



The Sixteenth International Congress on Sound and Vibration

Kraków, 5-9 July 2009

DIAGNOSTIC EVALUATION OF SYNTHETIC SPEECH USING SPEECH RECOGNITION

Miloš Cerňak, Milan Rusko and Marian Trnka

Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
e-mail: Milos.Cernak@savba.sk

The paper presents experiments on the use of automatic speech recognition for diagnostic evaluation of synthetic speech. Our previous work on the topic showed a strong correlation between the subjective and objective evaluation (ITU-T Rec. P.862 PESQ) of the quality of synthetic speech. The main drawback of the approach was the need for original human (reference) recordings in one to one mapping for each test token. Trying to overcome this problem, we now create HMM models from reference recordings, and use speech recognition of the synthetic speech to predict its intelligibility. Having developed a testing vocabulary of Slovak Diagnostic Rhyme Test, the prediction provides us also with a diagnostic evaluation of the synthesis. Measurements with the second diagnostic vocabulary designed for Slovak audiometry are presented as well. The paper presents the speech recognition technique, the correlation of the resulting subjective mean opinion scores from listening tests and objective MOS scores calculated for our male unit-selection voice. Results on different additive noise levels are presented. The main purpose of the new measure is to facilitate synthetic speech evaluation at the design state of system development providing a rapid and repeatable synthetic voice quality measurement technique.

1. Introduction

Although technology advances of last decade research and development on unit-selection speech synthesis resulted in high quality synthetic voices, there is still a lack of objective evaluation techniques of the voices, which would provide a rapid and repeatable measures. Subjective evaluation is considered as the only measure with the sufficient level of test outcome confidence.

Our previous work on the topic¹ showed a strong correlation between the subjective and objective evaluation (ITU-T Rec. P.862 PESQ) of the quality of synthetic speech. The main drawback of the approach was the need for original human (reference) recordings in one to one mapping for each test token. Trying to overcome this one to one mapping, we now create HMM models from reference recordings, and use speech recognition of the synthetic speech to predict its intelligibility. Having developed a Diagnostic Rhyme Test in Slovak, we synthesize the word pairs, evaluate the recognition rate, and we expect that it will provide us with a diagnostic evaluation of the synthesis. Measurements with the second diagnostic vocabulary – word test set designed for audiometry in Slovak are presented as well. Similar approach was already presented for Czech language², however the correlation subjective tests with speech recognition results was not shown in the study.

The paper presents the speech recognition technique, the correlation of the resulting subjective intelligibility scores from listening tests and ASR scores calculated for our male unit-selection voice. Results on different additive noise levels are presented.

The structure of the paper is the following. Next section 2 introduces diagnostic vocabularies that are used in this work. Section 3 describes ASR setup and Section 4 listening tests. Section 5 presents results and finally Section 6 concludes the paper.

2. Diagnostic vocabularies for Slovak language

There are two vocabularies described in this section, which we used in our work.

2.1 Slovak word test for audiometry

The set of test words for Slovak word audiometry (TWA) was designed by Bargár et. al. in 1986³. The authors claim that the selection of this set of 100 words organized in groups of 10 was made so that each group of words (decade) represents the entire language (Slovak) as well as possible from the linguistic and phonetic points of view (vowel formants, mono- and poly-syllabic words, different parts of speech, phoneme representation etc.). Every decade is of the same representativeness and of the same importance for the test. The reprint of the word set can be found in¹.

2.2 Slovak Diagnostic Rhyme Test

Slovak Diagnostic Rhyme test word lists were constructed recently⁴. It uses monosyllabic words that are constructed from a consonant-vowel-consonant sound sequence. In the DRT set, words are arranged in ninety-six rhyming pairs which differ only in their initial consonants differing by only one distinctive feature. One of the words in a pair starts with a consonant characterised by certain feature, and the second one by a consonant with a contrast feature. Listeners are shown a word pair, and then asked to identify which word is presented by the talker (reproduced). Carrier sentences are not used.

3. Automatic speech recognition of the synthetic speech

We constructed two tasks for our ASR system. In the first task, called *audiometry*, we performed isolated word recognition with audiometry vocabulary size of 100 words. In the second task, called *DRT*, we performed isolated word recognition as well, with dynamic vocabulary in size 2 words, which depended on a word pair in the test. Words in both tests were synthesized by our male unit-selection voice.

Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) based speech recognition system was trained using public domain machine-learning library TORCH on the training set that consisted of 1500 phonetically balanced utterances (TTS database of the unit-selection voice). Three states left-right HMM models were trained for each of the 54 phonemes in the TTS DB including silence as well. 16 Gaussian model mixtures were used in all experiments. Diagonal covariance matrices were used to model the emission probability densities of the 39 dimensional feature vectors - 13 cepstral coefficients and their derivatives (deltas) and double derivatives (double deltas). The phoneme HMMs were connected with no skip. We trained and tested the ASR system with MFCC feature set. All the features were calculated using HTK hcopy tool. We calculated MFCC vectors every 10 msec using windows of size 25 msec.

Our tests set consisted of 3 tests: the clean speech, the speech with additive white noise at SNR 25dB, and the speech at SNR 20dB. There were two reasons for that. Firstly, we recently worked on unit-selection speech synthesis in noise⁵, and we were interested how intelligible is our current voice in a noisy environment. Secondly, we planned to make a correlation between subjective tests (see next Section 4) and ASR results, and we needed evaluation on several levels.

ASR results are presented in next section, together with subjective evaluations.

4. Listening tests

Listening tests were carried out in order to evaluate the Slovak male unit-selection based speech synthesis developed at our department. We tested the intelligibility of the voice, in order to compare the results with accuracies of ASR system.

The subjects taking part in listening tests belong to the normal PC using population, with the provisos that:

- a) they have not participated in any subjective test whatever for at least the previous six months, and not in any listening-opinion test for at least one year;
- b) they have never heard the same word lists before.
- c) 19 subjects (10 males and 9 females) aged from 19 to 64 took part in the experiment .

A group of 100 words of TWA vocabulary and 347 words of DRT vocabulary were synthesized, while white noise was added at SNR 25 dB to 1/3 of the word group and white noise at SNR 20 dB to the next 1/3 of the word group. This was done to keep time feasibility of the test performance. The stimuli were played from the PC to the test participant via Sennheiser PX200 closed-system headphones in laboratory conditions.

Tests were written in files formatted in XML format, which were processed by our internal test evaluation software. TWA subjective tests were performed as open tests, where participants wrote what they heard. While performing DRT subjective tests, participants could choose between two presented words, the word pair defined by DRT test. Participants used the evaluation software by themselves after brief introduction.

5. Results

Both ASR and cumulative subjective test results are presented here.

5.1 Audiometry

Figure 1 presents both ASR accuracies and subjective tests scores for TWA vocabulary. We calculated Pearson's correlation coefficients for all accomplished tests according to the following formula⁶:

$$r \equiv r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} \quad (1)$$

where n is the number of measures, x and y , and s_{xy} is the sample covariance x and y , computed as

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n-1). \quad (2)$$

The Pearson's correlation coefficient for TWA vocabulary was $r = 0.975$. We used a significance test to determine the probability that the observed correlation is not achieved by chance. A one-tailed test was chosen, since we know the direction of the relationship between the PESQ and the MOS scores. We computed the t statistics using⁶:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (3)$$

where n is the length of each test block. Once we had the t statistics, we referred to Student's t distribution table to find the significance of the test.

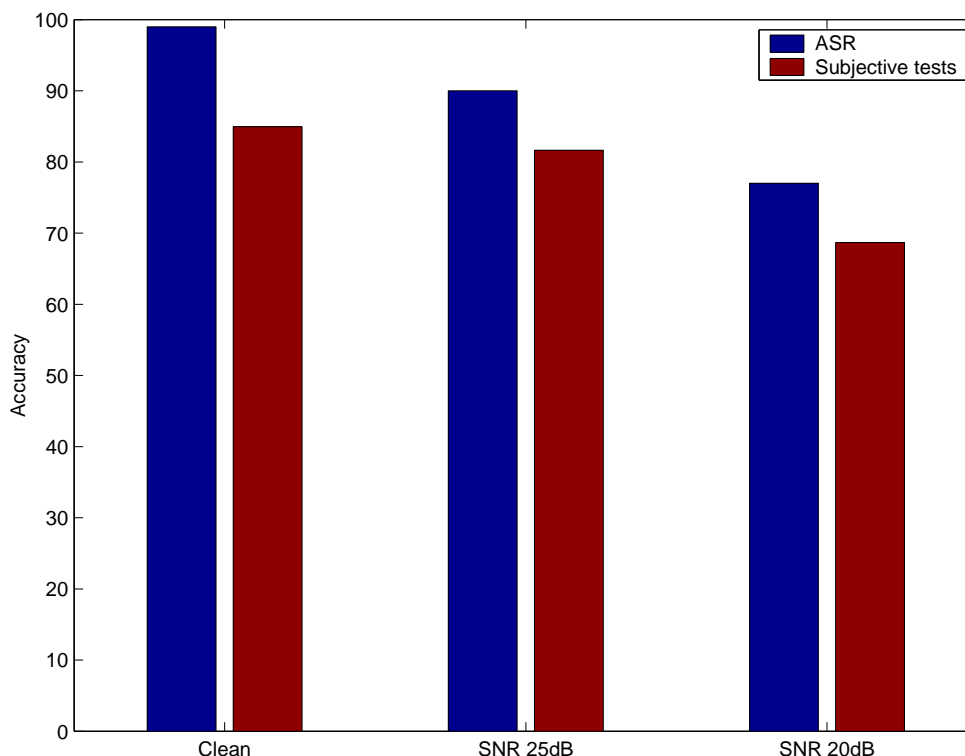


Figure 1. Accuracy of ASR and subjective listening tests for TWA vocabulary. The higher accuracy of ASR was achieved using closed 100 words vocabulary, while subjective tests were organized as the open test.

The Pearson's correlation coefficient for the voice is above the 10% significance level, since we had just 3 measures of ASR and subjective tests. However the trend is clear and we can predict that adding another measure would decrease the significance level.

5.2 DRT

Figure 4 presents both ASR accuracies and subjective tests scores for DRT vocabulary. Comparing to subjective test scores we can see significant accuracy drop for noisy stimuli. This is reflected also by the Pearson's correlation coefficient, which for DRT vocabulary is $r = 0.573$.

5.2.1 Analysis in terms of DRT categories

There are six DRT phonological distinctions (categories):

- Voicing (Vc)
- Nasality (N)
- Sustenation (O)
- Sibilantion (S)
- Graveness (A)
- Compactness (D)

ASR performance split over DRT distinctive features are shown in Figure 2, while Figure 3 presents split subjective tests. The overall correlation Pearson's correlation coefficient was quite low, but results for graveness phonological feature show high correlation $r_A = 0.989$. The correlation is above the 5% significance level.

We encountered during DRT evaluation some of the following problems. As we didn't evaluate all DRT word pairs in all noise versions, and the test stimuli were generated randomly, we got inconsistent results in some categories (see e.g. Fig. 3 categories D, S and O). Intelligibility of words with higher noise level was evaluated more intelligible than words with lower additive noise or without noise at all. Some of the words were also synthesized with lower quality, and caused significant accuracy drop for both listening test and ASR.

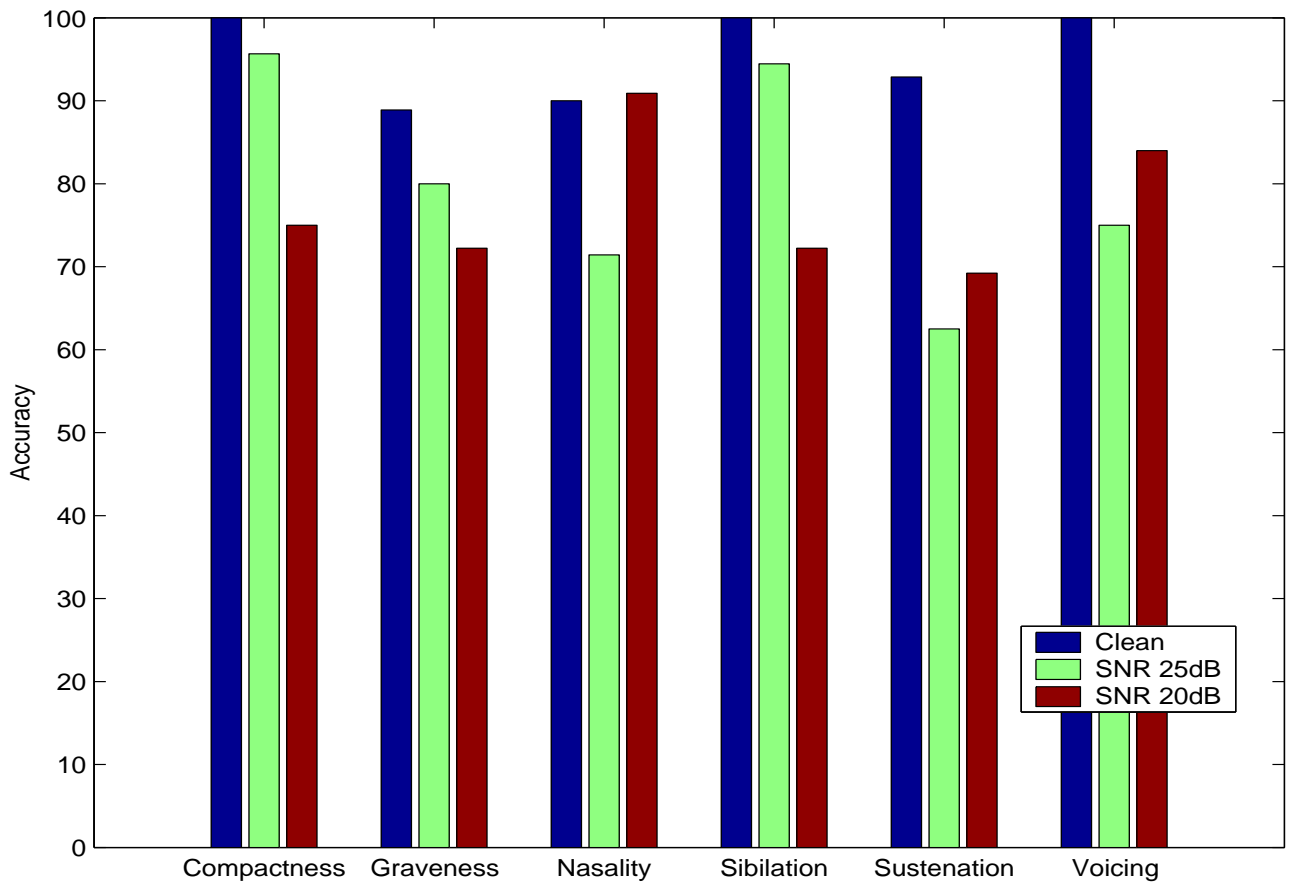


Figure 2. ASR accuracy split over DRT distinctive features.

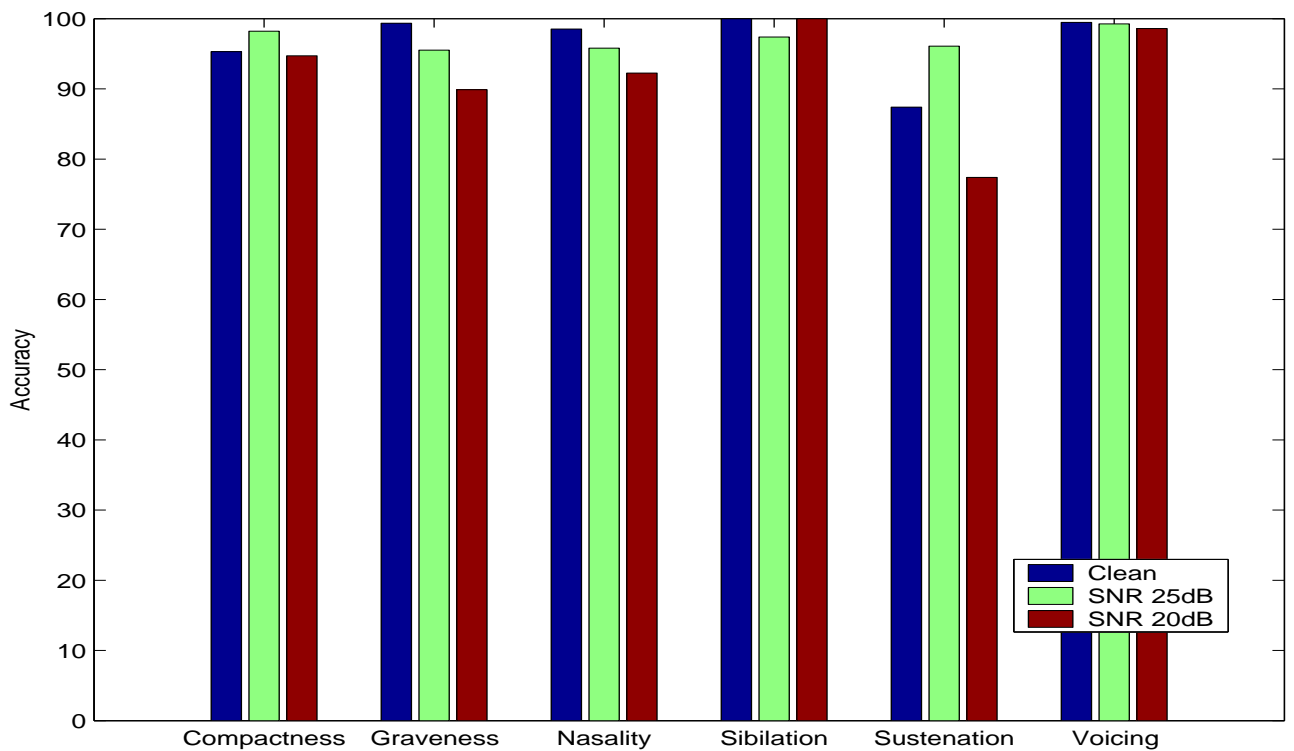


Figure 3. Subjective test accuracy split over DRT distinctive features.

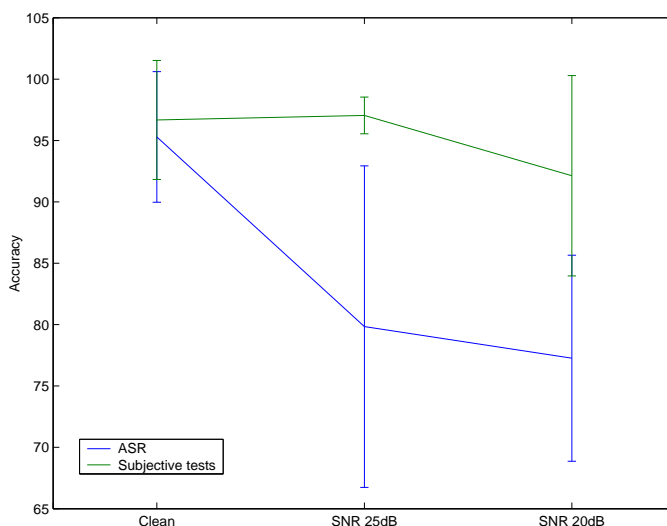


Figure 4. Overall accuracy of ASR and subjective listening tests for DRT vocabulary.

6. Conclusions

The results of using TWA vocabulary with ASR testing are quite encouraging. Our previous approach that used PESQ measure, even though was highly correlated with subjective testing, required reference audio recordings of stimuli. Now we can use ASR instead of PESQ, still maintaining high correlation with subjective tests. We proved that ASR testing can be used in place of PESQ measure, but this is valid just for TTS intelligibility. PESQ measures speech quality, which includes also speech intelligibility. Achieving automatic speech quality measure without reference recordings belongs to our future works.

However, less encouraged is using DRT tests with ASR testing. Although high correlation for graveness category was observed, the overall low correlation with subjective testing is significant, and we cannot recommend using this vocabulary for automatic testing of TTS intelligibility.

7. Acknowledgements

The work has been funded with support from the European Commission, project Euronounce (URL: <http://www.euronounce.net>), and with support from the VEGA grant No. 2/0138/08. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- ¹ M. Cernak and M. Rusko, *An Evaluation of Synthetic Speech Using the PESQ Measure*, ForumAcusticum 2005.
- ² 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*, pp. 136 – 148, Springer 2008.
- ³ Z. Bargár and A. Kollár, *Praktická audiometria*, Osveta, pp. 159-160, 1986 [In Slovak].
- ⁴ M. Rusko and M. Trnka, Word Tests for Speech Understandability Evaluation in Slovak, in: Garabík R. (ed): *Computer Treatment of Slavic and East European Languages*, VEDA, Bratislava, 2005. ISBN 80-224-0895-6, pp.186-192.
- ⁵ Cerňak M.: Unit Selection Speech Synthesis in Noise, Proceedings of ICASSP06, May 14-19, 2006, Toulouse, France.
- ⁶ J.P. Marques de Sá, *Applied Statistics Using SPSS, STATISTICA and MATLAB*. New York: Springer-Verlag Berlin Heidelberg (2003).