METHODS FOR REDUCING CONTEXT EFFECTS IN THE SUBJECTIVE ASSESSMENT OF SYNTHETIC SPEECH

Chaslav V. Pavlovic, Mario Rossi, and Robert Espesser

Institut de Phonetique, LA 261 CNRS, Universite´ de Provence,
Aix en Provence, FRANCE
& (1st. author only) University of Iowa, Iowa City, Iowa, USA.

ABSTRACT

The contextual invariance of categorical and magnitude estimates of speech quality could be improved by introducing a reference system (natural speech) and by appropriately normalizing the results with respect to it.

1. INTRODUCTION

A potential problem with subjective scaling of speech quality occurs when the rating of a certain system needs to be generalized outside the set of systems used in the experiment. Namely, the rating of a system may change depending on the selection of other systems evaluated at the same time ("context effect"). We evaluate here whether the context effects could be reduced by introducing a reference system (natural undistorted speech) common to all experiments, and by normalizing the rating of any given synthesizer in reference to the rating of the natural speech. Two subjective psychophysical techniques are evaluated: magnitude estimations (MEs) and categorical estimations (CEs).

The ratings of four systems labeled "A" and four systems labeled "B" were evaluated in two different types of context: "A and B" context and "A or B" context. Systems A were of superior quality to systems B. Both systems A and systems B were evaluated separately within their groups (A or B context), and together (A and B context). The research question is whether the ratings of the stimuli are invariant to these changes in context, both in the absolute and in the relative sense. These context effects were evaluated both with and without the reference condition. This particular design was selected because past research indicates that all scaling techniques may be particularly sensitive to it. It is hypothesized that subjects always use one restricted range of numbers regardless of the stimuli being evaluated. If this were indeed the case, there would be a strong tendency to use the same range of numbers for systems A only, systems B only, and systems A and B together. Given that systems A are superior in quality to systems B, the ratings of B will, therefore, be better when these systems are presented alone than together with A. The opposite would be true of systems A.

2. METHOD

The subjects were equally divided into 12 experimental groups. Six experimental groups gave ME and the other six CE judgments. The groups are identified by letters that correspond to the listening conditions they were exposed to. These six labels are ABR, AR, BR, AB, A, and B. Symbol A signifies that the group judged conditions A, symbol B that the group judged conditions B, and symbol R that the group judged the reference condition. The non-normalized group results for each condition were calculated as the means across subjects and condition repetitions. The arithmetic means were used for CEs, while the geometric means were used for MEs. Neither for the MEs nor for the CEs was the reference condition explicitly defined to the subject as such. Rather, it was treated as just another experimental condition. The subjects were required to judge how satisfied they were with the particular communication situation. For CEs the scale from 1 to 20 was used. Direct ME procedure and the sentence test material described in more detail in [1] were used.

3. RESULTS

In the tasks which did not incorporate the reference stimulus, relatively large AB context effects were seen (Fig. 1 for CEs; Fig. 2 for MEs). They seemed to be particularly severe in the case of CEs, where the mean rating of systems A and B were almost equal to each other in the "A or B" context, but quite different in the "A and B" context. When the reference condition was present, a large decrease in the AB context effect was seen in the CE (Fig. 3), while no improvement was demonstrated in the ME (Fig. 4). The introduction of the reference condition did not seem to have affected the relative ratings of the other systems neither for the MEs (Fig. 5), nor for the CEs. This indicates that some form of normalization may prove beneficial with regards to context effects.

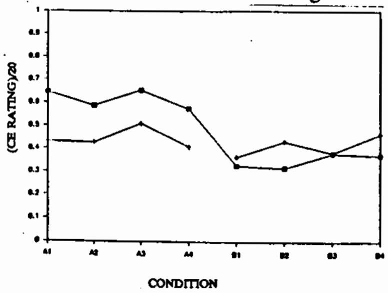

Fig. 1 CE ratings four groups AB (squares), A (pluses), and B (diamonds).

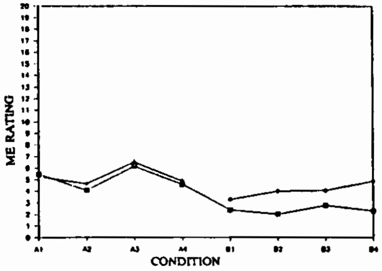

Fig. 2 ME ratings four groups AB (squares), A (pluses), and B (diamonds).

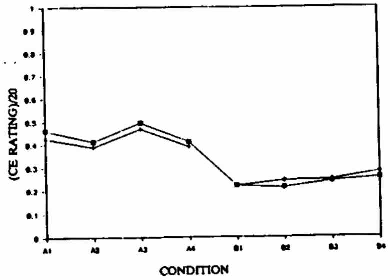

Fig. 3 CE ratings four groups ABR (squares), AR (pluses), and BR (diamonds).

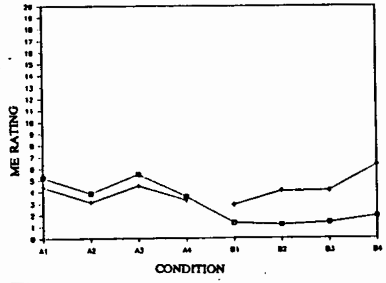

Fig. 4 ME ratings four groups ABR (squares), AR (pluses), and BR (diamonds).

4. NORMALIZATION

Two measures of the merits of normalization were used. These are the standard deviation (s), and the correlation (r) between the ratings of the eight experimental systems (A and B) observed, on one hand, in the "A and B" context, and on the other hand, in the "A or B" context. Measure s expresses the absolute proximity of the measurements made in the two contexts. Measure r is sensitive to how well relative ratings of the systems agree in various contexts. The smaller the s and the larger the r the more context-free the procedure is.

The application of measure s presumes that all results are on the same scale. This is indeed the case for all normalized values. This is also the case for the non-normalized CEs that are divided by the maximum scale value. However, in the case of the non-normalized MEs the scales are arbitrary and cannot be transformed to a 0 to 1 range. In the latter case, instead of s, the measure labeled s' was used. It is defined as s divided by the mean rating of the stimuli in the "A and B" context.

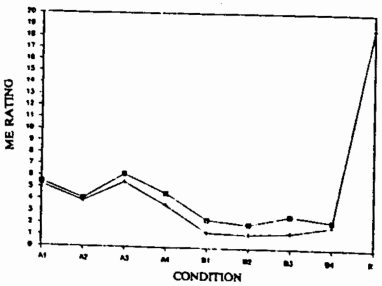

Fig. 5 ME ratings four groups AB (squares) and ABR (pluses).

The CE procedure typically results in an interval-type scale. Therefore, it is invariant to multiplication by a constant, or to addition of a constant. Thus, the results could be normalized by either of these operations. In addition, normalization could be performed on the group results, or on the results of individual subjects. In the case of normalization by multiplication, the rating of a stimulus is multiplied by the reciprocal of the rating of the reference stimulus. This operation applied to the mean group results is labeled "CE_MG," where M stands for "multiplication," and G for "group." Normalization by multiplication applied to the results of individual subjects is labeled "CE_MI," where I stands for "individual." In normalizing results by adding a constant, first the complement to 20 (maximum scale value) of the reference stimulus rating is added to the non-normalized value of the stimulus. Subsequently, these numbers are divided by 20. This procedure leads to the same results regardless of whether it is applied to the group or to the individual results. It will be labeled "CE_C," where C stands for "complement."

The measures of context effect s and r for CEs are given in Table I. All normalization procedures substantially reduced context effects with respect to the non-normalized results of the groups that did not judge the reference system. For example, in the case of method CE_MG correlation-type measure r increased from 0.48 (for the non-normalized results) to 0.98 (for the normalized results), while s decreased from 0.13 (for the non-normalized results) to 0.05 (for the normalized results). However, with respect to the non-normalized results obtained by the groups that judged the reference condition, the context effect was made somewhat worse with normalization.

The ME procedure results in a ratio-type scale, and is invariant to multiplication by a constant. Consequently the normalized results are obtained if the ratings of stimuli are multiplied by the reciprocal of the rating of the reference system. As was the case with CEs, this operation could be performed either on the group results, or on the individual subjects' results. In addition, the ME results could be calculated as "absolute" or "relative" [1]. The normalization procedures on the absolute group results is labeled "ME_AG" (symbols A and G represent "absolute" and "group," respectively), while the normalization procedure on the absolute individual results is labeled "ME_AI" (symbol I stands for "individual"). The normalization procedures either on the group or individual relative ratings yield the same values which are labeled "ME_R" (R stands for relative).

The measures of context effect r, s (if meaningful), and s' are given in Table II for these three normalization procedures, as well as for non-normalized results. Fig. 6 gives normalized MEs for the best of these procedures, i.e. ME_AI. All normalization procedures substantially reduce the context effects with respect to the non-normalized results of the groups that judged the reference system. However, the real benefit of normalization should be assessed against the non-normalized results obtained without the reference system. There, only the procedure ME_AI appears to reduce the context effects.

TABLE I. CONTEXT EFFECT MEASURES R AND S FOR CE.

PROCEDURE	R	S
CE (NON-NORMALIZED) (WITHOUT R)	0.48	0.13
CE (NON-NORMALIZED) (WITH R)	0.99	0.02
CE_MG	0.98	0.05
CE_MI	0.95	0.06
CE_C	0.88	0.08

TABLE II. CONTEXT EFFECT MEASURES R AND S, AND S' FOR ME

PROCEDURE	R	S	S'
ME (NON-NORMALIZED) (WITHOUT R)	0.84		0.35
ME (NON-NORMALIZED) (WITH R)	0.01		0.75
ME_AG	0.79	0.06	0.40
ME_AI	0.89	0.06	0.24
ME_R	0.66	0.05	0.47

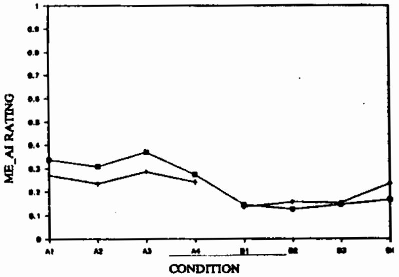

Fig. 6 Normalized ME ratings, method ME_AI, groups ABR, AR, BR

In the s value the best ME procedure (ME_AI) is practically equal (s = 0.6) to the best normalized CE procedures (CE_MG, CE_MI), but inferior to the non-normalized CE procedure when the reference stimulus is used (s = 0.2). In the r values the procedure is worse (r = 0.89) than both, better normalized CE procedures (r = 0.95 to 0.98), or the non-normalized CE procedure when the reference stimulus was presented (r = 0.99).

5. ACKNOWLEDGMENTS

This research was made possible by a grant from the EEC Esprit SAM project (Grant #2589).

6. REFERENCES

1] Pavlovic, C.V., Rossi, M., and Espesser, R. (1990). "Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems," J.Acoust.Soc.Am. 87, 373-382.