Comparative evaluation of synthetic prosody with the PURR method

Gerit P. Sonntag, Thomas Portele

Institut für Kommunikationsforschung und Phonetik (IKP), Universität Bonn
sonntag[portele]@ikp.uni-bonn.de

ABSTRACT

In order to evaluate the prosodic output of a speech synthesis system independently from its segmental quality, we have developed a special way to delexicalize speech stimuli which we call PURR (Prosody Unveiling through Restricted Representation). We compared the use of PURR stimuli for the evaluation of prosodic naturalness in three different test designs: magnitude estimation (ME), categorical estimation (CE), and ranking order (RO). Sentences of different types were synthesized by six German synthesis systems. The synthetic utterances and one human voice were comparatively judged by experienced listeners. On the whole the results of all three methods are in good agreement. Choice of stimuli seems to be more important than the choice of method.

1. INTRODUCTION

With improving intelligibility of synthetic speech it becomes necessary to assess other aspects of speech. The second major dimension of speech quality is naturalness [6]. Naturalness itself is composed of several factors such as voice quality, prosody etc. Though prosody is often modeled by a specific component within a synthesis system, there are inherent problems in assessing its performance separately. One major problem of prosody assessment in perception tests is to focus the listener's attention on the prosody alone. When using unmanipulated speech stimuli other factors (like meaning, intelligibility, voice quality etc.) may influence the listener's judgement. By a delexicalisation of the stimuli we ensure that the listener perceives only the prosody of a given utterance.

Synthetic speech evaluation methods are often described as being either 'diagnostic' or 'global'. The experiments in this paper can be regarded to be both, 'diagnostic' because only the prosody is assessed and 'comparative' because the final output of any synthesis system can be used as basis for the stimuli to be judged.

The condition for any perceptual evaluation is the existence of a relation between the perceived magnitude and the physical size of the stimulation. Stevens [12] divided perceptual continua into two general classes. Sensory perceptions like loudness, brightness or weight are mediated by an additive physiological process and termed "prothetic". Perceptual continua that are based on a substitution of excitation are termed "metathetic". The former indicate a change in quantity while the latter indicate a change in quality. The way to establish whether an attribute is prothetic or metathetic is to observe the relation between results of magnitude estimation (ME) and categorical estimation (CE) scales [3] or of ME and pair comparison (PC) scales [7]. For a metathetic continuum the CE/PC scale is linearly related to the ME scale, while for a prothetic continuum it is linearly related to the logarithm of the ME scale. We will relate to this in section 5.4.

2. METHODS OF PERCEPTUAL EVALUATION

Despite of attempts to standardize prosody evaluation [4] there is still no established procedure. A number of test paradigms are used to assess the overall quality of synthetic speech. One of them is the magnitude estimation (ME) technique which has proven suitable for synthetic speech evaluation [8]. It consists of a direct estimation of the perceived size of the attribute in question and its results are considered to be on a ratio scale. Furthermore an ME evaluation procedure takes less time than the PC procedure. Pavlovic et al. [7] compared ME with PC in a quality evaluation of four different synthesis procedures and three different prosodic rules. There were 16 subjects for each evaluation method; stimulus material was described as "sentence test material". They reported a good general agreement between the two methods.

Delogu et al. [2] evaluated the quality of a human voice (with and without distortion), three formant-based synthesizers and three vocoders. They compared the results of ME, PC, CE, and the reaction times (RT) of a word monitoring task with ten subjects and six sentence pairs as stimuli. Their conclusion was that PC had the best discrimination capability (dividing eight versions into six significantly different subgroups), ME discriminated better than CE and RT yielded the fewest statistically significant differences.

Categorical estimation (CE) (also known as 'mean opinion score', MOS) on a 5-point scale is proposed as a standardized method by [1]. However, there is still disagreement about the number of categories to be used. CE results are usually assumed to be on an interval scale, but it can be criticized that the categories are not necessarily equidistant. Even though we are aware of this problem we will continue to assume CE results to be on an interval scale in this paper.

Goldstein et al. [3] compared ME with a 5-point CE and an 11-point CE evaluating naturalness with different stimulus ranges. There were eight subjects, four sentence pairs, and four to six different versions, respectively. They concluded that both the method and the stimulus range affect naturalness ratings. Ratings on the 5-point scale seemed to systematically overestimate voices of poor naturalness as compared to the 11-point scale.

Salza et al. [9] compared CE of seven attributes on a 5-point scale with PC. They reported a good agreement of the results of both methods for a quality evaluation of three differently

configured synthesizers. Listener (18 subjects) and sentence (10 sentences) effect were reported to be significant in many cases.

We decided to replace the PC method by an overall ranking order (RO) task. Instead of comparing only two versions of one stimulus at one time, our subjects were asked to put all seven different versions of the same sentence in the appropriate ranking order. The results of this experiment are ordinal values and can therefore only be subjected to a restricted statistical analysis. We compared ME, CE and RO according to their discrimination capability.

3. STIMULI GENERATION AND MANIPULATION

In order to collect the stimuli independently from the system developers they were either downloaded from interactive web sites or generated by freely available demo versions. They were calibrated to the same mean intensity and stored digitally (16bit, 16kHz; one system had to be upsampled from 12kHz). The stimuli came from five concatenative synthesis systems, one formant based system and one human speaker. All seven versions were male voices. 16 sentences of each version were recorded, ranging from 3-21 syllables. The sentences comprised six questions, six statements and four orders. The orthographic input text was terminated with a question mark, a period or an exclamation mark, respectively.

In prosody research a number of different methods for the delexicalisation of speech stimuli have been applied [10]. We have compared several methods and found the PURR signals as described in [10] to fulfil the three requirements of adequate transportation of prosodic functionality, easy listening, and automatic generation. The manipulated stimuli contain information about the rhythmic organization of the utterance, the intonation (pitch movements) and intensity distribution. Segmental factors such as possible mistakes in the phonetic transcription of the orthographic input string or shortcomings in the unit inventory are not present in the signal and can therefore not influence the listener's judgement.

4. EXPERIMENTS

Usually ME is done numerically, i.e. subjects assign a number to the perceived magnitude of the attribute under observation. To prevent the influence of known numerical scales (e.g. school marks or percentages) on the intuitive estimation a line length estimation was carried out instead. 12 experienced listeners were divided into two groups. Each group evaluated half of the sentences (i.e. 8 sentences) twice by drawing a line on a sheet of paper. First all stimuli were presented to give an impression of the stimulus range. Judgments were made during the second presentation. Timing was fixed and each test session took about 30 minutes.

Due to time constraints six randomly chosen sentences were excluded from the CE and RO experiment. As a consequence we were not able to further study the influence of sentence type. In the CE experiment we compared the ratings of different attributes, namely ‘overall naturalness’, ‘accentuation’, ‘intonation’ and ‘rhythm’. 14 experienced subjects were asked to give their ratings on a 5-point scale on the adequacy of a) accent placement within sentence, b) the melody of the sentence, c) the temporal structure of the sentence and d) on the 'overall naturalness' in four different test sessions, respectively. The seven versions of the ten sentences were evaluated twice by each subject.

For the RO task subjects were asked to assign a ranking order (1-7) to the seven versions of each sentence according to the perceived naturalness. They could listen to each individual stimulus as many times as they wanted.

For all three experiments presentation order was randomized and during each audio presentation the orthographic form of the sentence appeared on a monitor screen; subjects had been instructed to read them while listening.

5. RESULTS

5.1. Magnitude estimation (ME)

The drawn lines were measured manually with an accuracy of 0.5cm and normalized by dividing each raw line length by the geometric mean of the line lengths of each subject. These normalized values formed the basis of all further analysis. An analysis of variance indicated that the effects of version, of sentence, and of subject were statistically significant ($p<0.05$). A post-hoc Scheffé test further analyzed the significant differences. Only one subject gave significantly different judgments than the others. For versions three subgroups could be significantly distinguished: the human version, versions a-c and versions d-f. The intraindividual consistency was measured by correlating first and second presentation and was $c=0.68$ (Pearson's correlation coefficient, $p<0.01$, $N=616$). The effect of sentence type was not statistically significant. Nevertheless a closer look at the different sentence types shows that system order changes according to the sentence type under observation (see Figure 1). The human version always comes first, but for

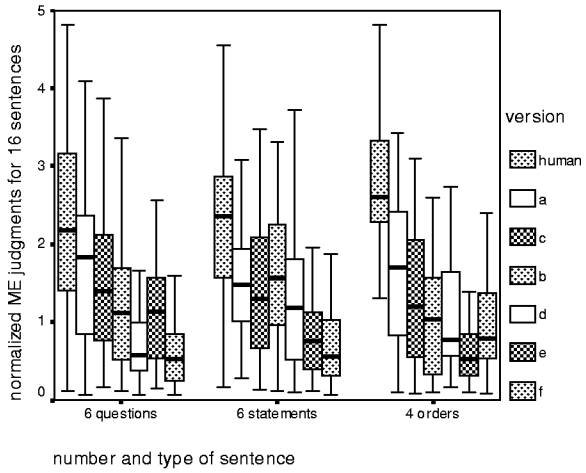

Figure 1: Median, interquartile range, minimum and maximum of the ME judgments according to sentence type.

example version b moves up to position two when we consider statements only. Concerning questions version e scores much better than versions d and f.

5.2. Categorical estimation (CE)

An analysis of variance of the CE results indicated that the effects of version, of sentence, and of subject were significant ($p<0.05$). A post hoc Scheffé test showed that the discrimination capability between the seven versions differs according to the different attributes. Whereas the ratings for rhythm significantly distinguish only three subgroups, the ratings for accentuation and naturalness distinguish four and the ratings for intonation even distinguish five different subgroups. In any case the human version constitutes the top subgroup on its own. As can be seen in Figure 2, the overall ranking order remains the same for all four attributes, except for the intonation ratings, where first and second synthetic version swap positions. Intraindividual consistency (Pearson's correlation coefficient of first and second presentation) was $c=0.6$ ($p<0.01$, $N=980$) in this experiment. Another point of interest was the relation between judgments of the individual attributes. In Table 1 we see that intonation correlates higher with naturalness than the other two attributes.

Correlation	naturalness	accentuation	rhythm	intonation
naturalness	-	0.65	0.69	0.79
accentuation	0.65	-	0.65	0.58
rhythm	0.69	0.65	-	0.61
intonation	0.79	0.58	0.61	-

Table 1: Correlation coefficients (Pearson's) of the different attributes evaluated ($p<0.01$, $N=980$).

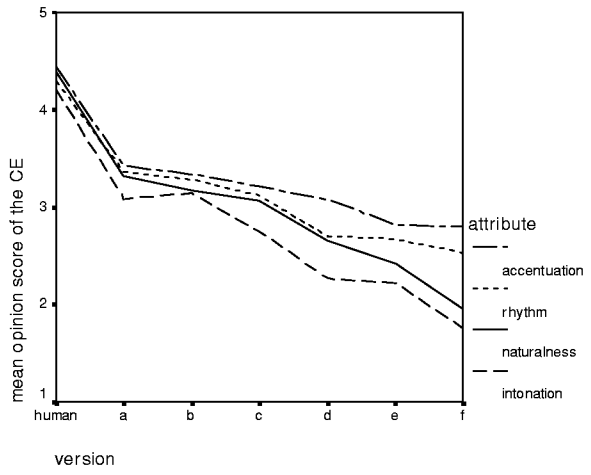

Figure 2: Mean opinion scores of the CE for each attribute evaluated across the seven versions.

5.3. Ranking order (RO)

We applied a non parametrical statistical analysis (Mann-Whitney-U test) to the results of the RO experiment. This analysis yielded significant differences ($p<0.05$) between all versions except for the two best and the two worst synthetic versions (see Figure 4). Subject and sentence effect could not be analyzed because of the test design where each subject gave scores of 1 to 7 for each sentence.

5.4. Comparison of the three methods

We correlated mean ME, mean CE and median RO judgments per version and sentence. Table 2 shows a slightly greater agreement between ME and CE than between RO and the other two methods. The linear relation of CE and ME results (Figure 3) seems to sustain the assumption that the attribute 'naturalness' (in this case: 'prosodic naturalness') is a metathetic continuum [3,7]. What may be more important is the agreement of the overall ranking order between all three methods. Figure 4 was computed by a linear transformation of the ME and CE results to match the lowest and highest value of the RO medians.

Correlation	ME	CE	RO
ME	-	0.87	0.74
CE	0.87	-	0.76
RO	0.74	0.76	-

Table 2: Correlation coefficients (Spearman's) of the different test methods' results ($p<0.01$, $N=70$).

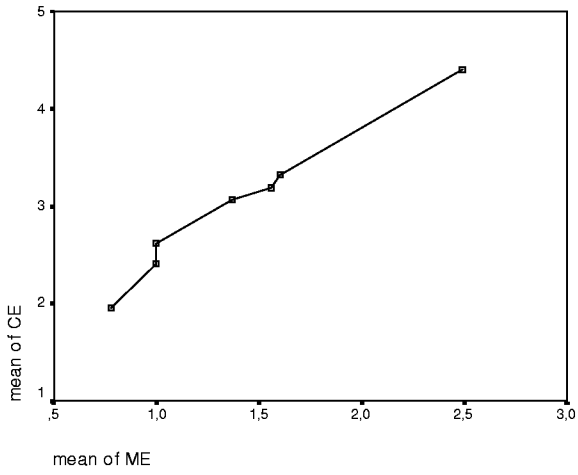

Figure 3: CE results plotted against ME results to show the linear relation between them (the different versions are indicated by squares).

6. DISCUSSION

Looking at the ten sentences that were used in all three experiments we find the same overall ranking order for all three methods (Figure 4). However, with the greater range of stimuli (16 sentences) in the ME experiment, version b and c swap place in the overall ranking order. Not only the number of stimulus sentences but also sentence type affect the overall ranking order. The ME ratings with 16 sentences clearly divides the seven versions into three subgroups: human, versions a-c

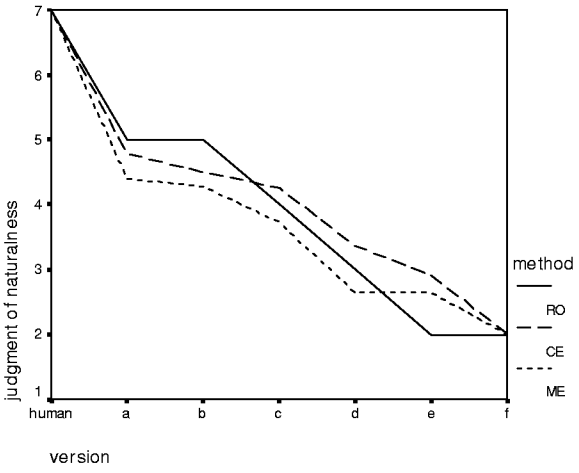

Figure 4: RO median, ME and CE mean of naturalness judgments across the seven versions.

and versions d-f. ME and CE naturalness ratings with 10 sentences result in four overlapping subgroups each. RO ratings even suggest five different subgroups. Taking all results into account we conclude that range and choice of stimuli can be more important than the choice of the evaluation method. We would therefore safely distinguish the human version from three 'good' systems and three 'bad' systems but make no further distinctions. Instead if a comparison of the systems within one of the subgroups is desired we suggest a separate evaluation of these systems only.

One thing that we ruled out from the very beginning is the segmental influence on the judgments. It must be emphasized that we did not evaluate the overall naturalness of the versions, but only their *prosodic* naturalness. A formant-based system which usually scores at the lower end of the scale when unmanipulated stimuli are used [5,6] was judged to be amongst the three 'good' systems in this evaluation (version c). It should be interesting for system developers to evaluate their prosodic component independently of other influences. Further evaluations should study the influence of the stimuli choice. Another interesting topic is the prosodic influence on speech comprehension [11].

7. ACKNOWLEDGMENTS

We would like to thank all our subjects for their participation and two anonymous reviewers for their helpful comments. This research was partially funded by the *Deutsche Forschungsgemeinschaft* in the project He 1019/9-1.

8. REFERENCES

1. CCITT "A method for subjective performance assessment of the quality of speech version output devices", Draft ITU-T Recommendation P.85, study group 12 – report R 6, 1993.
2. Delogu, C.; Paoloni, P.; Pocci, P.; Sementina, C. "Quality evaluation of text-to-speech synthesizers using magnitude estimation, categorical estimation, pair comparison and reaction time methods", *Proceedings of Eurospeech,* Genova, Italy, Vol.1, 353-355, 1991.
3. Goldstein, M.; Lindström, B.; Till, O. "Some aspects on context and response range effects when assessing naturalness of Swedish sentences generated by 4 synthesiser systems", *Proceedings of ICSLP'92*, vol.2, Alberta, Canada, 1339-1342, 1992.
4. Grice, M.; Vagges, K.; Hirst, D. "Prosodic form tests" and "Prosodic function tests", In: *SAM final report*, 1992.
5. Klaus, H.; Fellbaum, K. "Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthese-systemen", *ACUSTICA/Acta Acustica 83*, 124-136, Jan/Feb 1997.
6. Kraft, V.; Portele, T. "Quality Evaluation of Five German Speech Synthesis Systems", *Acta Acustica 3*, 351-365, 1995.
7. Pavlovic, C.V.; Sorin, C.; Roumiquière, J.P.; Lucas, J.P. "A Comparative Analysis of the Magnitude Estimation and the Pair Comparison Techniques for Use in Assessing Quality of Text-to-Speech Synthesis", *Proceedings of the ESCA workshop on Speech I/O Assessment and Speech Databases*, Noordwijkerhout, Netherlands, pp.3.1.1.-3.1.3., 1989.
8. Pavlovic, C.V.; Rossi, M.; Espesser, R. "Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems", *J. Acoust. Soc. Am.,* Vol.87(1), 373-382, 1990.
9. Salza, P.L.; Foti, E.; Oreglia, M. "MOS and Pair Comparison Combined Methods for Quality Evaluation of Text-to-Speech Systems" *ACUSTICA/Acta Acustica 82*, 650-656, 1996.
10. Sonntag, G.P.; Portele, T. "PURR – a method for prosody evaluation and investigation", to appear in: *Computer Speech and Language, Special Issue on Evaluation*, Vol.12(3), 1998.
11. Sonntag, G.P.; Portele, T.; Haas, F. "Measuring the comprehension of different synthetic versions in a dual task experiment" to appear in: *Proceedings of the ESCA workshop on Speech Synthesis*, Jenolan Caves, 1998.
12. Stevens, S.S. *Psychophysics*, New York: John Wiley & Sons, 1975.