Introduction

It is generally known that in natural reading materials, factors are confounded with each other, and that the effects of these factors on measures such as duration interact (Van Santen and Olive, 1990; Van Santen, 1992). Confounding refers to the fact that the occurrence of factor values is correlated so that many combinations of values are rare, e.g., a full vowel in an unstressed versus a stressed first syllable of a two syllable word in English. In reality, of course, many combinations cannot occur at all because they are contradictory (e.g., simultaneously being phrase-final and not word-final) or because of the constraints of the language (e.g., being a word-final /h/ in English). Other combinations do not occur in a given text corpus because their frequency in the language at large, or in the particular text genre, is small. The implication of confounding between factors is that the effects of a given factor cannot be measured by simply computing the mean durations (or other dependent measure, such as spectral balance) for each level of that factor. A good example of this are the effects of intra-word location and syllabic stress on vowel duration (Van Santen, 1992). In English two-syllable words, usually the first syllable is stressed (Cutler and Carter, 1987). Because of this, the average duration of the second syllable is shorter than that of the first syllable because the latter is generally stressed and the former not. However, the opposite is the case if one restricts measurements to stressed syllables in both positions in the word.

The term interaction in general refers to the effects of one factor being modulated by other factors, but is used in at least three different ways depending on how one defines effect. The standard additive definition is the definition used in the analysis of variance, and measures the effect of a factor by the difference in the dependent measure between the levels of the factor. The statement that syllabic stress has a 25 ms effect assumes this definition. An alternative multiplicative definition is as a percentage change: stress has a 14 percent effect. Finally, one can significantly broaden the concept, and only observe the order of the effect (lengthening vs. shortening). For example, the ordinal effect of intra-word location (encoded as first vs. second syllable) is reversed by word length factor when not confounded with stress, because, holding all else constant, in two-syllable words the second syllable is typically longer than the first syllable, whereas in three-syllable words usually the first syllable is longer than the second syllable (Van Santen, 1992). Thus, this presents an example of an ordinal interaction. We use the term directional invariance when ordinal interactions are absent and the direction of change (but not its size) due to one factor value is independent of the value of the other factors. Of course, directional invariance does not imply that no additive or multiplicative interactions can exist; it is a much weaker property.

The methodological challenge posed by the twin problems of confounding and interactions is that the former requires statistical methods that make assumptions about how missing factorial combinations can be inferred from combinations that are available in the corpus, whereas interactions imply that these assumptions cannot be
overly simple. A compromise was proposed in which it is assumed that one can in

certain situations judiciously split factors into two groups: factors of interest and

nuisance factors, and assume that the direction of change due to the factors of interest

is independent of the values of the nuisance factors and vice versa, i.e., we assume
directionally invariance of the two groups with respect to each other. Thus, one might
model the effects of stress (S), intra-word location (L), and consonantal identity (P) in
the presence of additional (nuisance) factors (X,Y,Z,...) as:

\[
\text{Duration}(S,L,P,X,Y,Z,...) = F[M(S,L,P), B(X,Y,Z,...)]
\]

where nothing is assumed about the \(M()\) and \(B()\) functions, and where \(F()\) is strictly
increasing in its two arguments (monotonic). Based on this model, one can search the
data for quasi-minimal pairs, which are pairs of subsets that are identical with respect
to the nuisance factors. (They are called quasi-minimal, because one may have re-
deфини́e some of the factors to combine certain factors levels.) These pairs can be

treated as independent statistical units, and hence lend themselves to standard
statistical analyses. However, often the data are not rich enough to contain a sufficient
number of such pairs; moreover, this analysis does not produce a measure of the
magnitude of the effects of, and the interactions between, the factors of interest.
Towards this end, we have proposed the corrected means analysis, which assumes
that \(F()\) is either additive or multiplicative:

\[
\text{Duration}(S,L,P,X,Y,Z,...) = M(S,L,P) + B(X,Y,Z,...)
\]

or

\[
\text{Duration}(S,L,P,X,Y,Z,...) = M(S,L,P) * B(X,Y,Z,...)
\]

This model allows one to estimate the \(M(S,L,P)\) parameters (e.g., "mean" values of
the (S,L,P) combinations) using standard least-squares methods, and to interpret these
parameters as the true mean durations, also called corrected mean durations, as a
function of \(S, L,\) and \(P,\) while holding the nuisance factors constant. In other words,
under the assumptions of this model, one can obtain the same observations that would
have been provided by the unobtainable, perfectly balanced, experiment.

Calculating corrected means

As informal speech material is never balanced, we are faced with widely varying
numbers of realizations for each of the phonemes with respect to all the other relevant
factors. This means that raw means of duration, pitch, or formant cannot be compared
between conditions (cf. discussion of this topic in Van Santen and Olive, 1990; Van
Santen, 1992). The large under sampling of possible combinations of factor values
and the variability in sample sizes precludes the use of normal ANOVA and
MANOVA statistics. To solve this problem we use a method developed by Van
Santen (1993a, see also Van Santen, 1992, 1993b).

The corrected means analysis model is a special case of the general linear model.
The standard method for estimating the corrected means, \(M()\), would be as follows.
First, we construct an incidence matrix (see Table 1) where rows correspond to
combinations of levels on the factors of interest, and columns to combinations of levels on the nuisance factors. We next would use Dodge’s R-method to determine which cell means in this matrix can be estimated (Dodge, 1981). In this method, estimability is determined iteratively, by filling at each step each empty cell \((c,r)\) when three filled cells can be found in locations \((c,r'), (c',r),\) and \((c',r').\) One fills this cell by adding the means in cells \((c'r)\) and \((c,r')\) and subtracting the mean in cell \((c',r').\) One then eliminates from the data any observations whose factor levels correspond to a non-full column in the resulting incidence matrix. It is known that once this is done, the remaining data allow unique estimation of the (remaining) parameters.

We found, however, that the standard method is not as robust to small violations of the corrected means model as we would like. The quasi-minimal pairs based method described next does not interpolate any cell means but only uses filled cells. This method eliminates a higher percentage of the data, but is more robust. In this method, we again use the incidence matrix as starting point. From this matrix, we obtain for each pair of rows (i.e., pairs of combinations of levels on the factors of interest) a list of cell pairs such that each cell pair is obtained from the same column (i.e., it is a quasi-minimal pair) and is not empty. For each such list, we can compute the mean difference between the means in the pairs. From this, we can construct a square matrix whose order is equal to the number of rows in the incidence matrix. This matrix contains the means of the within-quasi-minimal pair differences. The square means matrix can then be fitted with the additive model using standard least-squares methods (e.g., using Praat).

To calculate the means of the within-quasi-minimal pair differences, i.e., the mean differences between rows, we needed some way to account for the varying sample sizes that underlay each table cell mean. These varying sample sizes determined the variances of the within-quasi-minimal pair differences, i.e., table cell differences. The weighting of each difference should reflect that differences based on smaller samples had a higher variance, i.e., error. Under the assumptions of equal variance \((s^2)\) for the individual measurements, the variance of the differences between two table cell means \((C_{ik}, C_{jk})\) scales as

\[
\text{var}(C_{ik} - C_{jk}) = s^2 \times (1/N_{ik} + 1/N_{jk})
\]

To calculate the mean difference between rows \((i-j)\) from the individual cell differences \((C_{ik} - C_{jk})\) we choose as the weighting factor of the difference \((w_{i,j,k}):\)

\[
w_{i,j,k} = 1/(1/N_{ik} + 1/N_{jk})
\]

which corresponded to the scaling of the reciprocal of the standard error due to the sample sizes \((N_{ik}, N_{j} \) are the number of samples in each cell). It must be noted that the choice of weighting factors had only a small effect on the corrected mean values, as long as larger samples had larger weights.

The reason that this above method is more robust than Dodge’s R-method is that it only uses columns where at least two cells are filled. If the combination of nuisance factor levels corresponding to this column produces an unusually large duration, this
does not affect the difference score associated with the row pairs that do have filled cells in that column because both values in the pair are unusually large, nor does it affect the difference scores of row pairs that do not have filled cells in that column because the column plays no role in the overall difference score of this pair. By contrast, in the standard method, many columns where only one cell is filled are used in the estimation process, and many of these cells have large standard errors because of data sparsity, which then creates unreliable estimates for the corresponding row parameters.

The results are the relative Corrected Means of the rows, e.g., the corrected mean durations of the combinations of the position in the word and stress levels. For any fully balanced set of realizations, the result of this procedure would be identical to the raw means. Therefore, the corrected mean values can be interpreted as a least RMS-error approximation of 'balanced' means with an unbalanced data set. Because the corrected means are calculated from differences only, they need an absolute offset value to get 'real' means. We choose as the offset the overall mean duration of all realizations used for the calculation of the corrected means.

The above description is based on the assumption that the factors affected the segmental duration, pitch, or formants in an additive manner. However, if all values are replaced by their logarithm, the resulting model will be multiplicative. No further changes are necessary to cover a multiplicative model. Earlier tests showed that the results for a multiplicative model are, in general, more extreme with larger differences between factors than those for the additive model so we decided to use the more conservative additive model.

<table>
<thead>
<tr>
<th></th>
<th>Female, /N/, Prominence 2, 2-syllables</th>
<th>Male, /d/, Prominence 0, 3-syllables</th>
<th>Male, /f/, Prominence 1, 1-syllable</th>
<th>.....hundred s of further columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ &amp; I</td>
<td>-</td>
<td>mean</td>
<td>mean</td>
<td>.....</td>
</tr>
<tr>
<td>- &amp; I</td>
<td>-</td>
<td>mean</td>
<td>mean</td>
<td>.....</td>
</tr>
<tr>
<td>+ &amp; M</td>
<td>mean</td>
<td>mean</td>
<td>-</td>
<td>.....</td>
</tr>
<tr>
<td>- &amp; M</td>
<td>mean</td>
<td>mean</td>
<td>-</td>
<td>.....</td>
</tr>
<tr>
<td>+ &amp; F</td>
<td>mean</td>
<td>-</td>
<td>mean</td>
<td>.....</td>
</tr>
<tr>
<td>- &amp; F</td>
<td>mean</td>
<td>-</td>
<td>mean</td>
<td>.....</td>
</tr>
</tbody>
</table>

Table 1. Example of an incidence matrix used to calculate the corrected mean values of all six combinations of syllable stress (+ or -) and position in the word (Initial, Medial, and Final). There are 15 mean cell-by-cell row differences for 6 rows, e.g., (+&M) - (-&F) or (+&I) - (-&I), and therefore 15 linear ('normal') equations for the 6 hypothetical mean values. Solving these 15 equations gives the 'best estimates', in a least squares error sense, for the mean duration of each row. Note that, in general, less than 50% of the cells are filled, e.g., /N/ cannot be Word-Initial, monosyllabic words have no Word-Medial consonants, and /d/ cannot occur in Word-Final position in Dutch.

**Statistical analysis**

The original mean row differences are calculated from pair-wise cell differences.
The statistical significance of the size of the difference between each two rows can be tested on the collection of cell-pairs used to determine this difference. Because of the unbalanced distribution of realizations over the table cells, the statistical analysis is limited to a distribution-free test, the Wilcoxon Matched-Pairs Signed-Ranks test (WMPSR). Each pair of table cells is used as a single matched pair of mean values in the analysis. Distribution-free tests are generally considered to be less sensitive than tests based on the Normal distribution, e.g., the asymptotic relative efficiency of the WMPSR test with respect to the Student-t test is 0.95 when we assume a Normal distribution. However, consonant durations and pitch values are not normally distributed and we want to check the differences independent of the details of the chosen model and weighting function. Both facts together give the Wilcoxon Matched-Pairs Signed-Ranks test an advantage over the Student-t test. Using the WMPSR test on the set of differences between a pair of rows is completely independent of the weighting function used to calculate the mean difference between the rows.

### Correlations

Corrected means analysis of measurements gives the effects of the factor levels on the mean values, corrected for all nuisance factors. After such an analysis, the question arises on how correlations between measured values are distributed between
corrected means (isolating only the relevant factors) and within cells (i.e., variation unaccounted for by the factors used). The corrected means values can be correlated with each other, e.g., duration versus formants. To determine how strong a correlation remains after correcting for ALL the factors involved (relevant and nuisance), the cell means and variances have to be normalized. Two situations are important:

1) All variances are expected to be equal. In this case, cell-means are subtracted from all values, and a single degree of freedom is subtracted for every cell. This makes all cells have zero mean value. Cells with only a single value are discarded.

2) Variances are expected to differ (e.g., variances in duration differ between long and short vowels). After subtracting cell-means (as in 1), all values are divided by the cell standard deviation \((z=(x-\text{mean})/sd)\). The values now have mean = 0 and standard deviation = 1. Two degrees of freedom are subtracted for each cell.

After these procedures, the standard correlation coefficients are calculated. The resulting correlation coefficients \(R\) are insensitive to the factor effects. For example, the correlation between duration and formant reduction can now be calculated after accounting for vowel length, vowel quality, speaking style, speaker sex, and speaker identity.

**Example**

As an example we calculated the simple mean vowel duration (excluding schwa) for four speaking styles and lexical stress (Table 1). We also calculated the corrected means for these vowel realizations, with speaker, text type (fixed, variable), vowel identity, position in the word (Initial, Medial, Final), word-length in syllables, automatically determined prominence (0-4), and open/closed syllable as nuisance factors (Table 2). The vowel realizations available are given in Table 3. The number of quasi-minimal pair differences between the speaking styles (Total row of Table 2) are given in table 4.

From comparing tables 1 and 2, it is clear that the simple mean overestimates the differences between speaking styles, and underestimates the effect of lexical stress.

**Software**

Perl programs to calculate the corrected means and normalized correlations are available under the GNU General Public License from our web site (http://www.fon.hum.uva.nl/IFAcopus).

**Acknowledgements**

Copyrights for the IFA corpus, databases, and associated software are with the Dutch Language Union (Nederlandse Taalunie) and R.J.J.H. van Son. This work was made possible by grant nr 355-75-001 of the Netherlands Organization of Research (NWO) and a grant from the Dutch "Stichting Spraaktechnologie".
References


