Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 35 (1997) 115-119

In-Development Assessment of Concatenation Synthesis by Nonnative Speakers

D. Zboril, M. Libossek, F. Metze
Institut für Phonetik und Sprachliche Kommunikation (IPSK)
Schellingstr. 3 VG/II
80799 München
Germany
{zboril, marion, metze}@phonetik.uni-muenchen.de

Abstract

In-development assessment of speech synthesis is inevitable in industrial contexts. However, there is a strong need to minimize expenditure. Therefore, we present an evaluation procedure which involves considerably less time than other evaluations and may be carried out under laboratory conditions. Furthermore, we compare groups of raters which differ with respect to their competence concerning phonetics and their native language. The ratings of the different groups have proved to be consistent. Another interesting result is that it seems necessary to include measures of general impression as they differ from the single ratings of the more specific measures.

1. Introduction

For many applications a synthesis by concatenation of words (or even larger units) is the most appropriate solution [Pijper], [Eppinger & Herter]. The quality of speech output "directly affects the user's performance and satisfaction with the system" [Robert] that generates the speech output. Therefore special care has to be taken on the appropriate choice of a professional speaker and the recorded material, but also on an effective manipulation of the recorded speech units. There are many different approaches to synthesis evaluation [Tubach et al.]. Concerning the choice of a method or a combination of methods, " very much depends on the goals one has in mind" [Pols]. When designing real world applications a subjective evaluation is indispensable. Therefore it seems desirable to evaluate the system not only at its final stage but also during its development by human raters and not by a computational estimation of human ratings as done by [Halka]. An in-development assessment should meet the following requirements: (i) it should be substantially less time-consuming than a final evaluation; (ii) it should allow for an assessment under laboratory conditions. A promising approach is given by the semantic differential [Osgood] adopted for synthesis evaluation by the use of adjective pairs as described by [Robert] and others.
In our study 5 different groups of raters (4 of which consisted of nonnative speakers) were asked to evaluate American English and French sentences by 10 to 13 bipolar adjective scales with an even number of categories (implying a forced decision). All scales are ordinal. Therefore statistical analysis has to be performed by nonparametric methods [Lienert]. Additionally to (i) and (ii) (see above) we focus our examinations on:
(iii) the "sharpness" of the scales: Ideally, the subjects' choice would be concentrated on a few of the possible answer categories only. (Otherwise the possible answers would be an inadequate representation of the subjects' opinions.)
(iv) the correlations of subject groups: As there is no optimality criterion for the best choice of raters other than a representative sample of the target group, it is important to understand how the composition of subject groups affects their rating behavior. If one intends to replace one subject group by another (e.g. raters of different mother tongues in our study) in case of an in-development assessment, where a representative sample might not be available due to restrictions of time and money their ratings should be highly correlated in order to guarantee comparability.

2. Experiments

2.1. Speech data

The data to be assessed consisted of sentences in American English (Am) and French (Fr) which were to be used in a car navigation system (e.g. <Make a right turn eighty yards ahead.>). Each sentence consisted of a number of concatenated tokens (e.g. <make>, <a right>, <turn>, ...), which were recorded on DAT read in an anechoic chamber by professional speakers in an appropriate context (e.g. <Make a left turn fifty yards ahead.> for <make>). Out of this corpus the tokens to be concatenated for the different sentences where extracted. Special emphasis was laid on the possibility of using the same token in all sentences where it appeared in the same position (i.e. initial, medial, or final). Therefore, the speech signals of some tokens were modified with respect to amplitude and duration of the whole signal or its parts. Finally, the data were coded in ADPCM format.
The generated tokens allow for the construction of several hundred sentences in each language. Out of these, two corpora were stochastically generated, one for each language. In order to simulate real world conditions each of the corpora contained a

certain number of unmanipulated sentences. Thus, the first corpus consisted of 20 American English sentences with an average length of 8.58 words and average number of 4.0 concatenated units and the second corpus consisted of 20 French sentences with an average length of 10.3 words and an average number of 5.3 concatenated units. As in many real world applications, the length and number of tokens for the output had to be kept as small as possible while still allowing a pleasant sounding result.

2.2. The subject groups

Two subject groups were asked to assess the American English, three the French material. The two groups for American English consisted of 7 and 10 persons respectively. Group PhScGe (Phonetic Scientists with German as mother tongue) were German phoneticians with good working knowledge of English, group PhStGe (Phonetic Students with German as mother tongue) consisted of German students of an introductory course to phonetics with normal high school knowledge of English (about 6 years on average), both groups having German as first language.

The three groups rating the French material consisted of 12, 15 and 9 persons respectively. The first of these was the group PhStGe already used for the American material, merely with two persons more. This group had normal high school knowledge of French (about 3 years on average). Group FrStGe (French Philology students with German mother tongue) were German students of a "French Phonetics and Phonology" course at the University of Munich (2nd year of study on average). Finally, group PhStFr (Phonetic Students with French as mother tongue) were French phonetic students at the USHS in Strasbourg on an undergraduate advanced level with French as first language. Three of them acquired a regional language of France, namely Alsatian, but learned French before the age of six.

2.3. Design of the experiment.

The 20 sentences of each language were presented to the corresponding audience by a single loudspeaker using high quality equipment, where possible a Pfleid loudspeaker and DAT. Assessment was performed in a normal lecture room where no special sitting order was assumed. The sentences were put in random order on the tape. Every sentence was repeated twice, with a 2 sec pause between each repetition and a 5 sec pause between each different pair. As mentioned above, the subjects were to evaluate the sentences by 10 bipolar adjective scales. Regarding intelligibility (INT) this means that the subject had to assess the system by the following categories: "extremely intelligible", "very intelligible", "rather intelligible", "rather unintelligible", "very unintelligible", "extremely unintelligible". For the French version three more scales were included in order to have a better coverage of the items of the "Overall Quality Test" [EAGLES]. Unlike [Klaus et al.], evaluations were carried out by printed questionnaires. An English version of all scales and response categories may be found in Appendix 1, with the three additional scales marked in grey. The abbreviations used in this article are also given in Appendix 1.

2.4. Statistics

First of all, we calculated the mode for all scales in the different groups, which seems the most appropriate location parameter of the distribution due to the low number of subjects and its simple interpretation as most frequent judgement for a scale in a particular group. To simplify representation we assigned a numerical value to each answer category (e.g. extremely melodious = 1, very melodious = 2, ..., extremely grating =6). If the same frequencies were assigned to adjacent categories the mode was defined to lie in between, which is expressed by using the numerical mean. There were no bimodal distributions.

To estimate the sharpness (i.e. what would be called variance for metrically scaled data) of the judgements we used the H measure

$$H = \sum_i -(f_i) ld(f_i)$$

where *f* denotes the relative frequency of the answer category *i* and *ld* denotes the logarithm with base 2. We normalized H by its maximum $H_{\max}$ for a given number of subjects and categories:

$$H_{norm} = \frac{H}{H_{\max}}$$

Thus, H_{norm} lies on the closed real interval [0,1]. As H_{norm} is scaled logarithmically we declared in a somewhat heuristic manner all distributions with an H_{norm} value below 0.6 to be reasonably sharp.

For each subject group, all adjectives were brought into an order according to their mode values. The correlations between these orders for the different subject groups were calculated by Horn's rho [Lienert] for ordinal data with rank bindings. The ordering is to be interpreted as telling us which attributes are less problematic with respect to the data than others. For in-development evaluation, this is the crucial point as it is important to know which aspects need further improvement.

3. Results

As the focus of this study is on methodological aspects of in-development evaluation, we are less interested in the absolute values, but rather in the comparison of results obtained under different conditions. All experiments took place under laboratory conditions. The experiments described in this paper lasted not more than 20 minutes.

3.1 Modes

The mode values for the different experiments, subject groups and scales are compiled in Table 1. Generally speaking we find that the different groups are very similar in their respective rating. Differences greater than a single answer category can mainly be found between ratings for different languages. In detail, we have a mode at category two for scale NATURAL vs. ARTIFICIAL in group FrPhStGe, a mode between category three and four for scale FLUENT vs. HALTING in group AmPhStGe, a mode at category four for scale VIVID vs. MONOTONOUS in group AmPhScGe, and a mode at category five for scale VIVID vs. MONOTONOUS in group AmPhStGe which lie outside of the given range. Most of the modes lay around the categories two and three, which confirms the general tendency towards slightly positive group ratings.

3.2 H-Values

For an interpretation of these figures it is instructive to study the intragroup variation of ratings, i.e. did all participants vote for the mode or did they spread over several categories, the mode being determined only by a small margin? This is done using the entropy measure H_{norm} (see section 2.4.) in Table 2. The single values correspond, for most criteria, to a distribution over three or less of the six possible categories. H_{norm} values exceeding 0.6 (indicating a broad distribution) were obtained for the following categories: NATURAL vs. ARTIFICIAL (FrPhStGe), INTELLIGIBLE vs. UNINTELLIGIBLE (AmPhStGe), FLUENT vs. HALTING (AmPhStGe), PLEASANT vs. ANNOYING (AmPhScGe and FrPhStFr), MELODIOUS vs. MONOTONOUS (AmPhScGe and AmPhStGe), VIVID vs. MONOTONOUS (AmPhStGe), FRIENDLY vs. HOSTILE (AmPhStGe), and LOW LISTENING EFFORT vs. HIGH LISTING EFFORT (FrPhStFr).

Compilation of results:	Group:				
Criterion:	AmPhScGe	AmPhStGe	FrPhStGe	FrFrStGe	FrPhStFr
CLR	2	2	1.5	2	1
NAT	3	4	2	3.5	3
NSY	2	1	1	2	1
INT	2	2	2	2	1
SLW	3	3	3	4	3
FLU	2.5	3.5	2	2	2
PLS	2	3	3	2.5	3
MEL	3	4	3	3	3
VIV	4	5	2.5	3	3
FRN	2	3	2.5	3	2.5
LSE			2	1.5	1
ACC			2	1	2
PGI			2.5	2	2

Tab 1: Modes for the different experiments, subject groups and scales.

Criterion:	AmPhScGe	AmPhStGe	FrPhStGe	FrFrStGe	FrPhStFr	Average
CLR	0.46	0.46	0.46	0.29	0.39	0.41
NAT	0.55	0.39	0.41	0.44	0.66	0.49
NSY	0.23	0.54	0.33	0.31	0.28	0.34
INT	0.55	0.78	0.33	0.31	0.33	0.46
SLW	0.23	0.29	0.33	0.43	0.52	0.36
FLU	0.57	0.78	0.33	0.48	0.56	0.54
PLS	0.62	0.51	0.37	0.44	0.61	0.51
MEL	0.73	0.62	0.28	0.39	0.56	0.52
VIV	0.55	0.89	0.41	0.52	0.6	0.59
FRN	0.73	0.59	0.46	0.43	0.5	0.54
LSE			0.39	0.44	0.61	0.48
ACC			0.29	0.48	0.52	0.43
PGI			0.41	0.43	0.57	0.47
AVG	0.52	0.58	0.37	0.41	0.52	0.48
Hmax	2.52	2.52	3.58	3.17	2.56	

Tab. 2: H values for the different experiments, subject groups and scales.

	AmPhStGe	AmPhScGe	FrPhStGe	FrFrStGe	FrPhStFr
AmPhStGe		0.85			
AmPhScGe	0.41				
FrPhStGe				0.65	0.81
FrFrStGe			-0.04		0.84
FrPhStFr			0.19	0.66	

Tab. 3: Correlations of modes (above the principal diagonal) and H_{norm} values (below the principal diagonal).

3.3 Correlations

The correlations of the ranked scales (see section 2.4.) are given in Table 3 above the principal diagonal. Although correlations are generally high (rho>0.6) we include only those values that express meaningful correlations between the ratings of different subject groups for the same language version. It is remarkable that the correlation between FrFrStGe and FrPhStGe is relatively low compared to the other two correlations with French native speakers.

One can also look for correlation between scales ordered with respect to the H values for the different subject groups (Tab. 3 below the principal diagonal). A high correlation would then tell us that there are scales that are constantly rated with greater (or less) sharpness than others. (Although we deal with metrically scaled data it is reasonable to apply Horn's Rho because of the low number of subjects and the occurring rank bindings.) Only for the groups FrPhStFr and FrFrStGe does rho exceed 0.6 indicating a similar intragroup variability. However, all other correlations are below 0.5, in one case even slightly negative. This is also true for the correlations not represented in the table. In general, there seems to be little correlation in insecurity of rating.

4. Discussion

Despite the relatively low number of subjects, our data set can generally speaking be said to be "well-behaved" as we find that overall values of entropy are rather low (implying sharp distributions and a coherent rating process) and there seems to be no regular pattern in the data set which could hint at methodical weaknesses in our concept.

As high H-values have to be interpreted as uncertainty of the subject group (not necessarily of the single subject!) when making its decisions, it seems desirable not only to have high correlations for the ranking of the scales according to mode but also according to H_{norm}. However, this holds only for the groups FrFrStGe and FrPhStFr, where the concordance of intragroup variability indicates that the hypothesis that nonnative-speaker language students could approximate the behavior of native speakers to a great extent is quite reasonable. In fact, both groups are assumed to have approximately the same knowledge about French phonetics.

If we drop the requirement of correlation of H_{norm} values and even assume correlations of modes above 0.6 as significant (see [Lienert]), there still remains the problem of interpreting the results. The coherent answers and the high correlations for the modes allow for three explanations:

(a) The correlations are purely accidental.

(b) Our scales are not sensitive with respect to the differences between the two speakers and languages, but only with respect to general properties of the synthesis system, which were approximately equal for all experiments.

(c) Our scales are sensitive with respect to the differences between the two speakers and languages as well as to general properties of the system, but the latter ones are dominating here or the two types of speech material are similar to a great extent with respect to the parameters evaluated.

Though possible, (a) seems not to be very reasonable due to the high number of correlations and the low H values.

There are certain scales which are more associated with the speaker and his speaking style than with the synthesis system, of which they are nevertheless not independent. We get rather consistent ratings e.g. for the scales FRIENDLY vs. HOSTILE, CLEAR vs. NOT CLEAR, SLOW vs. FAST, and VIVID vs. MONOTONOUS; however, as the corresponding modes are similar for both languages (with exception of VIVID vs. MONOTONOUS) and some of the high H values occurred with these scales we cannot formally reject (b).

The variations in mode larger than one category occur only between the different versions discussed indicating that (c) is more likely to be true than (a) and (b). (c) includes the possibility of system specific factors overlaying speaker (or language) specific ones thereby explaining the results for FRIENDLY vs. HOSTILE, CLEAR vs. NOT CLEAR , SLOW vs. FAST, and VIVID vs. MONOTONOUS.

While our investigations provide a good empirical basis to formulate hypotheses on the basis of the given material, we can neither prove nor dismiss any hypothesis. It would need a further survey with a greater number of test-subjects before one can arrive at a final conclusion.

As a final point of our discussion we should mention the results for the scales LOW LISTENING EFFORT vs. HIGH LISTING EFFORT, ACCEPTABLE vs. UNACCEPTABLE, and POSITIVE GENERAL IMPRESSION vs. NEGATIVE GENERAL IMPRESSION. The corresponding modes are mostly less than or equal to the modes of the other scales. All scales concern the general impression, one of them explicitly. At least for our data we can say that the general impression is slightly better than the one that could be assumed according to the scales asking for the single properties of the speech data. If confirmed by further investigations this would be a rather important result.

5. Conclusion

All evaluations took place under laboratory conditions and were performed in about 20 min. Data processing could be considerably shortened by automatization. The proposed methods therefore meet requirements (i) and (ii) given in section 1. Furthermore, our scales have proven to be sharp, and correlations between different subject groups are high. It is a reasonable working hypothesis that our scales are sensitive with respect to the different versions (different speakers and different languages) as well as to the system properties. It seems important to include scales of general impression as they differ from the single ratings of the other scales.

Acknowledgments

We thank the subjects who volunteered for our experiments as well as Andre Bothorel and Monique Krötsch for their support.

6. References

[1] Eppinger, B. and Herter, E. (1993): Sprachverarbeitung; Carl Hanser, Berlin.

[2] Halka, U. (1991): Objektive Sprachqualitätsanalye durch Schätzung auditiver Attribut-Beurteilungen; in: DAGA'91, Bochu16m.

[3] Klaus, H. et al. (1993): An evaluation system for asserting the quality of synthetic speech based on subjective category rating tests; EUROSPEECH'93, Berlin.
[4] Lienert, G.A. (1986): Verteilungsfreie Methoden in der Biostatistik; Anton Hain, Meisenheim.
[5] Osgood, C.E. et al. (1971): The Measurement of Meaning, University of Illinois, Urbana.
[6] Pijper, J.R. de (1994) High-quality message-to-speech generation in a practical application. Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis, Mohawk Mountain House, New Paltz, New York.
[7] Pols, L. (1995): Hot topics in the Field of Speech Synthesis Assessment; in: Bloothoft, G. et al. (eds.): European Studies in Phonetics and Speech Communication; OTS Publishers; Utrecht.
[8] Robert, J.-M. et al. (1989): Subjective Evaluation of the Naturalness and Acceptability of Three Text-To-Speech Systems in French; EUROSPEECH'89, Paris.
[9] Tubach, J.P. et al. (eds.): La parole et son traitement; Masson, Paris.
[10] EAGLES Handbook on Spoken Language Systems (work in progress) (Version of 18 May 1995)

APPENDIX 1

Categories:

extremely	very	rather	rather	very	extremely

Scales:			**Abbreviations:**
clear	**vs.**	not clear	CLR
natural	**vs.**	artifical	NAT
noisy	**vs.**	not noisy	NSY
intelligible	**vs.**	unintelligible	INT
slow	**vs.**	fast	SLW
fluent	**vs.**	halting	FLU
pleasant	**vs.**	annoying	PLS
melodious	**vs.**	grating	MEL
vivid	**vs.**	monotonous	VIV
friendly	**vs.**	hostile	FRN
low listening effort (LSE)	**vs.**	high LSE	LSE
acceptable	**vs.**	unacceptable	ACC
positive general impression (GI)	**vs.**	negative GI	PGI