ForumAcusticum
2 0 0 5

An Evaluation of Synthetic Speech Using the PESQ Measure

Milos Cernak, Milan Rusko
Slovak Academy of Sciences, 84507 Bratislava , Dubravska cesta 9, Slovakia, **{Milos.Cernak, Milan.Rusko}@savba.sk**,

The paper presents experiments on the use of the perceptual objective measure – ITU-T Rec. P.862 Perceptual Evaluation of Speech Quality (PESQ), for the automatic evaluation of synthetic speech. The approach is based on the evaluation of the statistically significant correlation between the outputs of subjective and objective tests. We propose the following technique to evaluate the usage of the PESQ method for synthetic speech: Firstly, a list of the test words has to be defined for the entire language. Secondly, the tested synthesizers are used to generate synthetic speech signal for all the words in the list. Synthesizer engines of different quality were used for the generation of stimuli: LP synthesizer, RELP synthesizer and PSOLA synthesizer, both in female and male versions. We evaluated created stimuli by listening tests. Thirdly, the PESQ method with original human (reference) and synthesized (measured) recordings as inputs is used to evaluate the overall quality of the synthesized signals. Finally, a correlation of the resulting MOS and objective MOS scores is calculated for each voice. Our results indicate a strong correlation between the mentioned subjective and objective evaluation of the quality of synthetic speech. We plan to use the PESQ measure in automatic evaluation of new versions of synthetic voices, without a need of subjective tests. This approach can foster the life cycle of the development of new versions of synthetic voices tremendously. Using PESQ with "original voice" as reference represents a rapid and repeatable synthetic voice quality measurement technique that provides the developer with results in a few moments.

1 Introduction

Testing and evaluation of synthetic voices represent important steps in TTS systems building. We distinguish between two categories of approaches in speech quality evaluation. The first one is represented by a group of subjective evaluation methods [2, 3], producing Mean Opinion Score (MOS) for displayed speech stimuli. The second group uses an objective evaluation, which is mostly based on some kind of perceptual modeling.

Subjective testing has a dominant position in TTS evaluation. This is obvious, because the objective measures generally need an acoustical reference, and it is not a trivial problem to define a relevant reference for synthesized speech. Nevertheless, the objective measures have matured in the last years and their results show high correlation with those of subjective tests. Moreover, some new non-intrusive objective measures have appeared which do not need the acoustical reference at all [9].

In our experiment we checked the possibility of using evaluated PESQ (Perceptual Evaluation of Speech Quality) for synthesized speech quality evaluation. This very successful objective measure was developed primarily for the use in telecommunication [1]. Our effort was motivated by the recommendation that PESQ performance might be tested by artificial speech and concatenated real speech. In these tests the objective scores for the test signals in each condition serves as a prediction for the subjective condition MOS values [ITU-T Rec. P.862, section 8.1.1 'Choice of source material']. The problem is how to define the acoustical reference to synthetic signals. We decided to look at synthesized speech as at original speech passed through a communication channel. So we chose a recording of the test words uttered by the speaker, whose speech was originally used for the creation of the synthesized voice for the reference signal.

The paper is structured as follows. Section 2 describes proposed measurement technique, section 3 describes performed listening tests, section 4 introduces the usage of the PESQ measure, and section 5 presents the results achieved. Section 6 finally discusses the approach and results.

2 Measurement technique

We propose the following technique to evaluate the usage of the PESQ method for synthetic speech:

1. Firstly, a list of the test words has to be defined for the entire language. This should be a set of phonetically rich words reflecting as many phonetic and other important features of the language as possible. A test set of words in Slovak which has been defined for word audiometry [4] (TWA) was used in the experiment. The recordings of the TWA uttered by two "original" speakers (1 female and 1 male) were used as reference signals. These are the same speakers whose voices were used in the development of the synthesizers.

2. Secondly, the tested synthesizers are used to generate synthetic speech signal for all the words in the

list. Several versions of the Slovak diphone TTS system Kempelen were used to synthesize the whole TWA set in our experiment. Synthesizer engines of different quality were used for the generation of stimuli: LP synthesizer, RELP synthesizer and PSOLA [7] synthesizer, both in female and male versions. We evaluated them by subjective method ITU-T P.800, and we got Mean Opinion Scores (MOS) for each speaker and each synthesizer voice plus reference voices.

3. Thirdly, the PESQ method with original human (reference) and synthesized (measured) recordings as inputs is used to evaluate the overall quality of the synthesized signals. Several values of Objective MOS (OMOS) are obtained for each voice.

4. Finally, a correlation of the resulting MOS and OMOS scores is calculated for each voice.

3 Listening tests

Listening tests were carried out in order to evaluate three versions of our diphone text-to-speech system. The three synthesizers are based on linear predictive (LP) synthesis, residuum excited LP synthesis (RELP) and time-domain pitch synchronous overlap-and-add synthesis (PSOLA), respectively. All of them are in two versions – female and male voice. We tested the overall quality of voices and our aim was to reach MOS values for these synthetic speech signals.

3.1 List of words

The set of test words for Slovak word audiometry (TWA) was designed by Bargár et. al. in 1986 [4]. 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. Table 1 lists all used test words [in Slovak].

3.2 Listeners

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

a) they have not been directly involved in the work connected with assessment of the performance of speech synthesizers, or in related work;
b) 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;
c) they have never heard the same word lists before.
d) 8 subjects (2 males and 6 females) aged from 16 to 73 took part in the experiment .

Table 1: List of the test words for Slovak audiometry TWA according to [4].

1. decade	2. decade	3. decade	4. decade	5. decade
jazero	široký	dolina	adresa	opona
šíp	vek	tón	tyč	kmeň
takto	málo	žltý	tmavý	sto
nos	ľad	zrak	boj	dážď
vyrába	čaj	spieva	deň	heslo
lep	ozvena	lom	pozri	guľatý
živý	stan	nefajči	vlas	var
daň	zober	háj	uteká	zem
humno	pluh	múr	mnoho	búrka
česť	sedí	iste	cieľ	žije
6. decade	**7. decade**	**8. decade**	**9. decade**	**10.decade**
pokojne	farebný	povala	počúvaj	odvážny
stôl	hra	rok	vrch	loď
žiada	mäso	nedaj	cena	osem
dom	sud	meč	osobný	poháňa
vezmi	nevädza	nové	žiak	rak
deravý	pleť	plot	hrot	banka
tlak	soľ	zametá	dnes	svet
noc	lak	chýr	štýl	dym
ucho	domov	buk	múka	dub
sieť	šije	šiesti	ide	lice

3.3 Stimuli

A group of ten words (one decade of the TWA), prerecorded (either uttered by "original" speaker or synthesized by artificial voices) represents one stimulus. All the ten decades of TWA were synthesized by all the synthesizers and recorded by female and male speakers, which gives eight different realizations of each stimulus. Having ten decades we got eighty sound files in total, which were played from the PC to the test participant via Sennheiser HMD 25 closed-system headphones in laboratory conditions.

The testing person rated the speech quality of each stimulus by selecting one of the five Mean Opinion Score (MOS) values in the Microsoft Excel sheet. The interpretation of the MOS values is given in Table 2. The MOS is a live listener test [10] designed to yield a single numeric score that rates the perceived quality of speech of the analyzed audio sample.

Table 2: Interpretation of the MOS values

MOS	Description
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The mean averaged through all the listeners and all the decades of TWA was taken for the MOS value for the entire synthetic voice (see Table 3).

4 Objective tests using PESQ

4.1 PESQ Overview

Perceptual evaluation of speech quality (PESQ) is an objective method, designed for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

Figure 1 shows the overview of the basic philosophy used in PESQ [1]. A computer model of the subject, consisting of a perceptual and a cognitive modeling process, is used to compare the output of the device under test (e.g. a speech or a music codec) with the input (= reference), using any audio signal (speech, music or artificial test signal). A perceptual model, applying a time alignment for delay estimates, models original and degraded speech. From internal representation of the original and degraded speech the difference, which determines the audible difference of both signals is calculated. Finally a cognitive model simulate listening tests, models the difference. The output is a single value in -1 to 4.5 range.

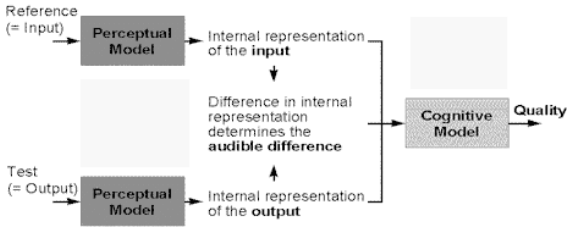

Figure 1: Overview of the basic model approach for an objective perceptual measurement [8]

4.2 Test procedure

We split each stimulus (a decade of TWA) using SFS software [9] automatically into words. In this way we processed all 6 synthetic voices (3 voices for male and 3 for female), and two original voices (one male and one female). Each voice consisted of 100 words (10 decades). We paired all synthetic words with their original-voice counterparts, creating 300 pairs for male and 300 pairs for female voices. All these pairs were then processed by PESQ at sampling frequency 8000 Hz.

We used ANSI C implementation of the P.862, as distributed by the ITU-T.

5 Evaluation results

We calculated Pearson's correlation coefficients for all accomplished tests according to the following formula [5]:

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

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

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

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 using [5]:

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

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.

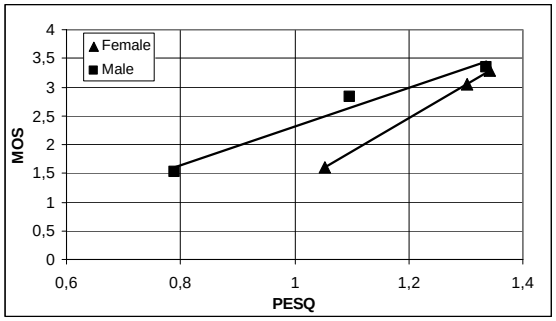

Figure 2: Calculated PESQ values versus MOS values achieved in the listening test, for all three versions of the tested TTS systems

Figure 2 shows calculated PESQ values versus MOS values in a scatter plot, applying linear interpolation for both male and female voices. Table 3 shows calculated coefficients for both voices and all three versions of the

TTS system. The correlation coefficient for female voice is above the 1% significance level, and the correlation coefficient for male voice is above the 8% significance level.

Table 3: Pearson's correlation coefficients for male and female voices. The PESQ values were calculated as mean values over all samples in the voice.

TTS	MOS	PESQ	*r*
LP-female	1.60	1.05	
RELP-female	3.05	1.30	**0.999**
PSOLA-female	3.23	1.34	
LP-male	1.53	0.79	
RELP-male	2.84	1.10	**0.985**
PSOLA-male	3.34	1.34	

6 Discussion

The results given in Table 3 show high correlation of final subjective MOS for each voice with final average PESQ for that voice. We can use the PESQ measure in the automatic evaluation of new versions of synthetic voices, without a need of subjective tests, what is the purpose of this study.

On the other hand, we found very low correlation between the decades of MOS and decades of PESQ. It is evident that PESQ cannot be used for the evaluation of diphone voice on small sample size (such as 10 words). We calculated the PESQ value for each word of TWA, and we then averaged them for getting one objective MOS value per voice. But we get also very low correlation calculating the PESQ values for whole decades of words. This is probably caused by worse functionality of the time alignment module between original and synthesized speech. Moreover, even though the PESQ values for the three TTS voices show high correlation with listening tests, these objective MOS cannot be directly used for TTS evaluation – at least a linear mapping must be applied first.

The approach described in this paper fosters tremendously the life cycle of the development of new versions of synthetic voices. The developer is provided with rapid and repeatable results, highly correlated with the equivalent listening tests. The future work will include testing of current non-intrusive objective methods, such as ITU-T rec. P.563 [9], which do not require acoustical reference at their input. Such a method could simplify the evaluation, because it will not be necessary to have the original-voice recordings of the test utterances available.

Acknowledgements

The work has been supported by the VEGA 2/5124/25 and by the Ministry of Education of the Slovak Republic in a frame of the state task of science and development Intelligent Speech Communication Interface.

References

[1] ITU-T P.862. Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of 3.1 khz Handset Telephony (Narrow-Band) Networks and Speech Codecs, February 2001.

[2] M. Goldstein, 'Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener'. *Speech Communication*, Vol. 16. pp. 225-244 (1995).

[3] V. Kraft and T. Portele, 'Quality Evaluation of Five German Speech Synthesis Systems'. *Acta Acustica* 3, pp. 351-365 (1995).

[4] Z. Bargár and A. Kollár, 'Praktická audiometria'. Osveta, pp. 159-160 (1986) [In Slovak].

[5] J.P. Marques de Sá, *Applied Statistics Using SPSS, STATISTICA and MATLAB*. New York: Springer-Verlag Berlin Heidelberg (2003).

[6] http://www.opticom.de/technology/pesq.html

[7] E. Moulines and F. Charpentier, 'Pitch-synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones'. *Speech Communication*, Vol. 9, No. 5, pp. 453-467 (1990).

[8] http://www.opticominstruments.com/psqm.shtml

[9] ITU-T rec. P.563, 'Perceptual non-intrusive single-sided speech quality measure' (2004).

[10] ITU-T P.800. Methods for objective and subjective assessment of quality, August 1996.

[11] SFS – Speech Filling System, Tools for Speech Research, www.phon.ucl.ac.uk/resource/sfs.htm.