processes in heterogeneous distributed systems in real time with the support of multimodal communication" – RPKOM (acronym RPKOM is a short-hand for the project name in Slovak). The goal of the activity is to conduct applied research in automatic speech recognition for adverse acoustic environments and propose algorithms and architecture for systems capable of 1) generating announcements of public information systems and utterances of spoken dialogue systems that are reliably understandable by humans, 2) recognizing spoken instructions in noisy environments, 3) synthesizing speech that is optimized to achieve high understandability in highly noisy environments, and 4) being included in a speech recognition system for Slovak that is robust in dealing with acoustic environmental noise and varied characteristics of the transmission channel.

Many authors have been trying to find methods to improve intelligibility of speech synthesis in noisy and reverberant environments [1][2][3]. Noisy and reverberant environments represent an issue also for speech recognition [4][5][6]. We, therefore, decided to develop a tool that will be able to generate synthesized speech signals, mix them with various noises and apply reverberation. This tool will be used for experiments with speech synthesis and speech recognition in simulated adverse acoustic conditions.

In this paper, we start by presenting a simplified model of speech communication that informed the design of RPKOM-GEN in Section II. The system architecture is described in

 ISBN: 978-1-61208-342-1

Section III. Section IV sketches the design of graphical user interface. Pilot experiments testing the effect of adding various types and levels of noise on speech recognition accuracy are discussed in Section V. Section VI concludes the paper.

II. SIMPLIFIED MODEL OF SPEECH COMMUNICATION

A detailed description of spoken communication using a complex model with a range of influencing factors is beyond the scope of this paper; see, for example, [7]. For our purposes, a simplified model is sufficient for achieving functional solutions that can be implemented and deployed in real life. Such a simplified model is depicted in Fig. 1 and we will briefly discuss its components. We limit the discussion to the first two components – the speaker, the channel, and the factors influencing them – since they are most closely linked to the core of the proposed approach.

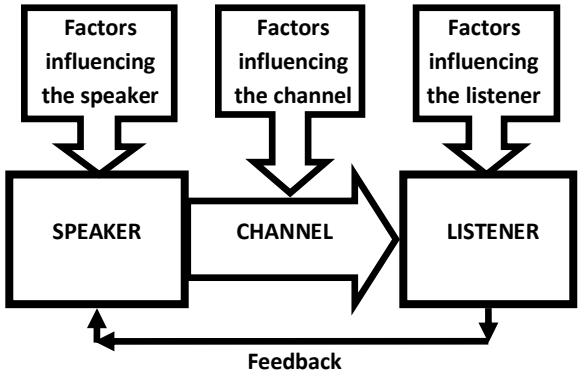

Figure 1. A schematic illustration of a simplified model of uni-directional transmission of information through the speech channel.

A. Speaker (Message sender)

The speaker sends the message through the acoustic signal produced by the articulatory speech production process. The characteristics of this signal are affected by multiple factors, some of which depend primarily on the speaker:

- Linguistic factors, such as the semantic content of the message, features of the lexicon, grammatical structuring, style, and others
- Paralinguistic factors, such as disfluencies, emotional states, and others
- Extra-linguistics, factors such as age, gender, speaker's health condition, and others.

B. Factors influencing the speaker

The speech signal is the primary carrier of information. Some external factors might induce changes in speaker's abilities, physical and mental conditions, or decision making processes. These factors then influence the final characteristics of the speech signal. In the RPKOM project, we focus mostly on the influence of external factors that increase speaker's stress.

C. Channel

In this paper, we consider the channel to represent the entire transmission process that the speech signal must undergo from speech generation by the speaker to speech decoding by the listener. We are thus concerned with the propagation of sound through air in some acoustic environment. Furthermore, we also include here the process of tracking the sound with a microphone, digitizing the analogue signal, coding, transmission of some tele-communication channel, such as internet cable, decoding, digital to analog conversion, and playback through speakers or headphones. Note that sound propagation through air is also involved when sound travels from the loudspeakers to the listener.

D. Factors influencing the channel

When the speech signal travels through the acoustic environment, is can be affected by the properties of barriers against which it bounces, or the presence and characteristics of background noise. Within the telecommunication channel, the quality of the signal is degraded by the noise of the channel itself and by the processes of digitizing and coding. There might also be signal distortions specific to a particular telecommunication transmission channel, such as delays or missing packets.

Our system covers primarily three types of acoustic signal degradation:

- Adverse influence of the acoustic environment such as acoustic bounces, reverberations, echo, and others
- Noise and non-speech sounds in the background such as pink, white, bubble, or cocktail-party noise
- Speech of other speakers

Other specific aspects of signal transmission such as microphone overload, specific factors of customer transmission channel, packet drop-outs, and others are left for future research.

III. RPKOM-GEN ARCHITECTURE

The architecture of the system is sketched in Fig. 2. The core of the RPKOM-GEN system is the *Signal Processing Unit* that has a functionality of mixing the speech signal with noise of various types while controlling for the signal-to-noise ratio. The database of noises (*Noise DB*) contains noise samples. The signal might be further distorted by adding the effects simulating adverse acoustic conditions such as *Delay*, *Reverberation*, *Echo*, and others.

The input of the signal processing unit is the speech signal that comes either from recorded human speech (*Speech Recordings*) or from synthesized speech produced by the *Speech Synthesis Unit*. This unit includes two types of synthesizers. The first one, *Unit TTS*, is based on corpus speech synthesis which selects and consequently joins the most suitable units of speech found in the speech corpus *Unit DB* [8][9]. Te second type of speech synthesis, *HMM TTS*, is based on Hidden Markov Models [10][11]. These statistic

 ISBN: 978-1-61208-342-1

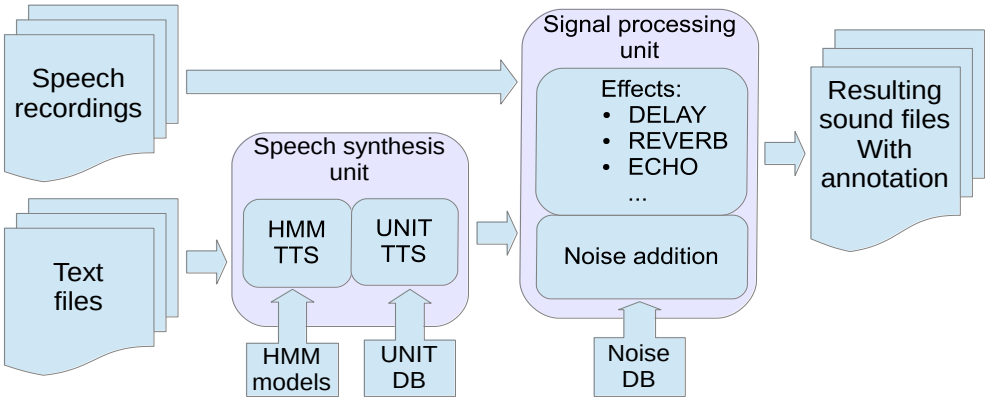

Figure 2. RPKOM-GEN system architecture.

parametric synthesizers produce the models of acoustic parameters (*HMM Models*). One of the differences between the two synthesizer types, that is relevant for this paper, is the control of speech effort, which is extremely important for synthesizing shouted speech and simulating the speech with the Lombard effect [12]. The Unit TTS offers the control of the speaker voice, the prosody model, primarily covering the rhythm and intonation of speech, mean fundamental frequency (F0), mean speech rate, and F0 range. The speech of HMM TTS voices, in addition to the above control parameters, also allows for manipulating speech effort exploiting our novel method for expressive speech synthesis [13]. Finally, the input for the text-to-speech synthesis is the set of pre-defined instructions and other texts contained in the database *Text Files*.

The system output is represented by *Resulting Sound Files* that contain detailed *Annotations* describing the content and the settings of all parameters applied during signal processing.

IV. User Interface

An example of graphical user interface design is shown in Fig. 3. The user interface is implemented in the Iron Python scripting environment. The user first selects whether real or synthesized speech should be used as the input.

In the latter, the user creates a *project* that collects all texts to be synthesized. The user then pairs each text with its own name of the testing voice. When prompted, the system reads in the project, lists all the testing voices (*List of tests*), and offers a list of available pre-defined synthetic voices (*Voice*). For each testing voice, i.e., for each *test*, the user selects from the available synthetic voices the one that is closest to his/her requirements and subsequently adjusts the individual parameters for signal processing in the Parameter panel (*Param*).

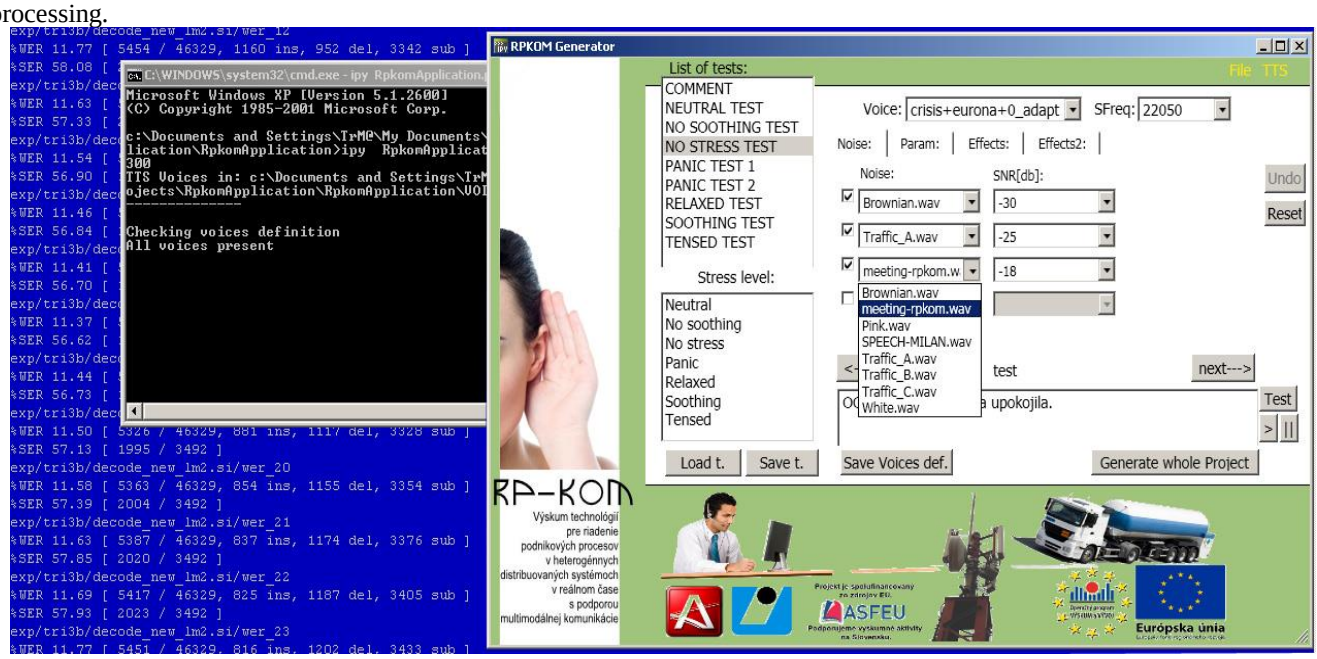

Figure 3. RPKOM-GEN system interface

Panel Noise offers a selection of noise types and the user might also control the resulting signal-to-noise ratio. Finally, the user might choose a type of a digital effect and the setting of its parameters in the Effects panel.

V. EXPERIMENTS

In this section, we present the results of pilot experiments testing the effect of added noise on speech recognition accuracy.

A. Methodology: adding noise

Depending on the setting of the parameters described in previous sections, mixing of noise components in the input signal is illustrated in Fig. 4. In the first step, the root mean square (RMS_S) value of the original acoustic signal is calculated using only those intervals that the *Voice Activity Detector* identifies as speech. The same method is applied for calculating RMS_N of the selected noise signal. Based on the ratio of the two RMS values and the selection of the required Sound-to-Noise Ratio (SNR), we calculate the coefficients for weighting the original and the noise signals. All processing of the signals is done on 32-bit-integer numeric fields to prevent overflow. Finally, the resulting signal is normalized to 16-bit.

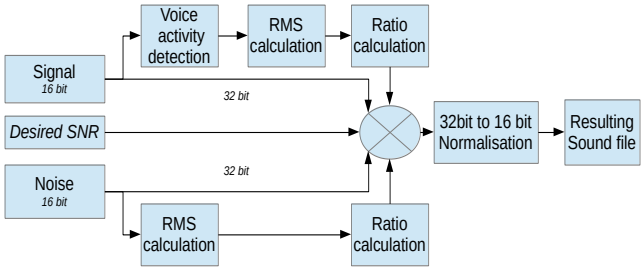

Figure 4. Noise addition scheme.

B. Results

We tested the possibility of assessing the effect of noise presence on the quality of Automatic Speech Recognition (ASR) using both synthetic and real speech. The reference sample of real speech consisted of the recording of 100 phonetically rich Slovak sentences collected for another project [14]. The same sentences were also generated using HMM TTS described in Section III above. Noise of four types (white, pink, brown, and traffic noise) and five SNR levels (-30 dB, -25 dB, -20 dB, -15 dB a -10 dB) were mixed with the all the original and synthesized speech signals. The resulting sentences then served as an input into our basic ASR system for Slovak [15]. The quality of recognition was assessed with a standard Word-Error Rate (WER) measure. The results are summarized Table I.

The table shows that brown noise deteriorates the speech signal the least, while the traffic noise distorts speech the most.

The results averaged for the type of noise are shown in Fig. 5. Two observations can be made. First, the degradation of ASR performance is non-linear. While increasing the noise levels between -30 and -20 dB results in rather moderate decrease in ASR performance, the last step produces a sharp decline in ASR performance.

TABLE I. WORD ERROR RATE (WER) RESULTS FOR DIFFERENT TYPES AND LEVELS OF ADDED NOISE

Test signal mixture		SNR					
		WER [%]					
Noise	Signal	clean	-30 dB	-25 dB	-20 dB	-15 dB	-10 dB
White	Human	11.0%	11.4%	13.9%	18.9%	27.2%	43.7%
	TTS	11.8%	12.1%	12.5%	16.0%	23.7%	48.9%
Pink	Human	11.0%	11.2%	14.1%	20.6%	28.7%	55.5%
	TTS	11.8%	12.0%	14.5%	18.5%	31.2%	65.7%
Brown	Human	11.0%	10.8%	11.0%	11.8%	15.2%	24.3%
	TTS	11.8%	10.6%	11.0%	12.1%	14.3%	21.0%
Traffic	Human	11.0%	14.1%	20.8%	25.6%	36.2%	64.6%
	TTS	11.8%	13.5%	16.6%	24.9%	44.9%	73.2%

Second, the results for human and TTS-produced speech are comparable with high correlation between them.

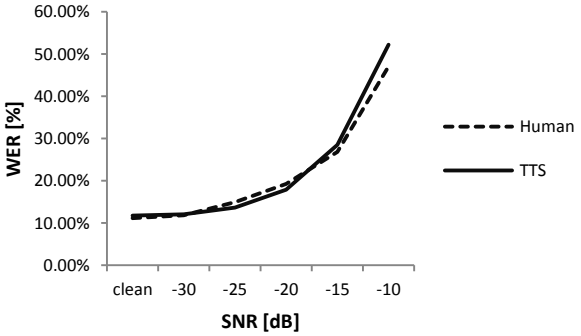

Figure 5. Comparison of Automatic speech recognition (ASR) performance in noise for human and synthesized (TTS) speech based on Word Error Rate (WER) for various levels of Sound to Noise Ratios (SNR).

This is an important observation since it provides a proof of concept that noise effects on these two types of speech result in comparable effects on understandability. This, in turn, will facilitate and accelerate significantly the production of test recordings for the evaluation of ASR systems in adverse acoustic environments.

VI. CONCLUSION

The paper outlined our work on a new system for generating speech samples that are suitable for testing the quality of speech recognizers deployed in adverse acoustic conditions. This system also facilitates parametric studies

 ISBN: 978-1-61208-342-1

and experiments with optimizing speech synthesis systems for high intelligibility in noise conditions or with distorting sound effects. The unique functionality of speech effort control allows simulating various vocal modes including shouted speech, Lombard speech, or long-distance speech. The user interface allows for fast online signal generation and the flexibility of the systems allows its implementation in various designing solutions.

ACKNOWLEDGMENT

This publication is the result of the project implementation: Technology research for the management of business processes in heterogeneous distributed systems in real time with the support of multimodal communication, ITMS 26240220064 supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

[1] A. W. Black and B. Langner, "Improving Speech Synthesis for Noisy Environments," Speech Synthesis Workshop 7 (SSW7), Kyoto, Japan, 2010, pp. 154-159.

[2] M. Cerňak, "Unit Selection Speech Synthesis in Noise," Proc. of ICASSP06, Toulouse, France, May 14-19, 2006, pp. 14-19.

[3] R. Vích, J. Nouza, and M. Vondra, "Automatic Speech Recognition Used for Intelligibility Assessment of Text-to-Speech Systems," Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, Springer 2008, pp. 136-148.

[4] R. H. Wilson and W. B. Cates, "A Comparison of Two Wordrecognition Ttasks in Multitalker Babble: Speech Recognition in Noise Test (SPRINT) and Words-in-Noise Test (WIN)," Journal of the American Academy of Audiology, vol. 19, no. 7, 2008, pp. 548-556.

[5] L. Couvreur and C. Couvreur, "Robust Automatic Speech Recognition in Reverberant Environments by Model Selection," Proc. of the International Workshop on Hands-Free Speech Communication, Kyoto, Japan, 2001, pp. 147-150.

[6] T. Yoshioka et al., "Making Machines Understand us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 114-126.

[7] J. H. L. Hansen et al., The impact of speech under 'stress' on Military Speech Technology, NATO RTO-TR-10, AC/323(IST)TP/5 IST/TG-01, 2000.

[8] A. D. Conkie, "Robust Unit Selection System for Speech Synthesis," Joint Meeting of ASA, EAA, and DAGA, paper 1PSCB_10, Berlin, Germany, 1999.

[9] S. Darjaa et al., "HMM Speech Synthesizer in Slovak," 7th International Workshop on Grid Computing for Complex Problems (GCCP), Bratislava, Slovakia, Institute of Informatics SAS, 2011, pp. 212-221.

[10] H. Zen et al., "The HMM-based Speech Synthesis System Version 2.0," Proc. of ISCA SSW6, Bonn, Germany, 2007, pp. 294-299.

[11] M. Rusko, M .Trnka, and S. Darjaa, "Three Generations of Speech Synthesis Systems in Slovakia," Proc. of the XI. International Conference SPECOM 2006, St. Petersburg, Russia, 2006, pp. 449-454.

[12] J. C. Junqua, "The Influence of Acoustics on Speech Production: A Noise-induced Stress Phenomenon Known As the Lombard Reflex," Speech Communication, vol. 20, no 1-2, 1996, pp. 13-22.

[13] M. Rusko, S. Darjaa, M. Trnka, and M. Cerňak, "Expressive Speech Synthesis Database for Emergent Messages and Warnings Generation in Critical Situations," Language Resources for Public Security Workshop (LRPS 2012) at LREC 2012 Proceedings, Istambul, 2012, pp. 50-53.

[14] O. Jokisch et al., "Multilingual Speech Data Collection for the Assessment of Pronunciation and Prosody in a Language Learning System," SPECOM´09, 13-th International Conference on Speech and Computer. Editor A. Karpov, Russian Academy of Science, St. Petersburg Institute for Informatics and Automation, State University of Aerospace Instrumentation, 2009, pp. 515-520.

[15] M. Rusko et al., "Slovak automatic transcription and dictation system for the judicial domain," Human Language Technologies as a Challenge for Computer Science and Linguistics, 5th Language & Technology Conference, Poznań, Fundacja Uniwersytetu Im. A. Miczkiewicza, 2011, pp. 365-369.

 ISBN: 978-1-61208-342-1